Paraninfo

Análisis multivariante aplicado con R

Joaquín Aldás • Ezequiel Uriel

2.ª EDICIÓN

alfacentauro

Estadísticos e-Books & Papers

Análisis multivariante aplicado con R

Joaquín Aldás • Ezequiel Uriel

Estadísticos e-Books & Papers

Estadísticos e-Books & Papers

Análisis multivariante aplicado con R

Joaquín Aldás • Ezequiel Uriel

Estadísticos e-Books & Papers

© 2017, Ediciones Paraninfo, S. A.

2.ª Edición

Calle Velázquez, 31, 3º dcha. / 28001 Madrid
Teléfono: 902 995240 / Fax: 914456218
clientes@paraninfo.es / www.paraninfo.es

© Joaquín Aldás y Ezequiel Uriel

Impresión: Liberdigital (Casarruebuelos, Madrid)

ISBN: 978-84-283-2969-9
Depósito legal: M-25519-2017
Impreso en España

Cualquier forma de reproducción, distribución, comunicación pública o transformación de esta obra solo puede ser realizada con la autorización de sus titulares, salvo excepción prevista por la ley. Diríjase a CEDRO (Centro Español de Derechos Reprográficos, www.cedro.org <http://www.cedro.org>) si necesita fotocopiar o escanear algún fragmento de esta obra.

Estadísticos e-Books & Papers

A Rubén y Mónica por el tiempo robado y al Colegio La Salle Paterna

por su sincera vocación de atención a la diversidad

Joaquín Aldás

A mis hijos Mónica y Jordi

Ezequiel Uriel

Estadísticos e-Books & Papers

Estadísticos e-Books & Papers

Índice general

Presentación **15**

1. Introducción **19**
1.1. Introducción 19
1.2. Niveles de medida de las variables 20
1.2.1. Escala nominal 20
1.2.2. Escala ordinal 21
1.2.3. Escala de intervalo 21
1.2.4. Escala de razón 22
1.2.5. Otras clasificaciones 23
1.3. Clasificación de las técnicas multivariantes 23
1.3.1. Técnicas de análisis de dependencias 23
1.3.2. Técnicas de análisis de interdependencia 26
1.4. Proceso de aplicación de una técnica multivariante 28

2. Análisis previo de los datos **31**
2.1. Introducción 31
2.2. Valores perdidos 31
2.2.1. Diagnóstico de la aleatoriedad de los valores perdidos 34
2.2.2. Tratamiento de los valores perdidos 37
2.3. *Outliers*: casos atípicos 41
2.3.1. Detección univariante de casos atípicos 43
2.3.2. Detección bivariante de casos atípicos 47
2.3.3. Detección multivariante de casos atípicos 47
2.4. Comprobación de los supuestos básicos del análisis multivariante 56
2.4.1. Normalidad 56
2.4.2. Homoscedasticidad 66
2.4.3. Linealidad 74
2.4.4. Independencia de las observaciones 74

3. Análisis de conglomerados **77**
3.1. Introducción 77
3.2. Medidas de similaridad 78
3.2.1. Medidas de similaridad para variables métricas 80
3.2.2. Medidas de similaridad para datos binarios 83
3.2.3. Estandarización de los datos 85

Estadísticos e-Books & Papers

3.3. Formación de los grupos: análisis jerárquico de conglomerados 88
3.3.1. Método del centroide 89
3.3.2. Método del vecino más cercano 93
3.3.3. Método del vecino más lejano 93
3.3.4. Método de la vinculación promedio 93
3.3.5. Método de Ward 96
3.4. Selección del número de conglomerados de la solución 97
3.4.1. Índice CH 98
3.4.2. Índice CCC 100
3.4.3. Índice Pseudo t^2 101
3.4.4. Índice DB 101
3.4.5. Índice de Dunn 101
3.4.6. Estadístico de Hubert 102
3.4.7. Índice Dindex 102
3.5. Formación de los grupos: análisis no jerárquico de conglomerados 106
3.6. Elección entre los distintos tipos de análisis de conglomerados 112
3.6.1. Elección entre análisis de conglomerados jerárquico y no jerárquico 112
3.6.2. Elección entre los distintos métodos de agrupación en el análisis de conglomerados jerárquico 113
3.7. Un ejemplo de aplicación del análisis de conglomerados 114

4. Escalamiento multidimensional 125
4.1. Introducción 125
4.2. El algoritmo básico del MDS 128
4.3. Recogida de datos para un escalamiento multidimensional . . 136
4.3.1. Similaridades directas 136
4.3.2. Similaridades derivadas 137
4.3.3. Consideraciones respecto a la recogida de los datos . . 139
4.4. Tipos de escalamiento multidimensional 140
4.4.1. Escalamiento multidimensional clásico 140
4.4.2. Escalamiento multidimensional ponderado 148
4.4.3. Escalamiento multidimensional clásico desdoblado . . 155

5. Análisis de correspondencias 159
5.1. Introducción 159
5.2. Funcionamiento del análisis de correspondencias 161
5.3. Fundamentación matemática del análisis de correspondencias 170
5.4. Incorporación de puntos suplementarios al análisis de correspondencias simple 178
5.5. Análisis de correspondencias múltiple 181

6. Análisis de la varianza 189
6.1. Introducción 189

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

6.2. Análisis de la varianza de un factor 190
6.2.1. Construcción del estadístico F 191
6.2.2. Supuestos del ANOVA 197
6.2.3. Medida de bondad del ajuste y tamaño del efecto . . . 202
6.2.4. Pruebas *post hoc*: comparaciones múltiples 203
6.3. Análisis de la varianza de dos factores 214
6.3.1. Construcción del estadístico F 215
6.3.2. Ilustración del efecto interacción 219
6.3.3. Medida de bondad del ajuste y tamaño del efecto . . . 222
6.3.4. Pruebas *post hoc*: comparaciones múltiples 224

7. Análisis multivariante de la varianza 227
7.1. Introducción 227
7.2. Análisis multivariante de la varianza con un factor 227
7.2.1. Descomposición de la matriz de covarianzas 228
7.2.2. Cálculo del estadístico de contraste 235
7.2.3. Comprobación de los supuestos en los que se basa el MANOVA 242
7.2.4. Bondad de ajuste y potencia de prueba 246
7.2.5. Pruebas *post hoc* 247
7.3. Análisis multivariante de la varianza con dos factores 250
7.3.1. Descomposición de la matriz de covarianzas 251
7.3.2. Cálculo del estadístico de contraste 251

8. Regresión lineal múltiple 261
8.1. Introducción 261
8.2. El modelo de regresión lineal simple y su estimación por mínimos cuadrados 261
8.3. El modelo de regresión lineal múltiple y su estimación por mínimos cuadrados 266
8.4. Contraste de hipótesis 272
8.4.1. Contraste para el conjunto de parámetros 272
8.4.2. Contraste para un parámetro individual 275
8.4.3. Contraste para un subconjunto de parámetros 277
8.5. Bondad de ajuste del modelo 281
8.5.1. Coeficiente de determinación 282
8.5.2. Coeficiente de determinación corregido 283
8.5.3. Estadístico AIC 283
8.5.4. Error estándar de la estimación 284
8.6. Supuestos del análisis de regresión múltiple 286
8.6.1. Multicolinealidad 288
8.6.2. Normalidad 293
8.6.3. Homocedasticidad 296
8.6.4. Linealidad 301
8.6.5. Independencia de los términos de error 302

Estadísticos e-Books & Papers

8.6.6. Valores atípicos 302
8.7. Modelos con variables ficticias 308
8.7.1. Variable ficticia con dos modalidades 309
8.7.2. Una variable cualitativa con más de dos modalidades . 310
8.7.3. La trampa de las variables ficticias 314
8.7.4. Interacción entre una variable cualitativa y una variable cuantitativa 315
8.7.5. Contraste de cambio estructural 316

9. Análisis discriminante 321
9.1. Introducción 321
9.2. Clasificación con dos grupos 323
9.2.1. Clasificación con dos grupos y una variable clasificadora 323
9.2.2. Clasificación con dos grupos y dos variables clasificadoras 327
9.3. Análisis discriminante con más de dos grupos 349
9.3.1. Obtención de las funciones discriminantes 352
9.3.2. Contrastes de significación 353

10. Regresión logística 365
10.1. Introducción 365
10.2. El modelo de regresión logística binomial 365
10.2.1. Estimación del modelo 368
10.2.2. Contraste de hipótesis para el modelo estimado 369
10.2.3. Interpretación de los coeficientes de regresión 371
10.2.4. Evaluando el ajuste del modelo 374
10.2.5. Capacidad discriminante del modelo 382
10.2.6. Calibración del modelo 385
10.3. Regresión logística multinomial 387

11. Análisis de componentes principales 395
11.1. Introducción 395
11.2. La geometría del análisis de componentes principales 396
11.3. Componentes principales de dos variables 403
11.4. Componentes principales para el caso general 410
11.4.1. Obtención de la primera componente 411
11.4.2. Obtención de las restantes componentes 413
11.4.3. Varianzas de las componentes 413
11.4.4. Correlación entre las componentes principales y las variables originales 414
11.4.5. Puntuaciones sin tipificar y tipificadas 415
11.5. Aspectos operativos en la estimación de un PCA 416
11.5.1. Efecto del tipo de datos sobre el análisis de componentes principales 416
11.5.2. Número de componentes principales que extraer . . . 419
11.5.3. Interpretación de las componentes principales 424

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

12. Análisis factorial exploratorio **431**
12.1. Introducción 431
12.2. Formulación del modelo de análisis factorial exploratorio . . . 433
12.2.1. Hipótesis del modelo 434
12.2.2. Propiedades del modelo 435
12.3. Métodos para la extracción de factores 437
12.3.1. Limitaciones a la extracción de factores 437
12.3.2. Método de las componentes principales 439
12.3.3. Método de los ejes principales 444
12.3.4. Método de máxima verosimilitud 446
12.3.5. Otros métodos de extracción 447
12.3.6. ¿Qué método elegir? 449
12.4. Determinación del número de factores que hay que retener . . 450
12.5. Rotación de la solución factorial 452
12.5.1. Rotación ortogonal 454
12.5.2. Rotación oblicua 459
12.5.3. Cambios que provoca la rotación oblicua en las propiedades del modelo 459
12.6. Indicadores de bondad de la solución factorial 461
12.6.1. Contraste de esfericidad de Bartlett 464
12.6.2. Medidas de adecuación muestral de Kaiser-Meyer-Olkin 464
12.6.3. Diferencias entre las correlaciones observadas y reproducidas 466
12.7. Puntuaciones factoriales 467
12.8. Un ejemplo de aplicación del análisis factorial exploratorio . . 469

13. Modelos de ecuaciones estructurales: análisis factorial confirmatorio **479**
13.1. Introducción 479
13.2. Formalización matemática del análisis factorial confirmatorio (CFA) 482
13.3. La identificación del modelo en un CFA 490
13.4. Estimación del análisis factorial confirmatorio 498
13.4.1. Estimación por mínimos cuadrados no ponderados . . 499
13.4.2. Estimación por mínimos cuadrados generalizados . . . 500
13.4.3. Estimación por máxima verosimilitud 500
13.4.4. Estimación por la teoría de la distribución elíptica . . 500
13.4.5. Estimación con libre distribución asintótica 501
13.4.6. Comparación de los distintos procedimientos de estimación 501
13.5. Bondad de ajuste del modelo estimado 505
13.5.1. Matriz residual de covarianzas 506
13.5.2. Estadístico χ^2 508
13.5.3. *Standardized Root Mean Residual* (SRMR) 510
13.5.4. *Root Mean Square Error of Approximation* (RMSEA) 512
13.5.5. *Tucker-Lewis Index* (TLI) 513

Estadísticos e-Books & Papers

13.5.6. *Comparative Fit Index* (CFI) 514
13.6. Interpretación del modelo 514
13.7. Reespecificación del modelo 516
13.8. Un ejemplo completo de CFA 518
13.9. Anexo 13.1 529

14. Modelos de ecuaciones estructurales: validación del instrumento de medida 531
14.1. Introducción 531
14.2. La medición en ciencias sociales 531
14.3. Análisis de la fiabilidad del instrumento de medida 533
14.3.1. Coeficiente α de Cronbach 533
14.3.2. Fiabilidad compuesta 538
14.4. Análisis de la validez del instrumento de medida 542
14.4.1. Validez de contenido 543
14.4.2. Validez convergente 543
14.4.3. Validez discriminante 546
14.4.4. Validez nomológica 550
14.5. Un ejemplo completo de evaluación del instrumento de medida 550
14.6. Guía para el desarrollo de escalas 556

15. Modelos de ecuaciones estructurales: modelos de estructuras de covarianza (CB-SEM) 567
15.1. Introducción 567
15.2. Formalización matemática del CB-SEM 570
15.3. Identificación del modelo de ecuaciones estructurales 576
15.4. Estimación del modelo de ecuaciones estructurales 583
15.5. Bondad de ajuste del modelo estimado 583
15.6. Interpretación del modelo 587
15.7. Reespecificación del modelo 589
15.8. Un ejemplo completo de modelo de ecuaciones estructurales . 590

16. Modelos de ecuaciones estructurales: modelos de estructuras de varianza (PLS-SEM) 601
16.1. Introducción 601
16.2. El algoritmo de estimación de los modelos PLS-SEM 603
16.3. Cuándo usar PLS-SEM: fortalezas y debilidades 609
16.3.1. Fortalezas 609
16.3.2. Debilidades 610
16.3.3. Criterios de elección entre CBSEM y PLS-SEM 611
16.4. Etapas en la estimación de un modelo estructural mediante PLS-SEM 612
16.4.1. Validación del instrumento de medida 616
16.4.2. Determinación de la significatividad de los parámetros: *bootstrapping* 625

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

16.4.3. Validez y fiabilidad del instrumento de medida (constructos formativos) 627
16.4.4. Evaluación del modelo estructural 632
16.4.5. El debate de los indicadores de ajuste global 638
16.5. Presentación de los resultados en una publicación 641

Bibliografía **661**

Estadísticos e-Books & Papers

Estadísticos e-Books & Papers

Presentación

Este libro es el resultado de muchos años de enseñanza de las técnicas de análisis multivariante en cursos de grado, master y doctorado. Nuestros estudiantes, procedentes fundamentalmente de los estudios de dirección de empresas y economía, desean, al acabar el curso, ser capaces de aplicar rigurosamente las técnicas a problemas concretos sin que la formulación y derivación matemática de dichas técnicas sean un obstáculo insalvable. La mayoría de manuales disponibles para estos cursos, en nuestra opinión, pecan de ser fundamentalmente "libros de recetas" o, en el extremo contrario, de ser excesivamente técnicos. Para cubrir ese hueco hemos escrito Análisis Multivariante Aplicado con R.

Esta obra busca el equilibrio entre rigor y aplicabilidad. Los libros basados solo en la formulación matemática hacen que el lector dude de su capacidad para aplicar la técnica estadística de su interés. Los excesivamente aplicados, basados en listas de "haz esto" y "no hagas aquello", hacen dudar al lector de que sea capaz de conocer las características internas de las técnicas con un mínimo de rigor como para aplicarlas a otros ejemplos fuera de los propuestos.

Para lograr este objetivo, el presente libro utiliza siempre un caso para ilustrar cada paso de los desarrollos matemáticos y justificar su interés. Pero una vez fundamentada matemáticamente la técnica, el lector tiene un segundo caso donde ésta se aplica volviendo a repasar los principales aspectos de sus condiciones de aplicabilidad, bondad del ajuste, forma de interpretar los resultados y generalizabilidad de los mismos. Al final el lector será capaz de saber cuándo y cómo aplicar la técnica, pero también estará seguro de conocer su interior con profundidad.

Otra dificultad con la que nos hemos encontrado en nuestra experiencia docente es que el recurso al *software* comercial como SPSS o SAS, por ejemplo, facilita mucho la labor docente al ser programas muy amigables y de fácil manejo, pero cuando el estudiante deja la universidad y sale del paraguas de las licencias contratadas por la institución, deja de tener acceso a esos programas y tiene que enfrentar sus retos profesionales o investigadores sin el apoyo del *software* con el que está familiarizado. Esta cuestión es especialmente delicada para los estudiantes de doctorado que, una vez finalizado el periodo de formación, se enfrentan a una tesis doctoral sin acceso a esos programas altamente especializados. Por esta razón, el libro que ahora tiene en sus manos, basa el análisis de los casos en distintos paquetes de `R`. `R` (`https://www.r-project.org`) es un entorno para el análisis estadístico que reúne varias ventajas: es gratuito, multiplataforma e incorpora los últimos avances en análisis estadístico fruto de la colaboración desinteresada de una amplia red de colaboradores. Esta decisión garantiza al lector o al profesor que adopte el manual, que el estudiante no

Estadísticos e-Books & Papers

tendrá ninguna limitación futura de aplicar lo aprendido y que, además, estará siempre en la frontera de los avances en análisis de datos.

Es importante señalar que Análisis Multivariante Aplicado con R no es un manual del lenguaje R. porque sería imposible enseñar herramientas estadísticas con la debida profundidad y, simultáneamente, convertir al lector en un experto en este lenguaje. Hay muy buenos libros para profundizar en el mismo (Crawley, 2013; Teetor, 2011) y remitimos al lector a ellos. En este manual ayudamos al lector a aplicar funciones pre-programadas de determinados paquetes de R, por lo que la complejidad potencial de este lenguaje no será nunca una limitación. La convención que utilizamos a lo largo del libro es nombrar a la función que corresponde a cada paquete con un tipo de letra distinto del siguiente modo: `función{paquete}`. En cualquier caso al lector se le proporciona siempre la sintaxis para poder adaptarla a su propia investigación.

Finalmente, este libro está basado en un enfoque multidisciplinar que nace de la diferente especialización de los dos autores. Así, cuando muchos manuales recurren exclusivamente a ejemplos de un campo del conocimiento determinado, en este libro se combinan los casos procedentes del mundo de la economía general, de la economía financiera, del marketing y de la investigación de mercados, permitiendo al lector ver la aplicación de las técnicas a campos muy diversos y apreciar así todo su potencial.

La estructura del libro es la siguiente. Los dos primeros temas introducen el concepto de técnicas multivariantes y explican cómo realizar una exploración previa de los datos que permita evaluar la aplicabilidad de las técnicas que se ofrecen en capítulos consecutivos, a saber, análisis de conglomerados, escalamiento multidimensional, análisis de correspondencias, análisis de la varianza, análisis multivariante de la varianza, regresión lineal múltiple, análisis discriminante, regresión logística, análisis de componentes principales, análisis factorial exploratorio, análisis factorial confirmatorio, validación de los instrumentos de medida de los modelos de ecuaciones estructurales, modelos de ecuaciones estructurales basados en covarianzas, y modelos de ecuaciones estructurales basados en varianzas (PLS-SEM). Los capítulos están apoyados en más de 30 casos completos y decenas de ejemplos. Los ficheros de datos y las sintaxis de R, necesarios para que el lector replique los resultados y explore nuevas posibilidades con cada técnica, aprovechando así todas las posibilidades didácticas de esta obra, se encuentran disponibles en la página web de la editorial: `http://www.paraninfo.es`.

Todo libro es una obra viva. Pese a todo el cuidado que hemos puesto en su edición, seguro que existirán erratas o posibles sugerencias para la mejora de futuras ediciones. Por esta razón, los autores agradecen por anticipado cualquier comentario sobre la obra tendente a mejorarla que pueden dirigir a:

Joaquín Aldás-Manzano
Universitat de València
Facultat d'Economia

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Avda. de los Naranjos s/n
46022-Valencia
joaquin.aldas@uv.es

Este libro no hubiera sido posible sin la ayuda y el apoyo de muchas personas. En primer lugar, y por encima de todo, queremos agradecer los numerosos comentarios ofrecidos por los alumnos de los distintos cursos en los que diversos capítulos de esta obra han sido probados. Especial mención queremos hacer a las muchas mejoras introducidas a partir de sugerencias de los alumnos de la asignatura de Análisis Avanzado de Datos del Doctorado Interuniversitario en Marketing y la de Técnicas Multivariantes de Investigación de Mercados del Master Universitario en Marketing e Investigación de Mercados, impartidos ambos en la Facultat d'Economia de la Universitat de València. Queremos agradecer también al equipo humano de la editorial Paraninfo que ha llevado a cabo una ingente tarea en el diseño, maquetación y revisión de los originales. A todos ellos, muchas gracias.

JOAQUÍN ALDÁS y EZEQUIEL URIEL

Estadísticos e-Books & Papers

Estadísticos e-Books & Papers

1. Introducción

1.1. Introducción

En las últimas décadas se ha producido un gran crecimiento del uso de las técnicas estadísticas multivariantes en todos los campos de la investigación científica. Podrían darse muchas razones para este uso creciente, pero quizás las dos más importantes sean las siguientes (Dillon y Goldstein, 1984):

- En la mayoría de las investigaciones científicas, es necesario analizar relaciones simultáneas entre tres o más variables. La investigación científica es un proceso iterativo. Primero es necesaria la formulación explícita de las hipótesis que después han de contrastarse mediante la recogida y el análisis de los datos. Estos análisis probablemente sugieran una modificación de las hipótesis. En este proceso se añaden y eliminan continuamente variables. La complejidad de los fenómenos analizados hace que sean muchas las variables implicadas y, por ello, las investigaciones sean necesariamente multivariantes.

- El desarrollo de ordenadores con capacidad de almacenamiento y potencia de procesamiento suficiente, acompañados de programas cada vez más fáciles de usar.

Pero ¿cómo definir el análisis multivariante? La tarea no es sencilla. Muchos autores (Hair *et al.*, 2014a) optan por la alternativa de mostrarlo como una extensión del análisis bivariante. Bajo esta perspectiva, el análisis multivariante sería el caso general y las técnicas univariantes o bivariantes serían los casos particulares de la anterior.

De una manera algo más formal, Kachigan (1991) define el análisis multivariante como la rama del análisis estadístico que se centra en la investigación simultánea de dos o más características (variables) medidas en un conjunto de objetos. En esta definición, voluntariamente laxa, el elemento central es la relación simultánea entre las variables. En otras palabras, las técnicas multivariantes difieren de las univariantes y bivariantes en que dirigen su atención no al análisis de la media y la varianza de una variable, o a la correlación entre dos variables, sino al análisis de las covarianzas o correlaciones que reflejan la relación entre tres o más variables.

A lo largo de este libro, utilizaremos el término *objetos* para referirnos a las personas, cosas o entidades de las que se toman las medidas. Las medidas, a

Estadísticos e-Books & Papers

las que casi siempre nos referiremos como *variables*, serán las características o atributos de los objetos que se consideran en la investigación.

1.2. Niveles de medida de las variables

Medir es el proceso mediante el cual se asocian números o símbolos a determinadas características de los objetos, de acuerdo con reglas preestablecidas (Sharma, 1996). Por ejemplo, a los individuos se les puede describir con respecto a características como la edad, la educación, los ingresos, el sexo o la preferencia por una marca u otra, y se deben buscar escalas adecuadas para medir esas características.

El tipo de escala utilizado para medir una variable es fundamental en la elección y aplicación correcta del análisis multivariante. A modo de ejemplo, si queremos establecer si existe una relación de dependencia entre el nivel de ingresos de un individuo y, por ejemplo, su edad, educación o sexo, no será lo mismo si los ingresos están medidos directamente en euros, que si se recurre a una escala donde la medición se hace por intervalos: $1 = [0, 600]$, $2 = [601, 1200]$ y $3 = [1201, \infty[$. En el primer caso podremos recurrir a una regresión lineal, en el segundo, probablemente debamos recurrir a una regresión logística multinomial.

Stevens (1946) consideró que cualquier escala de medida puede clasificarse en alguno de los siguientes cuatro tipos: nominales, ordinales, de intervalo o de razón. Esta clasificación es la más extendida y será la que adoptaremos en este texto. Sin embargo, como señala Sharma (1996), no podemos dejar de señalar que la aplicación de esta clasificación sigue generando debates no resueltos en la literatura estadística. Puede consultarse Velleman y Wilkinson (1993) para profundizar en esta cuestión.

1.2.1. Escala nominal

En este caso, los números asignados a cada característica se comportan como etiquetas, con tanta validez como letras del alfabeto, que de hecho también podrían asignarse. Su misión es distinguir entre diferentes valores; por ejemplo: sexo (hombre, mujer). En el proceso de codificación se puede asignar 1 al valor hombre y 2 al valor mujer. Esto no significa que la mujer sea mayor que el hombre ($2 > 1$) ni el doble ($2 = 1 \times 2$), ni que existan personas de sexo intermedio ($1,5$).

Por ello resulta totalmente inapropiado calcular estadísticos como la media o la varianza de una variable nominal, debiendo limitarnos a los recuentos de frecuencias, moda o tablas de contingencia cuando se cruce con otra variable nominal.

Una exigencia básica de las escalas nominales es que los objetos han de poder clasificarse en categorías que sean mutuamente excluyentes y exhaustivas, es

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

decir, cada individuo debe poder asignarse a una y solo una categoría y todos los individuos han de poder clasificarse en las categorías existentes.

1.2.2. Escala ordinal

No solo consigue distinguir entre valores, como la anterior, sino que además establece un orden entre ellos. Consideremos que a un individuo se le pide que ordene 4 modelos de coche (A, B, C y D) en función de que le gusten más o menos. Su respuesta es $[A = 1,\ D = 2,\ C = 3$ y $B = 4]$. Es obvio que el individuo no solo no prefiere igual al modelo A que al B, sino que, además, prefiere el modelo A más que el B.

Sin embargo es muy importante señalar que aunque las diferencias numéricas entre las categorías sean numéricamente las mismas, esto no quiere decir que las diferencias de preferencia también lo sean. La diferencia de preferencia entre el automóvil A y el D no tiene por qué ser la misma que entre el C y el B, aunque $(2 - 1 = 4 - 3)$. Tampoco el automóvil A se prefiere el doble que el D. Por lo tanto, en las escalas ordinales, tiene sentido distinguir y ordenar, pero no las diferencias ni las razones.

Los estadísticos que pueden calcularse en este tipo de escalas son, además de los que se calculaban en las nominales, medianas y percentiles.

1.2.3. Escala de intervalo

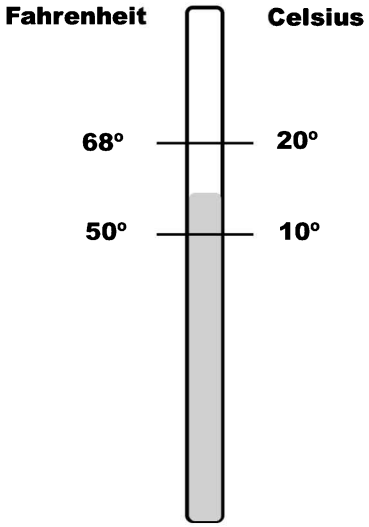
Una escala de intervalo posee las características de una nominal (diferentes valores representan diferentes características de los objetos) y de la ordinal (mayor valor representa mayor presencia de la característica, por ejemplo la preferencia). Sin embargo la escala de intervalo añade una nueva propiedad: las diferencias también tienen sentido. Un ejemplo, que además es útil para distinguir estas escalas de las de razón, es la medición de la temperatura mediante, por ejemplo, una escala Celsius. Si el termómetro marca 35°, marca una temperatura distinta a si marca 30° (como la nominal), pero también marca más temperatura (como la ordinal). Sin embargo, además, entre 35° y 30° hay la misma diferencia de temperatura que entre 30° y 25°: las diferencias iguales en los códigos se traducen en diferencias iguales en el atributo medido.

Pero ¿tienen sentido las razones? Esto no es tan sencillo de ver. Aparentemente 20° es el doble de temperatura que 10°. Sin embargo veamos esta misma medición en una escala Fahrenheit (figura 1.1). El atributo que se está midiendo es el mismo, la temperatura, pero decir que 20° es el doble de 10° en Celsius implicaría decir que 68° es el doble de 50° en Fahrenheit, y esto no es así.

Siempre que el atributo que se esté midiendo no tenga un cero absoluto, sino que este sea arbitrario, estaremos ante escalas de intervalo. Es el caso, por ejemplo, de los calendarios, donde el cero se ha colocado arbitrariamente en el nacimiento de Cristo, pero hay calendarios con otros ceros y, de nuevo, las diferencias de años tendrán sentido pero las razones, cuando se comparen, no.

Estadísticos e-Books & Papers

Figura 1.1.: Ilustración de una escala de intervalo



En estas escalas pueden calcularse todos los estadísticos menos los que están basados en ratios, como el coeficiente de variación.

En investigación de mercados es muy habitual el recurso a escalas de intervalo para medir, por ejemplo, acuerdos o desacuerdos con determinadas afirmaciones (1 = totalmente en desacuerdo, 5 = totalmente de acuerdo). Aunque no es evidente, es importante que se tenga en cuenta que en el diseño de estas escalas se está asumiendo que diferencias iguales en la codificación implican diferencias iguales en el grado de acuerdo pues, de no ser así, nos encontraríamos ante una escala ordinal.

1.2.4. Escala de razón

Las escalas de razón tienen las mismas propiedades que las de intervalo pero, además, las razones sí que tienen sentido. Estas escalas tienen un valor base 0 natural: la edad, los ingresos, una escala de temperatura Kelvin. Si un individuo tiene 20 años y otro tiene 10, no solo tienen distintas edades (nominal), el primero es mayor que el segundo (ordinal) y hay la misma diferencia de edad entre el primero y el segundo que entre el primero y un sujeto de 30 años (intervalo) sino que podemos afirmar sin problemas que el primero tiene el doble de edad que el segundo.

No hay ninguna restricción respecto a los estadísticos que pueden calcularse en este tipo de escalas.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

1.2.5. Otras clasificaciones

Como se ha señalado, la expuesta no es la única de las clasificaciones de variables posible, aunque sí, la más implantada. Es necesario, sin embargo, precisar sobre la base de la clasificación presentada algunas otras formas de referirse a las escalas que el lector puede encontrarse.

Es muy habitual simplificar la clasificación de Stevens dejándola en dos grupos, el que se correspondería con variables *no métricas* (nominales y ordinales) y el de variables *métricas* (de intervalo y razón).

También es habitual distinguir entre variables *discretas* y *continuas*. Esta distinción se basa en los posibles valores que la variable puede tomar. Una variable discreta solo puede tomar un número determinado de valores en un intervalo determinado: errores en un contraste, crímenes en una ciudad, admisiones en un hospital, etc. Una variable continua, por el contrario, puede tomar potencialmente cualquier valor numérico en un intervalo dado. El peso de un individuo puede ser tanto de 70,0 kg como de 70,054556 kg.

1.3. Clasificación de las técnicas multivariantes

La importancia de una adecuada clasificación de las técnicas multivariantes no reside tanto en la necesidad tipológica, sino en que es necesario presentar al lector una guía que le permita la elección adecuada de la técnica que debe aplicarse en función del problema que pretenda resolver. Este es el fin de este epígrafe.

Antes de plantearnos la elección de una técnica u otra, es necesario que seamos capaces de responder a las siguientes preguntas básicas (Dillon y Goldstein, 1984):

1. ¿Nuestra investigación responde a un problema de dependencia entre variables o de interdependencia entre las mismas?

2. ¿Cómo están medidas las variables implicadas, en escalas métricas o no métricas?

3. Si estamos ante un problema de dependencia, ¿cuántas relaciones se plantean entre las variables dependientes e independientes? ¿Cuántas variables dependientes existen?

Si se es capaz de estructurar el problema de análisis para responder a las preguntas planteadas, la elección de la técnica se simplifica bastante.

1.3.1. Técnicas de análisis de dependencias

Supongamos que nos encontramos ante dos grupos de variables. Las técnicas de análisis de dependencias buscarán la existencia o ausencia de relaciones entre los dos grupos de variables. Si el investigador, basándose en un experimento

Estadísticos e-Books & Papers

controlado o gracias a una base teórica previa, clasifica los dos grupos de variables en dependientes e independientes, entonces el objetivo de las técnicas de dependencia será establecer si el conjunto de variables independientes afecta al conjunto de dependientes de manera conjunta o individualmente.

Si de un conjunto de individuos se conocen sus ingresos, nivel de estudios, edad y sexo, podemos plantearnos si existe una relación entre los ingresos (variable dependiente) y el resto de variables. Estaríamos ante un problema de *análisis de dependencia* y sería necesario ver cómo están medidas las variables para elegir entre una técnica u otra.

Sin embargo, podemos encontrarnos ante un problema en el que sea imposible distinguir conceptualmente entre variables dependientes e independientes. Nos interesa simplemente saber cómo se relacionan entre sí todas las variables del problema. Los métodos estadísticos que abordan estas cuestiones serían los denominados *métodos de interdependencia.*

Siguiendo el ejemplo anterior, el investigador puede querer saber si considerando todas las variables que caracterizan a los individuos (ingresos, nivel de estudios, edad y sexo) pueden encontrarse grupos de individuos que se parezcan mucho entre sí respecto a estas variables y que difieran de otros grupos. Aquí no nos encontramos ante dos grupos de variables, sino que se consideran todas juntas. La técnica que se elija para resolver este problema deberá pertenecer al grupo de métodos de interdependencia.

La figura 1.2 ilustra el proceso de elección de cada técnica de dependencia atendiendo a las preguntas que se planteaban al principio de este epígrafe. Asimismo, se muestra el capítulo del presente libro en el que cada técnica es analizada.

Siguiendo este esquema, el **análisis de regresión lineal múltiple** se empleará cuando se pretenda analizar la relación existente entre una variable dependiente métrica y un conjunto de variables independientes que pueden o no ser métricas, de acuerdo con el esquema simplificado:

$$Y \Leftarrow (X_1, X_2, X_3, \ldots, X_m)$$
$$métrica \Leftarrow métricas,\ no\,métricas$$

A modo de ejemplo, si queremos saber si existe o no relación entre el nivel de ingresos de un individuo medido en euros (variable dependiente métrica) y las ya mencionadas variables de nivel educativo, sexo y edad.

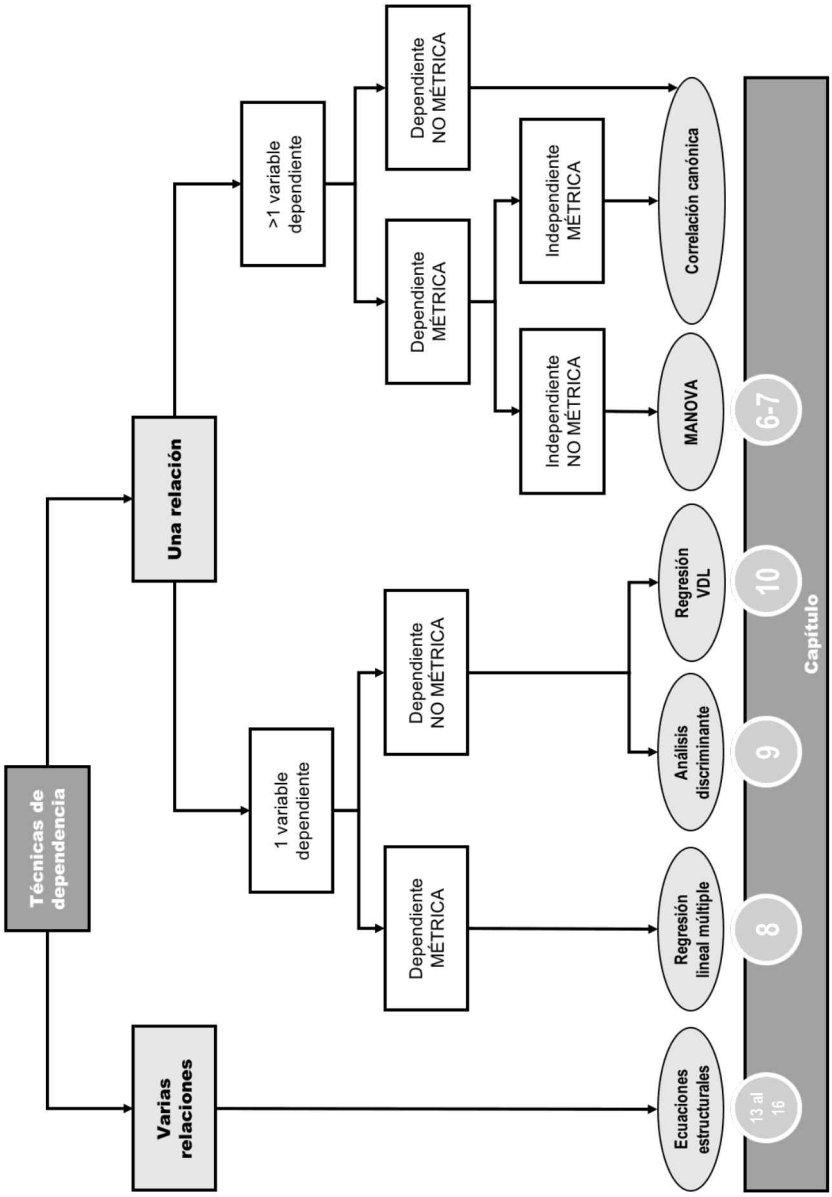
Pero ¿qué ocurre si la variable dependiente no es métrica? Por ejemplo si el nivel de ingresos está codificado como 1 = por debajo de la media, 2 = por encima de la media. Entonces ya no se puede recurrir a una regresión lineal y se ha de optar por el **análisis discriminante** o la **regresión de variable dependiente limitada**. Aun siendo el mismo el objetivo de análisis, el modo en que está medida la variable dependiente condiciona la elección de la técnica.

El **análisis de correlación canónica** pretende determinar la existencia de asociación lineal entre un conjunto de variables independientes y otro conjunto de variables dependientes, de acuerdo con el esquema simplificado:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 1.2.: Técnicas de análisis de dependencia



Fuente: Adaptado de Dillon y Goldstein (1984, p. 21)

Estadísticos e-Books & Papers

$$(Y_1, Y_2, Y_3, \ldots, Y_n) \Leftarrow (X_1, X_2, X_3, \ldots, X_m)$$
$$\textit{métricas},\ \textit{no métricas} \Leftarrow \textit{métricas},\ \textit{no métricas}$$

Siguiendo el ejemplo que venimos empleando, el investigador puede querer establecer cómo influyen el nivel educativo, el género y la edad no solo sobre el nivel de ingresos, sino, por ejemplo, sobre el nivel de satisfacción con el empleo actual (medido, por ejemplo, mediante una escala de intervalo donde 1 = totalmente insatisfecho, 5 = totalmente satisfecho). Nótese que la regresión lineal múltiple sería un caso particular del análisis de correlación canónica cuando solo se dispone de una variable dependiente. Esta técnica, en cuanto que poco utilizada, no se incluye en esta obra.

En el ejemplo que hemos utilizado, las variables independientes son métricas (edad, nivel educativo) y no métricas (sexo), mientras que ambas dependientes son métricas. En el caso en que todas las independientes fueran no métricas y las dependientes siguieran siendo métricas, el mismo objetivo logrado con el análisis de correlación canónica podría conseguirse mediante un **análisis multivariante de la varianza (MANOVA)**.

En todos los casos expuestos hasta ahora el investigador buscaba evaluar la intensidad de una sola relación entre dos conjuntos de variables, pero ¿existen alternativas cuando no es una única ecuación la que recoge las relaciones sino varias, esto es, se analizan varias relaciones? Este objetivo correspondería al siguiente esquema simplificado:

$$\begin{gathered} Y_1 \Leftarrow (X_{11}, X_{12}, X_{13}, \ldots, X_{1m}) \\ Y_2 \Leftarrow (X_{21}, X_{22}, X_{23}, \ldots, X_{2m}) \\ \vdots \\ Y_n \Leftarrow (X_{n1}, X_{n2}, X_{n3}, \ldots, X_{nm}) \\ \textit{métricas} \Leftarrow \textit{métricas} \end{gathered}$$

Este conjunto de relaciones es el objeto del conjunto de técnicas denominadas **sistemas de ecuaciones estructurales**, tres de las cuales, análisis factorial confirmatorio, modelos de estructuras de covarianzas y modelos basados en varianzas (PLS-SEM), se abordan en este texto.

1.3.2. Técnicas de análisis de interdependencia

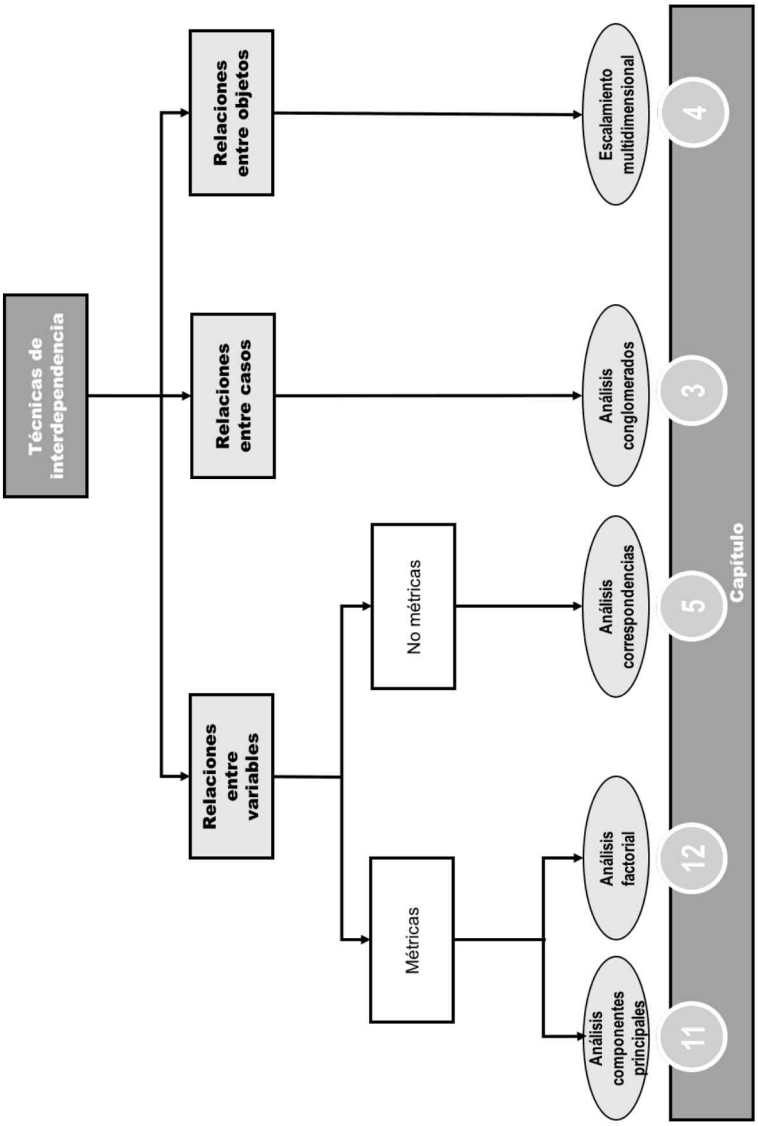
Como se ha comentado con anterioridad, existen situaciones de investigación en las que es imposible separar las variables en dependientes e independientes, y lo que interesa es determinar cómo y por qué las variables están correlacionadas entre ellas. Un esquema de la lógica de las mismas se presenta en la figura 1.3.

El **análisis de componentes principales** es una técnica de reducción de datos cuyo objetivo fundamental es construir combinaciones lineales de las variables originales que expliquen la mayor parte posible de la información recogida en las variables originales. Cada combinación lineal se extrae de tal forma que está incorrelacionada con las anteriores. Cada combinación lineal

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 1.3.: Técnicas de análisis de interdependencia



Fuente: Elaboración propia a partir de Hair *et al.* (2014a).

Estadísticos e-Books & Papers

aporta, cada vez, menos información que la anterior. Por ejemplo, un analista contable puede extraer del balance muchas ratios financieras para medir la salud de una empresa. Puede emplearlas todas o, por el contrario, obtener digamos dos combinaciones de ellas. Es más fácil comparar empresas con dos datos que con un centenar. En este sentido, el análisis de componentes principales le permite obtener de manera adecuada esos dos indicadores.

El **análisis factorial** también es una técnica de reducción de datos pero, a diferencia de la técnica anterior, el interés se centra en establecer qué causas latentes (factores) pueden estar causando la correlación entre las variables manifiestas, es decir, entre las variables observadas. Puede verse como una técnica que pretende identificar grupos de variables de tal forma que las correlaciones entre las variables de cada grupo sea superior a las correlaciones de las variables entre los grupos.

El **análisis de conglomerados** lo que pretende, a diferencia del análisis factorial, no es agrupar variables, sino observaciones, de tal forma que las observaciones de cada grupo sean parecidas respecto a las variables que se han utilizado para agrupar y las observaciones entre los grupos sean lo más diferente posible respecto a esas variables.

El **análisis de escalamiento multidimensional** permite al investigador hacer aflorar los criterios subyacentes que utilizan los individuos para considerar que distintos objetos son parecidos o distintos. Una de las principales utilidades de este análisis son mapas, normalmente bidimensionales, donde los objetos están tanto más cercanos cuanto más parecidos son para el conjunto de individuos entrevistados.

Finalmente, el **análisis de correspondencias** permite visualizar gráficamente tablas de contingencia. Imaginemos que deseamos ver si existe relación entre dos variables no métricas, cada una de las cuales tiene, digamos, 20 niveles. Por ejemplo si existe relación entre el tipo de ocupación y la universidad donde el sujeto se licenció. Esta tabla de 20×20 no será fácil de analizar, sin embargo, el análisis de correspondencias permitirá mostrar un mapa, probablemente bidimensional, en el cual una universidad estará tanto más cerca de un tipo de ocupación, cuanto mayor sea la frecuencia de esa celda en la tabla de contingencia.

1.4. Proceso de aplicación de una técnica multivariante

Las técnicas multivariantes son, en general, herramientas muy poderosas que permiten al investigador extraer muchísima información de los datos disponibles. Estas técnicas son, en sí mismas, relativamente complejas y requieren para su utilización un conocimiento profundo de sus fundamentos y condiciones de aplicabilidad. El desarrollo de programas informáticos de manejo sencillo, como SPSS, ha provocado su uso indiscriminado y, muchas veces, no se utilizan adecuadamente.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

El objeto de este epígrafe es proporcionar al lector una guía que le permita la aplicación correcta de las técnicas multivariantes y le facilite el llegar a conclusiones razonables. Este epígrafe sigue en su desarrollo el proceso planteado por Hair *et al.* (2014a).

Paso 1. Definición del problema que está investigando, sus objetivos y decisión sobre la técnica multivariante que se debe utilizar. No se puede pretender utilizar una técnica multivariante sin una adecuada aproximación teórica al problema que se está abordando. El investigador debe analizar conceptualmente su objeto de investigación, definir los conceptos e identificar las relaciones fundamentales que se pretenden investigar. Primero hay que centrarse en el tema que se investiga, y no en las técnicas que se van a emplear, lo que evitará que dejemos fuera del análisis conceptos importantes. Una vez, y solo una vez realizado esto, el lector puede seguir los esquemas de la sección 1.3 para seleccionar la técnica más adecuada.

Paso 2. Desarrollo del plan de análisis. Una vez establecido el modelo conceptual, el énfasis se centra en aplicar adecuadamente la técnica elegida, lo que hace referencia fundamentalmente a los tamaños muestrales mínimos que permiten su aplicación, y a asegurarse de que el procedimiento de recogida de datos (por ejemplo, los cuestionarios) miden las variables con las escalas oportunas (métricas *vs.* no métricas).

Paso 3. Análisis las condiciones de aplicabilidad de la técnica elegida. Una vez recogidos los datos, es necesario conocer cuáles son las hipótesis en que se basan las técnicas multivariantes y, que si no se cumplen, hace que carezca de sentido aplicarlas. En las técnicas de dependencia, por ejemplo, suele ser necesario que los datos cumplan las hipótesis de normalidad, linealidad, independencia del término de error y homoscedasticidad. Una aproximación a estas comprobaciones se realizará en el capítulo 2.

Paso 4. Estimación del modelo multivariante y ajuste global del mismo. Aplique la técnica multivariante elegida. Pero fíjese si el nivel de bondad del ajuste es adecuado. Si no es así, deberá reespecificarse el modelo, incorporando o eliminando variables. No todas las técnicas tienen indicadores de bondad de ajuste.

Paso 5. Interpretación de los resultados. Una vez logre un nivel de ajuste aceptable, interprete el modelo. Fíjese en los efectos de las variables individuales examinando sus coeficientes, cargas factoriales, utilidades... La interpretación puede conducirle a nuevas reespecificaciones del modelo.

Estadísticos e-Books & Papers

Paso 6. Validación del modelo. Antes de aceptar los resultados a los que haya llegado, debe aplicar una serie de técnicas de diagnóstico que aseguren que estos resultados son generalizables al conjunto de la población.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

2. Análisis previo de los datos

2.1. Introducción

En el capítulo anterior ya se ha indicado que es necesario dar una serie de pasos previos antes de aplicar una técnica multivariante determinada. Algunos de ellos tienen que ver con la propia técnica y la comprobación del cumplimiento de sus hipótesis subyacentes, por ejemplo, normalidad, linealidad, homocedasticidad, etc. Otras comprobaciones son, incluso, previas a la técnica y tienen que ver con la fiabilidad de los datos de partida: existencia de valores perdidos y de observaciones anómalas.

A la realización de estas comprobaciones previas se dedica el presente capítulo. Debe señalarse que algunas de las técnicas de análisis que se expondrán en capítulos posteriores tienen sus propios procedimientos para la comprobación del cumplimiento de sus hipótesis o, por ejemplo, la detección de las observaciones anómalas y así serán presentadas en su momento (piénsese, por ejemplo en la regresión lineal múltiple). El presente capítulo pretende ofrecer los necesarios procedimientos previos a la aplicación de cualquier técnica, tanto más necesarios cuanto que algunas de ellas no disponen de herramientas específicas.

2.2. Valores perdidos

La existencia de valores perdidos es algo prácticamente inevitable en la investigación en Ciencias Sociales. Los entrevistados en una encuesta se niegan a declarar su nivel de ingresos, el entrevistador no recoge una respuesta en la casilla adecuada o, simplemente, la tasa de paro no está disponible todavía para el semestre que se analiza.

Las consecuencias para la investigación de la existencia de valores perdidos depende del patrón que siguen estos datos ausentes, cuántos son y por qué están perdidos. Como señalan Tabachnick y Fidell (2001), el patrón de los valores perdidos es más importante que su cuantía. Si su distribución es aleatoria en la matriz de datos, no pueden causar mucho daño al análisis. Sin embargo, si responden a un patrón determinado, sí. Veámoslo con un ejemplo.

Caso 2.1 Actitud hacia el tabaco de los jóvenes

Un investigador desea saber cuál es la actitud hacia el tabaco de los jóvenes, para ello les pide que expresen su acuerdo o desacuerdo con un conjunto de afirmaciones (la escala de respuesta es 1 = estoy en total desacuerdo, 5 = estoy en total acuerdo), tal y como se recoge en el cuadro 2.1. Junto con estas

Estadísticos e-Books & Papers

Cuadro 2.1.: Preguntas de actitud hacia el tabaco

Variable	Afirmación
V1	Fumar perjudica la salud
V2	No debe permitirse fumar en lugares públicos
V3	A los poderes públicos solo les interesa recaudar impuestos
V4	Deben aumentarse los impuestos sobre el tabaco
V5	Debe informarse más sobre los efectos del tabaco
C1	Edad (años)
C2	Sexo (1=Hombre; 2=Mujer)
C3	Hábito (1=Fumador; 2=No fumador)

preguntas se realizan otras de clasificación, como son la edad, el sexo y el hábito, esto es, si se es o no fumador.

El cuadro 2.2 recoge un conjunto de respuestas simuladas a este cuestionario donde los valores perdidos de la variable V4 se han asignado aleatoriamente (V4a), mientras que, por el contrario, en V4b los valores perdidos de V4 siguen un patrón: los fumadores se niegan a contestar en mucha mayor medida que los no fumadores, por ejemplo porque pueden pensar que la opinión vertida en la encuesta puede tener alguna influencia en la Administración y puede conllevar una nueva subida del precio. Las variables V4a_d, V4b_d y V2_d serán generadas en el proceso de análisis de los valores perdidos y su construcción se explicará en su momento.

Si el objetivo del investigador es determinar cuál es el nivel de acuerdo con que se suban los impuestos sobre el tabaco, en el primer caso (distribución aleatoria de los valores perdidos), la media de V4a no debería diferir sustancialmente de la media muestral con la muestra completa. Sin embargo, si son los no fumadores (que probablemente estén más en desacuerdo con esta medida) los que principalmente no responden, la media de v4b puede elevarse artificialmente (mayor valor, más acuerdo).

Como se comprueba en el cuadro 2.3, esto es exactamente lo que ocurre, al ser los no fumadores quienes mayoritariamente no responden, el acuerdo con que se suban los impuestos es superior (V4b) a cuando la distribución de los valores perdidos era aleatoria (V4a).

Aunque la tentación es asumir que los valores perdidos se han generado de manera aleatoria, las graves consecuencias para la investigación de que esto no sea así obligan a desarrollar estrategias para determinar la aleatoriedad de los valores perdidos. Se verán dos procedimientos planteados por Tabachnick y Fidell (2001): (1) comprobar si los casos con valores perdidos tienen valores medios de otras variables relacionadas distintos a los casos sin valores perdidos y (2) comprobar si existe relación entre la tendencia a no contestar a dos variables que tengan valores perdidos.

Estos análisis se corresponden con la comprobación de las dos condiciones que han de darse para lo que Rubin (1976) y Little y Rubin (1987) definen

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 2.2.: Respuestas simuladas al cuestionario

Caso	V1	V2	V3	V4a	V4b	V5	C1	C2	C3	V4a_d	V4b_d	V2d
1	5	5	4	.	5	5	21	2	2	0	1	1
2	5	5	4	4	4	5	21	2	2	1	1	1
3	5	5	4	2	2	5	21	1	2	1	1	1
4	5	4	3	.	3	4	20	2	2	0	1	1
5	5	5	2	5	5	5	24	2	2	1	1	1
6	5	5	5	5	5	5	26	2	1	1	1	1
7	5	.	5	1	.	4	22	2	1	1	0	0
8	5	4	3	3	3	5	23	1	2	1	1	1
9	4	4	4	1	1	5	22	2	2	1	1	1
10	5	2	3	.	.	3	21	2	1	0	0	1
11	5	5	3	3	3	5	23	1	2	1	1	1
12	5	4	2	4	4	5	21	1	2	1	1	1
13	5	3	4	2	2	4	23	2	2	1	1	1
14	5	4	5	1	1	3	22	2	1	1	1	1
15	5	5	5	3	3	4	24	2	2	1	1	1
16	5	.	3	2	.	5	27	1	1	1	0	0
17	5	.	5	1	.	3	21	1	1	1	0	0
18	5	4	4	.	.	3	20	1	1	0	0	1
19	5	4	2	4	4	4	21	2	2	1	1	1
20	1	5	3	4	4	5	23	2	2	1	1	1
21	5	4	3	4	4	5	20	2	2	1	1	1
22	5	5	3	4	4	5	20	2	2	1	1	1
23	5	4	5	4	4	5	20	2	1	1	1	1
24	5	3	1	5	5	5	22	1	1	1	1	1
25	5	3	5	1	1	3	24	1	1	1	1	1
26	4	.	5	1	.	5	23	2	2	1	0	0
27	2	3	5	1	.	3	20	2	2	1	0	1
28	4	4	5	.	2	5	22	2	2	0	1	1
29	5	5	4	.	5	5	22	1	1	0	1	1
30	5	5	5	5	5	5	23	2	2	1	1	1

Estadísticos e-Books & Papers

Cuadro 2.3.: Media de V4 en función del tipo de valor perdido

```
     caso            v1              v2              v3          v4a
 Min.   : 1.00   Min.   :1.000   Min.   :2.000   Min.   :1.0   Min.   :1.000
 1st Qu.: 8.25   1st Qu.:5.000   1st Qu.:4.000   1st Qu.:3.0   1st Qu.:1.000
 Median :15.50   Median :5.000   Median :4.000   Median :4.0   Median :3.000
 Mean   :15.50   Mean   :4.667   Mean   :4.154   Mean   :3.8   Mean   :2.917
 3rd Qu.:22.75   3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:5.0   3rd Qu.:4.000
 Max.   :30.00   Max.   :5.000   Max.   :5.000   Max.   :5.0   Max.   :5.000
                                 NA's   :4                     NA's   :6
     v4b             v5              c1              c2
 Min.   :1.000   Min.   :3.000   Min.   :20.00   Min.   :1.000
 1st Qu.:2.500   1st Qu.:4.000   1st Qu.:21.00   1st Qu.:1.000
 Median :4.000   Median :5.000   Median :22.00   Median :2.000
 Mean   :3.435   Mean   :4.433   Mean   :22.07   Mean   :1.667
 3rd Qu.:4.500   3rd Qu.:5.000   3rd Qu.:23.00   3rd Qu.:2.000
 Max.   :5.000   Max.   :5.000   Max.   :27.00   Max.   :2.000
 NA's   :7
      c3            v4a_d          v4b_d              v2_d
 Min.   :1.000   Min.   :0.0   Min.   :0.0000   Min.   :0.0000
 1st Qu.:1.000   1st Qu.:1.0   1st Qu.:1.0000   1st Qu.:1.0000
 Median :2.000   Median :1.0   Median :1.0000   Median :1.0000
 Mean   :1.633   Mean   :0.8   Mean   :0.7667   Mean   :0.8667
 3rd Qu.:2.000   3rd Qu.:1.0   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :2.000   Max.   :1.0   Max.   :1.0000   Max.   :1.0000
```

como aleatoriedad completa o MCAR (*Missing Completely at Random*), esto es, que los valores perdidos sean independientes tanto de los valores observados del resto de variables del problema como de los valores perdidos de esas mismas variables.

2.2.1. Diagnóstico de la aleatoriedad de los valores perdidos

El **primer procedimiento** para establecer si los valores perdidos guardan o no un patrón sistemático se basa en la lógica de la investigación. Si el patrón es sistemático, los casos con valores perdidos deberán tener un comportamiento distinto respecto a otras variables que los casos sin valores perdidos. En nuestro ejemplo, como son los fumadores quienes no han querido contestar principalmente a la pregunta sobre si deben aumentarse los impuestos (V4b), es probable que los casos con valores perdidos (principalmente fumadores) estén más en desacuerdo con que, por ejemplo, no se permita fumar en lugares públicos (V2) que los casos sin valores perdidos.

Es evidente que el investigador no puede tener esta hipótesis a priori, y deberá comprobar qué variables se comportan de manera distinta en los dos grupos para deducir la existencia o no de un patrón. De no existir variables cuya media sea distinta en los casos con y sin valores perdidos, habrá que asumir la

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 2.4.: Prueba t para muestras independientes

	VP aleatorios (V4a_d)		
	1 (sin VP)	0 (Con VP)	t
V2 (media)	3,96	3,83	0,23

$^{**}p < 0,01$

	VP sistemáticos (V4b_d)		
	1 (sin VP)	0 (Con VP)	t
V2 (media)	4,30	2,71	-3,95**

$^{**}p < 0,01$

aleatoriedad de los mismos.

En nuestro ejemplo habíamos generado dos variables V4 con valores perdidos generados aleatoriamente (V4a) y respondiendo a un patrón (V4b). Veamos si otras variables (V2: no debe permitirse fumar en lugares públicos) tienen el mismo comportamiento en el grupo de casos con valores perdidos y los que no lo tienen. Para ello es necesario crear una variable ficticia que tomará el valor 1 si el caso tienen un valor perdido en V4 y 0 si no lo tiene. En el cuadro 2.2 estas variables aparecen etiquetadas como V4a_d y v4b_d.

Para contrastar si la media de la variable V2 es igual o distinta en el grupo de casos con valores perdidos respecto al que no los tiene, efectuamos una prueba t para muestras independientes, donde la variable dependiente es V2 y el factor serán las variables que especifican si estamos ante el grupo de valores perdidos o el que no los tiene (V4a_d y V4b_d, respectivamente). Los resultados se muestran en el cuadro 2.4. La hipótesis nula es que las medias son iguales en los dos grupos. Valores de t significativos implicarán el rechazo de esa hipótesis.

Como se deduce del cuadro 2.4, la variable V2 (no debe permitirse fumar en lugares públicos) no tiene una media significativamente distinta en los grupos con y sin valores perdidos para V4 cuando estos son aleatorios. Si esta conclusión se obtuviera para la mayoría de las variables, podríamos concluir que los valores perdidos no siguen un patrón dado.

Sin embargo, se constata como, cuando los valores perdidos corresponden a un patrón dado (fumadores), la media del grupo con valores perdidos es significativamente inferior a la del grupo sin valores perdidos, esto es, están más en desacuerdo con que se prohíba fumar en lugares públicos. Si el investigador obtuviera este resultado para más variables, debería concluir que los valores perdidos responden a un patrón sistemático.

El **segundo procedimiento** para evaluar la aleatoriedad de los valores perdidos consiste en ver si existe una coincidencia significativa entre los casos concretos en que las variables toman un valor perdido. Pérez (2004) denomina a este procedimiento prueba de las correlaciones dicotomizadas, siguiendo la terminología que utiliza el programa `BMDPAM` según lo presentan Tabachnick y Fidell (2001).

Estadísticos e-Books & Papers

Cuadro 2.5.: Matriz de correlaciones

```
 Pearson correlations:
         v2_d   v4a_d  v4b_d
v2_d   1.0000 -0.1961 0.7110
v4a_d -0.1961  1.0000 0.1182
v4b_d  0.7110  0.1182 1.0000

 Number of observations: 30

 Pairwise two-sided p-values:
      v2_d   v4a_d  v4b_d
v2_d         0.2990 <.0001
v4a_d 0.2990        0.5338
v4b_d <.0001 0.5338

 Adjusted p-values (Holm's method)
      v2_d   v4a_d  v4b_d
v2_d         0.5979 <.0001
v4a_d 0.5979        0.5979
v4b_d <.0001 0.5979
```

En nuestro ejemplo, podemos plantearnos si los entrevistados que no contestan a V4 son más o menos los mismos que los que no contestan a otras variables, por ejemplo, a V2. Si por ser fumador no se quiere declarar que se es favorable a que se suban los impuestos, por si se hace, es posible que tampoco se quiera declarar de acuerdo con que se impida fumar en lugares públicos. De ser así estaríamos ante una situación en que los valores perdidos de distintas variables tienen una causa común y, por tanto, comparten un patrón.

El procedimiento para detectar esta relación es sencillo. Basta con convertir las variables que se quieren analizar en variables ficticias que tomarán el valor 1 si para ese caso la variable original no toma un valor perdido y 0 en caso contrario, es decir, lo mismo que se hizo en el caso anterior con V4 y que, ahora, haremos también con V2 (en el cuadro 2.2 aparece esta nueva variable como V2_d). A continuación se calcula la matriz de correlaciones entre las variables implicadas y se analiza la significatividad de los coeficientes (cuadro 2.5). En nuestro ejemplo cabe esperar que, cuando la generación de los valores perdidos ha sido aleatoria (V4a), su variable dicotomizada no guarde correlación significativa con la que muestran los casos perdidos de V2 (V2_d), mientras que, cuando la generación de los valores perdidos responde a una causa común (ser fumador), cabe esperar que la matriz de correlaciones haga aflorar esta relación.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

El hecho de que la correlación sea significativa y fuerte entre los casos en que V4 y V2 toman valores perdidos debería hacer sospechar al investigador que puede existir un motivo subyacente (caso de V4b que, recordemos, no fue una generación aleatoria, sino que respondía a una negativa de responder de los fumadores que se repetía en V2). Si, por el contrario, la situación fuera la de una correlación no significativa (V4a, que fueron valores perdidos generados aleatoriamente), el investigador puede suponer razonablemente que se encuentra ante una deseable situación de MCAR.

Cuando los valores perdidos responden a un patrón, nos encontramos ante un grave problema pues, según indica Byrne (2001), (a) no hay medios estadísticos conocidos para reducir el número de valores perdidos y (b) se imposibilita la generalizabilidad de los resultados. Sin embargo, ante una situación de MCAR, sí que se dispone de estas estrategias, las cuales desarrollaremos a continuación.

2.2.2. Tratamiento de los valores perdidos

Básicamente existen dos grandes procedimientos para tratar los valores perdidos: la eliminación de los casos que los contienen o la imputación de un valor estimado a la variable en el caso en que tome un valor perdido.

La eliminación de todos los casos que tengan un valor perdido es el procedimiento más utilizado debido a que es el que la mayor parte de programas estadísticos tienen por defecto. Esto provoca que, si el investigador no realiza una exploración previa de los datos, el programa puede estar eliminando casos sin su conocimiento. Es más, se eliminan los casos con valores perdidos aunque estos estén en variables que no se usan en el análisis.

La generalización del uso de este procedimiento se debe a que algunas técnicas (los modelos de ecuaciones estructurales, por ejemplo) dan muchos problemas cuando sus matrices de varianzas-covarianzas se basan en datos incompletos (véase Bentler y Chou, 1997; Boomsma, 1985).

La principal limitación de este procedimiento es, obviamente, la pérdida de información que se produce al trabajar con una muestra más reducida, sobre todo si la muestra de partida no era muy amplia y los valores perdidos no se concentran en unos casos determinados sino que se distribuyen por muchos de ellos. Asimismo, este procedimiento asume una distribución MCAR de los valores perdidos, pues, de no ser así, las estimaciones estarán sesgadas, independientemente del tamaño muestral.

En el cuadro 2.6 mostramos el cálculo de las medias de las variables eliminando todos aquellos casos que tienen un valor perdido, lo que la mayoría de programas denomina eliminación según lista o *listwise deletion*. Dado que la eliminación según lista exige MCAR, solo consideraremos en el ejemplo V4a como medición de V4 y todos los análisis de aquí en adelante obviarán V4b que, como ya vimos, no seguía una distribución aleatoria.

Como se observa en el cuadro 2.6, de utilizar el procedimiento de eliminación según lista, perderíamos 10 de los 30 casos, al quedar solo 20 en los que ninguna de las variables toma un valor perdido (desaparecen los casos: 1, 4,

Estadísticos e-Books & Papers

Cuadro 2.6.: Estadísticos descriptivos

	mean	sd	n
c1	22.15	1.6630663	20
v1	4.60	1.0954451	20
v2	4.25	0.7863975	20
v3	3.65	1.2680279	20
v4a	3.25	1.4464112	20
v5	4.55	0.7591547	20

7, 10, 16, 17, 18, 26, 28 y 29, compruebe el lector como ejercicio en el cuadro 2.2 el porqué, teniendo en cuenta, recordemos, que, al no considerar V4b, las dos únicas variables con valores perdidos son V4a y V2). En el cuadro 2.6 no aparecen C2 y C3 por ser variables no métricas y no tener sentido el cálculo de la media.

Una alternativa distinta de eliminación es la **eliminación de casos por parejas**. La filosofía es la misma, solo que se eliminan los casos únicamente en el supuesto en que contengan un valor perdido en las variables que se están utilizando en un análisis determinado. La limitación de este enfoque es que el tamaño muestral varía para cada uno de los análisis efectuados, provocando serios problemas en algunas técnicas. Tomando de nuevo el ejemplo de los modelos de ecuaciones estructurales, este procedimiento puede provocar (Byrne, 2001): (1) que la matriz de varianzas covarianzas no sea definida positiva impidiendo la convergencia, (2) los indicadores de bondad de ajuste basados en el estadístico chi cuadrado pueden estar sesgados. En cualquier caso, la aplicabilidad de este método está condicionada al supuesto de que los valores perdidos se distribuyen MCAR.

En el cuadro 2.7 se muestran las medias de las variables de nuestro ejemplo, resultantes si se aplicara el procedimiento de eliminación por parejas. Fijemos nuestra atención en la estimación de la media de V4a. Si calculamos su media cuando la variable, por ejemplo, V1 no toma valores perdidos, esta da 2,92, al igual que ocurre cuando V3, V5, C1, C2 y C3 no toman valores perdidos (lo que siempre ocurre, luego los casos considerados son los mismos). Sin embargo, la media calculada solo con los casos en que V2 no toma valores perdidos difiere sustancialmente (3,25), coincidiendo con la media que se obtenía en la eliminación según lista, pues estas dos variables eran las únicas que aportaban valores perdidos.

La alternativa a la eliminación es la **imputación**, es decir, sustituir el valor perdido por alguna estimación de su valor. Probablemente lo más habitual es sustituir el valor perdido por la media de la variable calculada con los casos disponibles. Uno de los atractivos de este procedimiento es que es conservador; la media de la distribución no cambia. Sin embargo, la varianza de la variable se reduce, pues, con casi toda seguridad, la media estará más próxima a sí misma

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 2.7.: Descriptivos en eliminación de casos por parejas

```
#Cuando V2 no toma valores perdidos

mean   n NA
v4a  3.250000 20  6
v1   4.653846 26  0
v2   4.153846 26  0
v3   3.692308 26  0
v5   4.461538 26  0
c1  21.884615 26  0

#Cuando V4a no toma valores perdidos

mean   n NA
v4a  2.916667 24  0
v1   4.625000 24  0
v2   4.250000 20  4
v3   3.791667 24  0
v5   4.500000 24  0
c1  22.333333 24  0

#Cuando resto variables no toma valores perdidos

mean   n NA
v4a  2.916667 24  6
v1   4.666667 30  0
v2   4.153846 26  4
v3   3.800000 30  0
v5   4.433333 30  0
c1  22.066667 30  0
```


Estadísticos e-Books & Papers

Cuadro 2.8.: Imputación por regresión

```
Call:
lm(formula = v4a ~ v1 + v3 + v5 + c1, data = Datos_2_1_Caso)

Residuals:
     Min       1Q   Median       3Q      Max
-2.22343 -0.71214  0.06567  0.81012  2.09377

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.01624    3.67031  -0.004   0.9965
v1           0.13737    0.25477   0.539   0.5960
v3          -0.37199    0.23381  -1.591   0.1281
v5           0.92648    0.37710   2.457   0.0238 *
c1          -0.02065    0.14240  -0.145   0.8862
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.229 on 19 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.4671,    Adjusted R-squared:  0.355
F-statistic: 4.164 on 4 and 19 DF,  p-value: 0.01375
```

que la de los valores perdidos que sustituye, lo que hará que las correlaciones con el resto de variables también caigan (Brown, 1994).

En nuestro ejemplo, los valores perdidos de V4a se sustituirían por 2,92, y los de V2, por 4,15 (como se desprende de los resultados del cuadro 2.7).

Un método alternativo de imputación es el de regresión. La variable cuyos valores perdidos se quiere estimar actúa como variable dependiente, mientras que el resto de variables actúan como independientes (normalmente con la condición de que sean métricas y no tengan a su vez valores perdidos). La recta de regresión obtenida, que ha sido estimada lógicamente solo utilizando los datos en los que la variable dependiente no tomaba valores perdidos, se utiliza para estimar esos valores, puesto que las variables que han actuado como independientes sí que son casos completos.

En nuestro ejemplo, podemos estimar los valores perdidos de V4a regresando los casos no perdidos de esta variable frente a V1, V3, V5 y C1 (no incluimos V2 por tener también valores perdidos ni C2 y C3 por ser variables no métricas. Podríamos transformarlas en ficticias y utilizarlas, pero el procedimiento estándar de sustitución por regresión en la mayoría de programas no lo contempla). Efectuando la regresión sobre los datos del cuadro 2.2, los resultados aparecen recogidos en el cuadro 2.8.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Por lo tanto, para estimar los valores perdidos de V4a utilizaríamos la función:

$$V4a = -0,016 + 0,137 \cdot V1 - 0,372 \cdot V3 + 0,926 \cdot V5 - 0,021 \cdot C1$$

por ejemplo, para el valor perdido del caso 1:

$$V4a = -0,016 + 0,137 \cdot 5 - 0,372 \cdot 4 + 0,926 \cdot 5 - 0,021 \cdot 21 = 3,38$$

En este caso hemos realizado la estimación y la imputación de manera manual para ilustrar el procedimiento, la mayoría de programas existentes lo realizan de forma automática.

Este procedimiento de imputación es mucho más razonable que la imputación simple de la media, por cuanto tiene en cuenta mucha más información. Sin embargo no está exento de limitaciones. Tabachnick y Fidell (2001) señalan las siguientes: (a) las estimaciones pueden ser "demasiado" coherentes con las variables utilizadas como independientes, pues al estimarse a partir de ellas serán más consistentes con las mismas que lo serían probablemente las contestaciones reales; (b) la varianza se reduce al estar las estimaciones muy próximas a las medias; (c) si en el contexto teórico las variables independientes no son regresores coherentes, la estimación de los valores perdidos puede ser en el fondo tan elemental como imputar una media, y (d) solamente se pueden utilizar los resultados de la estimación si esta cae en el rango permitido a la variable (en nuestro caso la variable V4 está acotada entre 1 y 5 y no podríamos imputar un valor fuera de ese rango).

Aunque está fuera del alcance de este manual, `R` proporciona varios paquetes específicos para la imputación de valores perdidos con procedimientos mucho más sofisticados de los expuestos. Por ejemplo, el paquete `MICE`[1] (Multivariate Imputation via Chained Equations) operativiza la imputación mediante regresión, solo que varía el tipo de regresión en función de la variable dependiente (*Predictive Mean Maching,* regresión logística, regresión politómica bayesiana). `AMELIA`[2] es un paquete que opta por el *bootstrapping* en el proceso de imputación.

2.3. *Outliers*: casos atípicos

Los *outliers* o casos atípicos son aquellos casos para los que una, dos o múltiples variables toman valores extremos que los hacen diferir del comportamiento del resto de la muestra y sospechar al investigador que han sido generados por mecanismos distintos al resto (Hawkins, 1980).

[1] `https://cran.r-project.org/web/packages/mice/mice.pdf`

[2] Llamado así en recuerdo a Amelia Earhart, la primera aviadora que cruzó el Atlántico y que desapareció (*missing*) misteriosamente, desaparición de valores que pretende resolver el paquete. `https://cran.r-project.org/web/packages/Amelia/Amelia.pdf`

Estadísticos e-Books & Papers

¿Por qué es importante detectar los valores atípicos? Fundamentalmente por sus consecuencias (Rasmussen, 1988; Schwager y Margolin, 1982; Zimmerman, 1994): (1) distorsionan los resultados al oscurecer el patrón de comportamiento del resto de casos y obtenerse conclusiones que, sin ellos, serían completamente distintas; (2) pueden afectar gravemente a una de las condiciones de aplicabilidad más habituales de la mayor parte de técnicas multivariantes, la normalidad.

Las **causas** que generan la existencia de valores atípicos en un fichero de datos pueden ser diversas. Anscombe (1960) clasifica estos valores atípicos en dos grandes grupos: los ocasionados por errores de los datos y los ocasionados por la inevitable y necesaria variabilidad de esos datos. De una manera más extensa, las causas pueden ser:

- Errores en los datos: tanto en su recogida como en la introducción de los mismos en la base de datos.
- Errores intencionados en la contestación al cuestionario por parte del entrevistado.
- Errores en el muestreo que se concretan en introducir en la muestra a individuos pertenecientes a una población distinta a la objetivo.
- *Outliers* legítimos, es decir, casos pertenecientes a la población objetivo que se quería muestrear pero que por la variabilidad inherente a las muestras difieren del resto en sus opiniones, actitudes o comportamientos.

Posteriormente veremos procedimientos para identificar univariante y multivariantemente esos valores atípicos, pero la pregunta es ¿qué hacer con ellos una vez identificados? Evidentemente la respuesta depende del tipo de *outlier*. Si corresponde a un error en la introducción de los datos, puede consultarse el cuestionario original y corregirlo. Si el error está en el registro y la encuesta no es anónima, puede optarse por reentrevistar. Si es anónima, una alternativa es la imputación del valor medio de la variable. Pero en todos estos casos hemos de estar seguros de que es un error de recogida o introducción y no una respuesta legítima.

Nadie discrepa de la conveniencia de eliminar los valores atípicos en caso de error evidente, sin embargo, el debate es mucho más intenso cuando se trata de qué hacer con un valor atípico legítimo. Algunos autores como Judd *et al.* (2009) consideran que la mejor alternativa es su **eliminación** para asegurar que las estimaciones son correctas para la mayoría de la población. Otros autores, sin embargo, creen que la eliminación es el último recurso y que se puede intentar suavizar su influencia **transformando** las variables mediante raíces cuadradas o logaritmos, lo que reduce su rango (Hamilton, 1992). Sin embargo esta solución puede no ser teóricamente razonable. Si la variable original es una escala cuyos valores tienen un sentido teórico para el investigador (una escala estandarizada, por ejemplo), su transformada puede no ser fácil de interpretar (Newton y Rudestam, 2013). Una última alternativa es intentar la utilización

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 2.9.: Descripción de la base de datos

Variable	Afirmación
SYS	Sueldos y salarios del directivo. Miles de euros
EDAD	Edad del directivo
EXP_PTO	Experiencia como alto directivo en cualquier empresa (años)
EXP_EMP	Experiencia en la empresa. Años
VENTAS	Ventas de la empresa. Millones de euros
BENEF	Beneficio de la empresa. Millones de euros

de técnicas de análisis estadístico que sean lo más **robustas** posible frente a los valores atípicos, como los contrastes estadísticos no paramétricos.

2.3.1. Detección univariante de casos atípicos

Planteada la importancia que pueden tener los *outliers* para la realización de un análisis estadístico veremos ahora distintas alternativas para su detección. Esta detección puede producirse desde una perspectiva univariante (analizando para una variable dada si algunos casos toman valores anormalmente altos) o multivariante (el vector de datos difiere del centroide). Ha de tenerse en cuenta que un caso puede no tomar valores anormales en dos variables consideradas individualmente, pero sí si se consideran conjuntamente. Un individuo de 14 años puede ser un elemento muestral lógico de nuestra investigación, un sujeto con un doctorado también, pero un sujeto de 14 años con un doctorado será, casi con toda seguridad, una caso atípico.

Caso 2.2. Retribución de altos directivos

Un investigador desea saber cuáles son las causas que explican la distinta remuneración de los altos directivos de las empresas. Dispone de los datos de 100 altos directivos que se sintetizan en el cuadro 2.9

Para ilustrar los distintos métodos de detección univariante nos centraremos en la variable que contempla la remuneración total del directivo (SYS). Se recomienda replicar los análisis que se desarrollan a partir de la base de datos suministrada.

El procedimiento más extendido consiste en considerar atípicos aquellos casos cuyo valor estandarizado de la variable analizada (z_i) supere un umbral determinado. Al estandarizar la variable x mediante la siguiente expresión:

$$z_i = \frac{(x_i - \bar{x})}{\sigma} \tag{2.1}$$

donde:

Estadísticos e-Books & Papers

Cuadro 2.10.: Observaciones atípicas para cada variable

Variable	Caso
SYS	14, 82, 88
EDAD	100
EXP_PTO	Ninguno
EXP_EMP	28, 50
VENTAS	2, 21, 42
BENEF	5, 14, 97

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{2.2}$$

siendo n el número de casos de nuestra base de datos. Por lo tanto, para establecer si un valor x_i determinado es atípico solo es necesario determinar el umbral que debe superar y que normalmente se expresa en número de desviaciones típicas k. Así, x_i será un valor atípico si su valor estandarizado z_i:

$$z_i \geq \bar{z} \pm k\sigma \tag{2.3}$$

Dado que son valores estandarizados, la media es 0 y la desviación típica 1, con lo que la expresión anterior se simplifica a $z_i \geq k$. Lo normal es considerar $k = 2,5$ para muestras pequeñas (menos de 80 casos) y $k = 3$ o $k = 4$ para muestras mayores (Hair *et al.*, 2014a). Si la distribución fuera normal, $k = 3$ implicaría que menos del $0,25\,\%$ de los casos pueden caer fuera de ese intervalo, por lo que no es ilógico considerarlo como atípico.

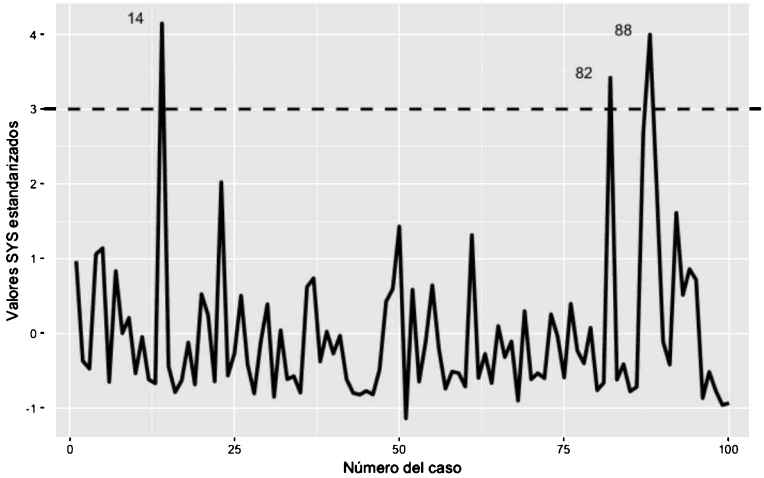
La figura 2.1 muestra la distribución de la variable SYS estandarizada (ZSYS) tal y como se ha indicado. Puede comprobarse que tres casos (14, 82 y 88) superan en más de 3 desviaciones típicas el valor medio. Si repitiendo el análisis para todas las variables esos casos tomasen también valores atípicos, deberíamos comenzar a pensar que probablemente lo sean de manera multivariante. Como se comprueba en el cuadro 2.10, donde se muestran los casos que, aplicando este procedimiento, resultarían univariantemente atípicos, no se observa reincidencia de casos. De momento, por tanto, el investigador debería comprobar si estamos ante errores en la toma o introducción de datos o si, por el contrario, son *outliers* legítimos.

Un procedimiento menos utilizado, pero muy recomendado por algunos autores (Pérez, 2004), para la detección univariante de *outliers* es el **test de Grubbs** (Grubbs, 1969; Stefanksy, 1971), basado también en la asunción de normalidad de la distribución. Partiendo de la hipótesis nula de que no hay caso atípico entre los datos, se calcula el estadístico G como sigue:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 2.1.: Valores de la variable SYS estandarizados (ZSYS)



$$G = \frac{\text{máx}\,|x_i - \bar{x}|}{\sigma} \tag{2.4}$$

donde todas las variables han sido definidas. La hipótesis nula de ausencia de valores atípicos se rechaza si:

$$G > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{(\alpha/2n,n-2)}}{n-2+t^2_{(\alpha/2n,n-2)}}} \tag{2.5}$$

donde $t^2_{(\alpha/2n,n-2)}$ es el valor crítico de la distribución t con $n-2$ grados de libertad y un nivel de significatividad de $\alpha/2n$.

El cuadro 2.11 recoge algunos de los casos de la base de datos para ilustrar el proceso de cálculo del test de Grubbs. Las dos primeras columnas son, directamente, datos del fichero y la tercera es un cálculo elemental que requiere únicamente del cálculo previo de la media, que figura al pie de la tabla.

De acuerdo con 2.4 y los datos del cuadro 2.11 el estadístico G del test de Grubbs sería:

$$G = \frac{3535,33}{852,72} = 4,15$$

De la expresión 2.5 el único cálculo no inmediato corresponde con el valor de la distribución t. Es obvio que resulta complicado encontrar tablas estadísticas para cualquier número de grados de libertad y, aunque existieran, sería demasiado laboriosa su consulta. Por ese motivo creemos que lo más sencillo es recurrir a las macros ya programadas de las hojas de cálculo. En nuestro caso

Estadísticos e-Books & Papers

Cuadro 2.11.: Observaciones atípicas para cada variable

Caso	SYS	$\lvert x_i - \bar{x}\rvert$
1	1.948	826,33
2	809	312,67
3	721	400,67
4	2.027	905,33
5	2.094	972,33
6	570	551,67
7	1.833	711,33
8	1.126	4,32
9	1.300	178,33
10	668	453,67
11	1.082	39,67
12	597	524,67
13	554	567,67
14	4.657	3.535,33
$\vdots$	$\vdots$	$\vdots$
97	684	437,67
98	466	655,67
99	307	814,67
100	329	792,67
$\bar{x} = 1121,67$		
$\text{máx}\,\lvert x_i - \bar{x}\rvert = 3535,33$		
$\sigma = 852,72$		

© Ediciones Paraninfo

Estadísticos e-Books & Papers

se han utilizado las funciones implementadas en R, concretamente la expresión `qt(a,b,lower.tail=c)`, donde a es el nivel de significación, b es el número de grados de libertad y c es una opción que señala si estamos optando por una (`lower.tail=TRUE`) o dos colas de la distribución (`lower.tail=FALSE`). En nuestro caso concreto, la expresión sería:

```
qt((0.05/200),98,lower.tail=FALSE)=3,6008
```

Nótese que hemos decidido trabajar a un nivel de significación del 5 %. Sustituyendo en 2.5:

$$G > \frac{n-1}{\sqrt{n}}\sqrt{\frac{t^2_{(\alpha/2n,n-2)}}{n-2+t^2_{(\alpha/2n,n-2)}}} = \frac{100-1}{\sqrt{100}}\sqrt{\frac{3,6008^2}{100-2+3,6008^2}} = 3,384$$

Como $4,15 > 3,38$, rechazamos la hipótesis nula de ausencia de *outliers,* con lo que la observación 14, a la que se corresponde el mayor $|x_i - \bar{x}|$, sería un caso atípico. El proceso se repetiría para el siguiente (caso 88, como se comprueba en el gráfico 2.1) hasta que el estadístico dejara de ser significativo, con lo que habríamos aislado todos los valores atípicos univariantes de la variable SYS.

2.3.2. Detección bivariante de casos atípicos

La detección bivariante tiene utilidad cuando, con posterioridad, vamos a realizar algún análisis en el cual una variable vaya a actuar como dependiente pues, de esta forma, podemos realizar una inspección de la relación que cada independiente guardará previsiblemente con ella.

El proceso es sencillo, basta con obtener un gráfico de dispersión y realizar una regresión simple. Esto nos permitirá superponer al gráfico las bandas de un intervalo de predicción individual (por ejemplo al 95 %) y ver qué casos quedan fuera. Si sistemáticamente son los mismos, querrá decir que sus valores de variable dependiente son anormales aunque se contemplen con cualquier variable explicativa.

En el ejemplo del caso 2.2 que estamos siguiendo, podemos regresar la variable de ingresos SYS sobre, sucesivamente, la edad (EDAD), la experiencia (EXP_PTO, EXP_EMP), las ventas (VENTAS) y los beneficios de la empresa (BENEF). Los gráficos 2.2, 2.3 y 2.4 ilustran algunas de estas regresiones.

Parece que las remuneraciones de los casos 14, 82, 87, 88 y 89 son casos atípicos bivariantes, pues sistemáticamente aparecen como tales.

2.3.3. Detección multivariante de casos atípicos

Igual que en muchas ocasiones las técnicas que se aplican no exigen normalidad univariante, sino multivariante (caso de la regresión lineal múltiple, por ejemplo), respecto a los casos atípicos ocurre lo mismo. No es tan problemático

Estadísticos e-Books & Papers

Figura 2.2.: Relación ingresos-edad

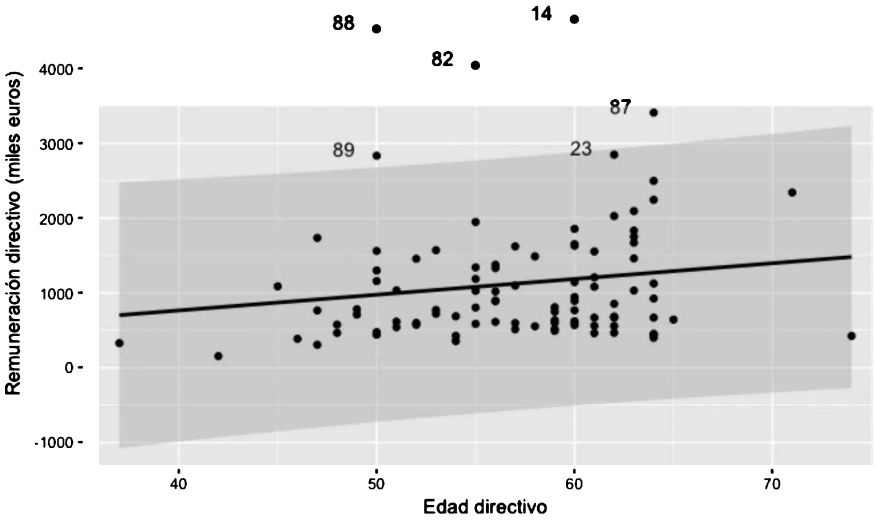
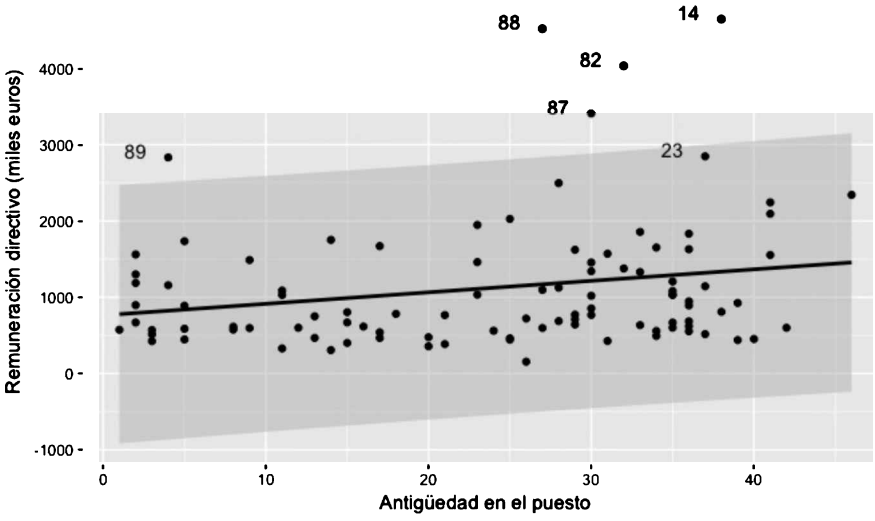


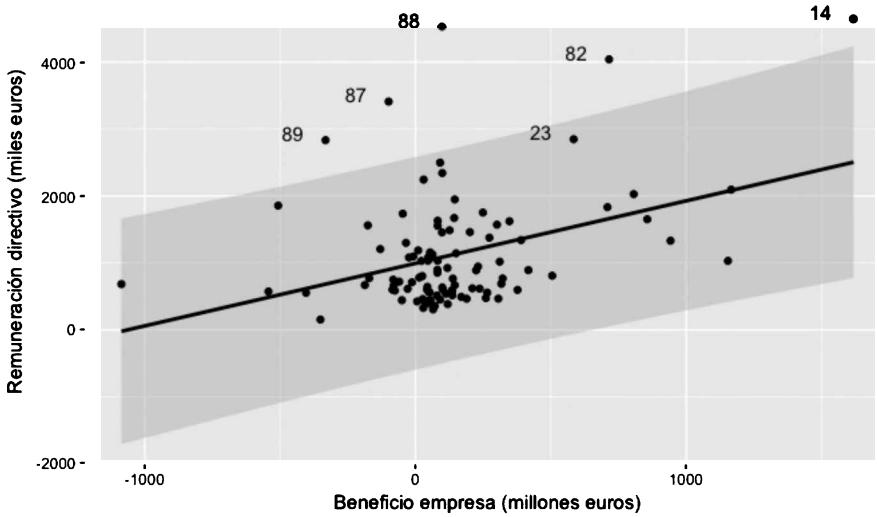
Figura 2.3.: Relación ingresos-antigüedad en el puesto



© Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 2.4.: Relación ingresos-beneficio de la empresa



que un caso sea atípico respecto a una variable sino respecto al conjunto de las que se incorporan al análisis. Ello hace necesario buscar un procedimiento que contemple simultáneamente todas las variables para determinar si algún caso tiene un comportamiento anómalo.

Un procedimiento bastante intuitivo consiste en calcular la distancia de cada caso al centroide del conjunto de los datos. Cuanto más lejos esté un caso de la media de los casos más probable es que sea un caso atípico. La distancia habitualmente empleada en la detección de casos atípicos es la distancia de Mahalanobis (D). Ilustraremos su cálculo con un ejemplo sencillo (caso 2.3) y luego lo aplicaremos al caso 2.2.

Caso 2.3. Motores de distintos modelos de automóviles

Sea $\mathbf{X}$ la matriz de datos formada por n casos y m variables donde quiere detectarse la existencia de casos atípicos. En nuestro ejemplo la matriz $\mathbf{X}$ aparece en el cuadro 2.12, siendo $n = 10$ y $m = 5$. Son datos de 10 modelos de automóviles de los cuales se da su consumo (l/100km), cilindrada (cc), potencia (cv), peso (kg) y aceleración (segundos de 0 a 100 km/h).

Deseamos saber cuánto dista cada caso del centroide de los datos, es decir, del vector de medias de las variables implicadas que denotaremos como $\bar{\mathbf{X}}$ que se transforma en una matriz $n \times m$ repitiendo el vector de medias $1 \times n$ en las m filas de esa matriz para que la resta con $\mathbf{X}$ sea compatible.

Pues bien, el cuadrado de la distancia de Mahalanobis $\left(\mathbf{D^2}\right)$ para cada caso se obtiene como sigue:

$$\mathbf{D}^2 = diag\left\{\left[\mathbf{X} - \bar{\mathbf{X}}\right]\mathbf{S}^{-1}\left[\mathbf{X} - \bar{\mathbf{X}}\right]'\right\} \tag{2.6}$$

Estadísticos e-Books & Papers

Cuadro 2.12.: Matriz X de datos para la detección de *outliers*

Caso	Consumo	Motor	CV	Peso	Aceleración
1	13	5.031	130	1.168	12
2	16	5.735	165	1.231	12
3	13	5.211	150	1.145	11
4	15	4.982	150	1.144	12
5	14	4.949	140	1.149	11
6	16	7.030	198	1.447	10
7	17	7.440	220	1.451	9
8	17	7.210	215	1.437	9
9	17	7.456	225	1.475	10
10	16	6.391	190	1.283	9
Media	15,4	6.144	178,3	1.293	9,5

donde toda la notación es conocida salvo $\mathbf{S}^{-1}$, que es la inversa de la matriz de varianzas covarianzas de $\mathbf{X}$ y que, como es conocido, se obtiene:

$$S = \frac{1}{n-1}\left[\mathbf{X} - \bar{\mathbf{X}}\right]\left[\mathbf{X} - \bar{\mathbf{X}}\right]' \tag{2.7}$$

Con los datos de nuestro ejemplo y con operaciones muy fácilmente implementables mediante cualquier hoja de cálculo, podemos obtener las matrices intermedias:

$$\mathbf{X} = \begin{bmatrix} 13 & 5031 & 130 & 1168 & 12 \\ 16 & 5735 & 165 & 1231 & 12 \\ 13 & 5211 & 150 & 1145 & 11 \\ 15 & 4982 & 150 & 1144 & 12 \\ 14 & 4949 & 140 & 1449 & 11 \\ 16 & 7030 & 198 & 1447 & 10 \\ 17 & 7440 & 220 & 1451 & 9 \\ 17 & 7210 & 215 & 1437 & 9 \\ 17 & 7456 & 225 & 1475 & 10 \\ 6 & 6391 & 190 & 1238 & 9 \end{bmatrix}$$

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

$$\overline{\mathbf{X}} = \begin{bmatrix} 15,4 & 6144 & 178,3 & 1293 & 10,5 \\ 15,4 & 6144 & 178,3 & 1293 & 10,5 \\ 15,4 & 6144 & 178,3 & 1293 & 10,5 \\ 15,4 & 6144 & 178,3 & 1293 & 10,5 \\ 15,4 & 6144 & 178,3 & 1293 & 10,5 \\ 15,4 & 6144 & 178,3 & 1293 & 10,5 \\ 15,4 & 6144 & 178,3 & 1293 & 10,5 \\ 15,4 & 6144 & 178,3 & 1293 & 10,5 \\ 15,4 & 6144 & 178,3 & 1293 & 10,5 \\ 15,4 & 6144 & 178,3 & 1293 & 10,5 \end{bmatrix}$$

$$\mathbf{X} - \overline{\mathbf{X}} = \begin{bmatrix} -2,4 & -1113 & -48,3 & -125 & 1,5 \\ 0,6 & -409 & -13,3 & -62 & 1,5 \\ -2,4 & -933 & -28,3 & -148 & 0,5 \\ -0,4 & -1162 & -28,3 & -149 & 1,5 \\ -1,4 & -1195 & -38,3 & -144 & 0,5 \\ 0,6 & 886 & 19,7 & 154 & -0,5 \\ 1,6 & 1296 & 41,7 & 158 & -1,5 \\ 1,6 & 1066 & 36,7 & 144 & -1,5 \\ 1,6 & 1312 & 46,7 & 182 & -0,5 \\ 0,6 & 247 & 11,7 & -10 & -1,5 \end{bmatrix}$$

$$\mathbf{S^{-1}} = \begin{bmatrix} 3,102 & 0,004 & -0,259 & -0,007 & -0,963 \\ 0,004 & 0,000 & -0,002 & -0,001 & 0,009 \\ -0,259 & -0,002 & 0,058 & 0,006 & 0,047 \\ -0,007 & -0,001 & 0,006 & 0,004 & -0,046 \\ -0,963 & 0,009 & 0,047 & -0,046 & 2,945 \end{bmatrix}$$

Efectuando las operaciones especificadas en 2.6, se obtiene la D^2 de Mahalanobis:

$$D^2 = \begin{bmatrix} 4,94 \\ 6,45 \\ 6,43 \\ 4,99 \\ 4,23 \\ 3,80 \\ 2,00 \\ 2,15 \\ 4,90 \\ 5,11 \end{bmatrix}$$

Para determinar si alguno de los 10 casos especificados es un *outlier*, dado que la D^2 de Mahalanobis se distribuye como una χ^2 con tantos grados de libertad como variables implicadas –5 en nuestro caso–, bajo la hipótesis nula de que el caso i no es un caso atípico. Para calcular la significatividad o los

Estadísticos e-Books & Papers

Cuadro 2.13.: Resultado del test de Mahalanobis

Caso	D^2	D crítico (99 %, gl=5)	Sig
1	4,94	15,09	0,42
2	6,45	15,09	0,26
3	6,43	15,09	0,27
4	4,99	15,09	0,42
5	4,23	15,09	0,52
6	3,80	15,09	0,58
7	2,00	15,09	0,85
8	2,15	15,09	0,83
9	4,90	15,09	0,43
10	5,11	15,09	0,40

valores críticos puede recurrirse, de nuevo, a una hoja de cálculo, pero todo el proceso es fácilmente implementable en R, tal y como muestra la figura 2.5.

El cuadro 2.13 resume los resultados de la ejecución de la sintaxis de la figura 2.5. De acuerdo con él, ningún caso tiene una distancia de Mahalanobis que supere el valor crítico y no existirían *outliers*. Diversos autores (Hair *et al.*, 2014a; Tabachnick y Fidell, 2001) señalan que hay que ser muy prudentes a la hora de clasificar un caso como *outlier*, recomendando clasificarlo como tal solo cuando la significatividad del test sea $p < 0,001$. En esta misma línea, autores como Field (2005), tras manifestar la dificultad en establecer un nivel de corte para que la D^2 de Mahalanobis permita clasificar un caso como atípico, recomiendan seguir las tablas proporcionadas por Barnett y Lewis (1994), en las que, a grandes rasgos para muestras grandes ($N = 500$) y cinco variables, valores de la D^2 superiores a 25 deben considerarse como *outliers*, 15 para muestras más pequeñas ($N = 100$) y alrededor de 11 para muestras muy reducidas ($N = 30$). Únicamente por facilidad de análisis, dado que los valores críticos de una distribución t son más conocidos, autores como Hair *et al.* (2014a) suelen utilizar la ratio D^2/gl porque sigue aproximadamente la distribución de una t.

Si aplicamos ahora el procedimiento descrito a los datos del caso 2.2, observamos en la figura 2.6 que cabría considerar atípico los casos 14, 42, 88 y 97 siguiendo la recomendación de utilizar una $p < 0,001$, mientras que si se es un poco más laxo, con una $p < 0,01$, a los anteriores se añadiría el caso 44.

Suponiendo que asumiéramos $p < 0,001$ y consideráramos los casos, por ejemplo, 14, 42, 88 y 97 como atípicos, dado que tenemos 100 casos en la base de datos no sería ningún problema su eliminación, sin embargo, antes de hacerlo, deberían analizarse los motivos que causan que lo sean y ver cómo puede afectar a la generalizabilidad de los resultados su eliminación. Un procedimiento sencillo es comparar los valores que toman los casos atípicos (por ejemplo, el caso 14 y el 97) con las medias del resto de casos.

Comprobamos en el gráfico 2.7 cómo los casos 14 y 97 son atípicos por razones

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

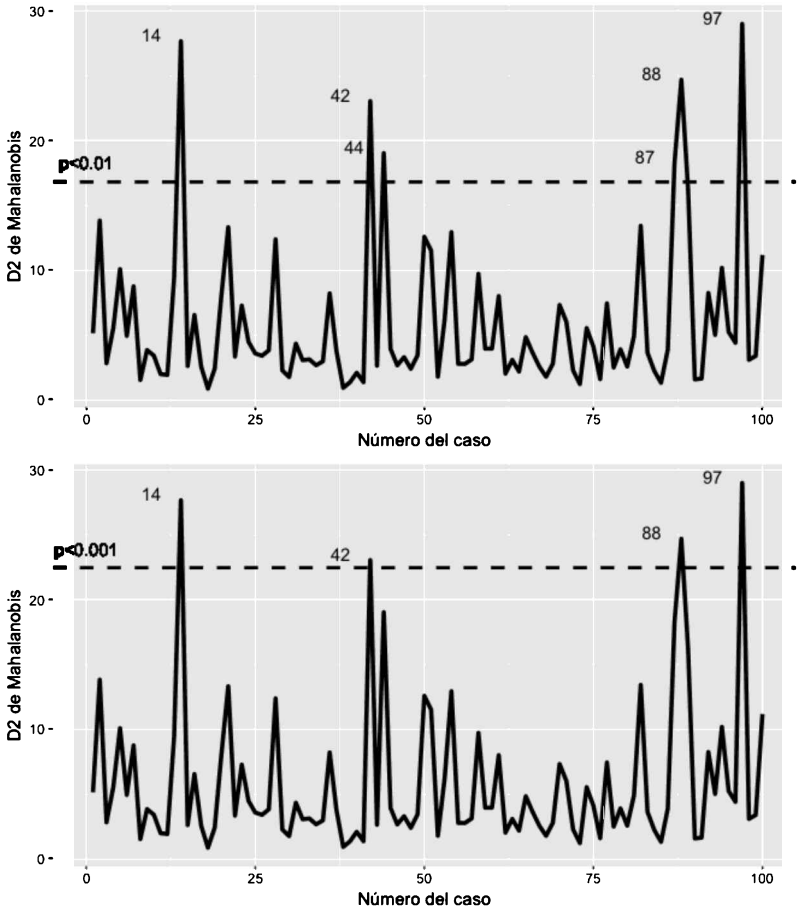
Figura 2.5.: Sintaxis para la obtención de la D^2 de Mahalanobis y su significatividad en R

```
mean<-colMeans(Datos_2_3_Caso)                          #Cálculo del vector de medias de las variables
Sx<-cov(Datos_2_3_Caso)                                 #Cálculo de la matriz de covarianzas
D2<-mahalanobis(Datos_2_3_Caso,mean,Sx,inverted = FALSE) #Cálculo de la D2 de Mahalanobis
                                                        #Como le hemos dado la matriz de covarianzas Sx
                                                        #actúa por defecto el modificador inverted=FALSE
                                                        #si le diéramos Sx ya invertida: inverted=TRUE
print(D2)                                               #Muestra los valores de la D2
pchisq(D2, df=5, lower.tail=FALSE)                      #df=5 porque hay 5 variables. Muestra la signficatividad
qchisq(.99, df=5)                                       #Calcula el valor crítico
```

© Ediciones Paraninfo

Estadísticos e-Books & Papers

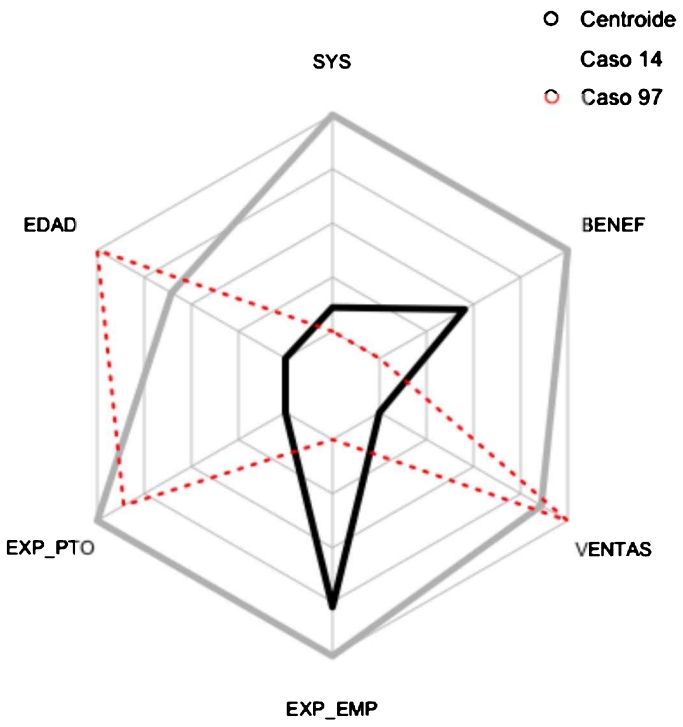
Figura 2.6.: Casos atípicos de acuerdo con la D^2 de Mahalanobis



© Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 2.7.: Casos atípicos de acuerdo con la D^2 de Mahalanobis



Estadísticos e-Books & Papers

distintas. El caso 14 corresponde a un directivo que supera los promedios que marca el centroide en todas las variables, es decir, tiene más edad y experiencia y también está en una empresa que le paga un sueldo superior porque también tiene mayores ventas y beneficios.

Por su lado, el caso 97 corresponde a un directivo que, pese a que tiene más edad que el promedio y lleva tiempo trabajando en el sector, se acaba de incorporar a la empresa que, pese a tener una cifra de ventas elevada, le paga un salario inferior al promedio por la juventud en el puesto.

2.4. Comprobación de los supuestos básicos del análisis multivariante

Como se verá en cada uno de los capítulos correspondientes, cada técnica multivariante está basada en una serie de supuestos estadísticos. Algunas de ellas son muy exigentes, por ejemplo, la regresión múltiple asume la normalidad multivariante de los datos, la homocedasticidad, linealidad e independencia de las observaciones. Otras técnicas, sin embargo, son mucho menos exigentes respecto a estas propiedades. Así, el análisis de conglomerados, por ser una técnica algorítmica, no exige propiedades estadísticas a sus datos.

En cada capítulo se abordarán las exigencias de cada técnica. En este apartado, dada la importancia que tiene la comprobación de estas propiedades, se abordará una primera aproximación al contraste de las más importantes, a saber, normalidad univariante y multivariante, la homocedasticidad, la linealidad y la independencia de las observaciones.

2.4.1. Normalidad

Sharma (1996) plantea el siguiente razonamiento para justificar la relevancia de comprobar la normalidad univariante y multivariante de las variables implicadas en el análisis. Cuando se pretende contrastar una hipótesis nula se pueden cometer dos tipos de errores. El error tipo I (al que normalmente nos referimos como α o significatividad del estadístico utilizado para el contraste) es la probabilidad de equivocarnos al rechazar la hipótesis nula. El investigador normalmente elige un nivel estándar para este error (digamos 0,05), lo que implica que, si se repitiera infinidad de veces el estudio, nos equivocaríamos un 5 % de las veces al rechazar la hipótesis nula. Sin embargo, si se está violando alguno de los supuestos del modelo, por ejemplo la normalidad, el número de veces que nos estaríamos equivocando al rechazar la hipótesis nula sería superior a ese 5 % teórico.

El error tipo II (β) es la probabilidad de no rechazar la hipótesis nula cuando esta es, de hecho, falsa. Se define la potencia de un test como $1 - \beta$, que es la probabilidad de acertar al rechazar la hipótesis nula cuando esta es falsa. Cuanto más baja es la potencia de un test, más se reduce la posibilidad de encontrar resultados significativos.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Obviamente un investigador deseará tener valores α pequeños y test potentes. Sin embargo, ambas variables pueden verse afectadas por la violación de las hipótesis subyacentes, entre ellas, la normalidad.

Como señala Sharma (1996), diversos autores (Everitt, 1979; Glass *et al.*, 1972b; Hopkins y Clay, 1963; Olson, 1974) han constatado que la violación de la hipótesis de normalidad no tiene un efecto apreciable sobre el error tipo I, sin embargo sí que lo tiene, e importante, sobre el error tipo II, de ahí la importancia de su comprobación.

Aunque las técnicas que analizaremos posteriormente suelen exigir normalidad multivariante, veremos a continuación cómo contrastar primero la univariante por varias razones: (1) los test multivariantes son más complejos y didácticamente se entienden mejor viendo primero los univariantes (Sharma, 1996); (2) aunque es teóricamente posible que siendo todas las variables univariantemente normales, no lo sean multivariantemente, es bastante improbable, por lo que es difícil que la no normalidad multivariante no sea detectada a través de la no normalidad univariante (Gnanadesikan, 1977), y (3) si la distribución no es multivariantemente normal, entonces hay que indagar qué variables están causando este problema y para ello es necesario conocer los tests univariantes.

A. Análisis univariante de la normalidad

El primer paso es analizar la **asimetría** y la **curtosis** (apuntamiento) de las distribuciones de cada variable. Estos son dos componentes de la normalidad. Cuando una distribución es normal, los valores de asimetría y curtosis son cero (realmente la distribución normal de una curtosis es 3, pero todos los paquetes estadísticos restan este valor para tomar 0 como referencia). La figura 2.8 ilustra lo que indican valores positivos y negativos de asimetría y curtosis.

Para ilustrar los distintos cálculos que realizaremos utilizaremos los datos del sencillo caso 2.3, para después, una vez desarrolladas todas las técnicas de comprobación de las hipótesis, aplicarlos a un caso más complejo.

Para contrastar si el coeficiente de asimetría (A) o el de curtosis (C) es o no estadísticamente distinto de cero, se calcula la probabilidad de que sus valores estandarizados Z_A y Z_C que se distribuyen según una $N(0,1)$ difieran de la normal para un nivel de significación preestablecido (que para muestras pequeñas, como en nuestro ejemplo, se considera que sea un nivel conservador: 0,01 o 0,001, como señalan Tabachnick y Fidell, 2001).

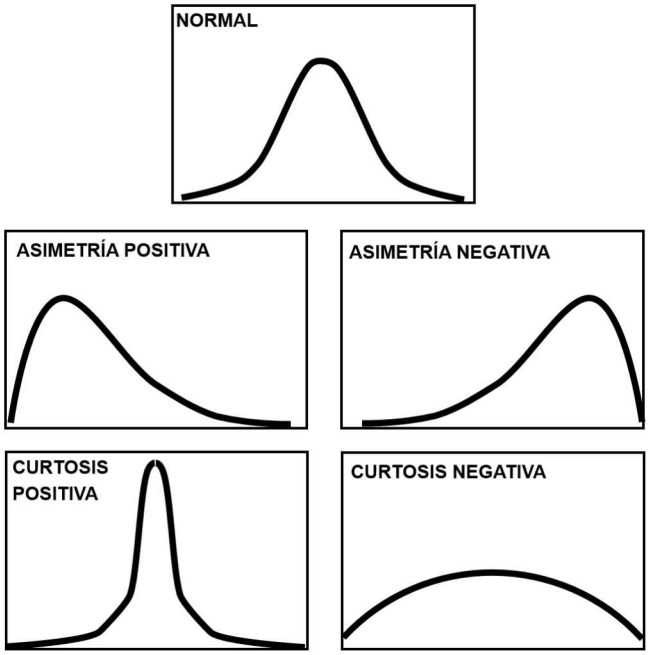
Calculamos los valores estandarizados restando la media (0) y dividiendo por la desviación típica del estadístico:

$$Z_A = \frac{A-0}{SE_A} \; Z_c = \frac{C-0}{SE_C} \tag{2.8}$$

siendo SE_A y SE_C los errores estándar de los coeficientes de asimetría y curtosis. Un valor absoluto de estos estadísticos normalizados superiores a 1,96 —digamos 2— mostraría una desviación significativa de la normal para $p < 0,05$. Si estos valores normalizados fueran superiores a 2,58 lo serían para $p < 0,01$,

Estadísticos e-Books & Papers

Figura 2.8.: Distribuciones normal, asimétricas, platicúrticas y leptocúrticas



y superiores a 3,29, para $p < 0,001$. Como señalan Field *et al.* (2012), en muestras grandes será habitual tener errores estándar pequeños, siendo significativas incluso pequeñas desviaciones de la normalidad, recomendando para muestras superiores a 200 casos observar directamente la forma de la distribución antes que recurrir a estos valores normalizados de asimetría y apuntamiento.

El cuadro 2.14 muestra los resultados del análisis descriptivo de los datos del caso 2.3 mediante la función `stat.desc()` de `R`.

```
round(stat.desc(Datos_2_3_Caso
[,c("consumo","motor","cv","peso","acel")],
basic=FALSE,norm=TRUE),digits=3)
```

Quizás la forma más sencilla de obtener una interpretación sea fijarse en los valores que aparecen etiquetados como `kurt.2SE` y `skew.2SE`. Habíamos señalado que valores del estadístico normalizado superiores a 2 implicarían una desviación significativa de la normal. `R` lo que nos ofrece son esos valores normalizados divididos por 2, por lo tanto, valores superiores a 1 implicarán una desviación de la normal significativa. Podemos comprobar en la salida que no se aprecian, con este criterio, desviaciones significativas de la normalidad.

Otra alternativa para establecer la normalidad univariante es el recurso a los **gráficos Q-Q**. Aunque todos los paquetes estadísticos los calculan auto-

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 2.14.: Resultado del test de Mahalanobis

	consumo	motor	cv	peso	acel
median	16.000	6063.000	177.500	1257.000	10.500
mean	15.400	6143.500	178.300	1293.000	10.500
SE.mean	0.499	340.557	11.240	45.577	0.401
CI.mean.0.95	1.129	770.394	25.426	103.101	0.908
var	2.489	1159791.833	1263.344	20772.222	1.611
std.dev	1.578	1076.936	35.544	144.126	1.269
coef.var	0.102	0.175	0.199	0.111	0.121
skewness	-0.446	0.067	0.037	0.163	0.000
skew.2SE	-0.325	0.049	0.027	0.119	0.000
kurtosis	-1.543	-1.968	-1.830	-1.986	-1.820
kurt.2SE	-0.578	-0.738	-0.686	-0.744	-0.682
normtest.W	0.855	0.841	0.909	0.812	0.852
normtest.p	0.067	0.046	0.273	0.020	0.061

máticamente, ilustraremos su cálculo con el ejemplo que venimos utilizando, concretamente para la variable peso. El gráfico Q-Q se obtiene como sigue:

1. Se ordenan de menor a mayor los n valores de la variable analizada (peso del coche, en nuestro ejemplo). Como la variable es continua y es improbable que se repitan valores, cabe esperar que haya j observaciones iguales o inferiores al valor X_j de la variable considerada (columna 2 del cuadro 2.15).

2. La proporción de observaciones que son inferiores a X_j se estiman mediante la expresión $(j - 0,5)/n$, donde el término $0,5$ es un factor de corrección, puesto que lo lógico sería j/n. Johnson y Wichern (1998) indican que otros autores como Filliben (1975) o Looney y Gulledge (1985) sugieren la correción $(j - 3/8)/(n + 1/4)$. Esto aparece recogido en la columna 3 del cuadro 2.15.

3. Los valores anteriores se asume que son los percentiles o niveles de probabilidad de la función de distribución normal acumulativa estandarizada si los datos muestrales siguen una normal. Si se calculan los niveles teóricos directamente de tablas (nosotros los hemos calculado mediante la función de R `qnorm(seq(from=.05,to=.95,by=.1))`) y se representan frente a los valores observados Xj, la relación entre ambos debería ser lineal si los datos muestrales proceden de una normal (gráfico 2.7).

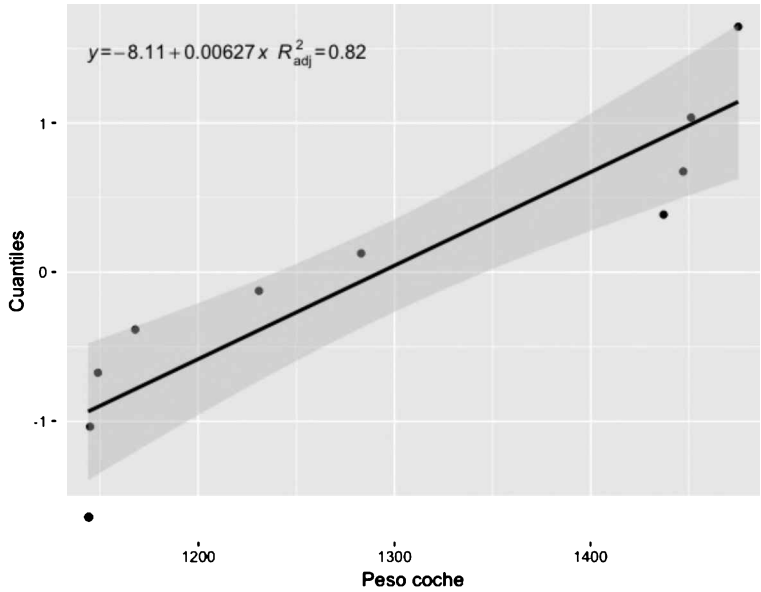
Como se observa en la figura 2.9, la relación lineal no es evidente. En cualquier caso, los gráficos Q-Q no son especialmente útiles salvo para muestras de cierto tamaño, $n < 20$ (Johnson y Wichern, 1998), y su interpretación es muy subjetiva (Sharma, 1996). Por ese motivo este último autor recomienda objetivarlos mediante el siguiente procedimiento: calcular el coeficiente de correlación

Estadísticos e-Books & Papers

Cuadro 2.15.: Cálculos necesarios para el gráfico Q-Q

Observación j	Peso X_j	Nivel de probabilidad $(j-1/2)/n$	Cuantiles normal estandarizada
1	1.475	0,05	-1,64
2	1.451	0,15	-1,04
3	1.447	0,25	-0,67
4	1.437	0,35	-0,39
5	1.283	0,45	-0,13
6	1.231	0,55	0,13
7	1.168	0,65	0,39
8	1.149	0,75	0,67
9	1.145	0,85	1,04
10	1.144	0,95	1,64

Figura 2.9.: Gráfico Q-Q para la variable peso



© Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 2.16.: Valores críticos para el coeficiente de correlación del gráfico Q-Q bajo el supuesto de normalidad

Tamaño muestral n	Significatividad α		
	.01	.05	.10
5	.8299	.8788	.9032
10	.8801	.9198	.9351
15	.9126	.9389	.9503
20	.9269	.9508	.9604
25	.9410	.9591	.9665
30	.9479	.9652	.9715
35	.9538	.9682	.9740
40	.9599	.9726	.9771
45	.9632	.9794	.9792
50	.9671	.9768	.9809
55	.9695	.9787	.9822
60	.9720	.9801	.9836
75	.9771	.9838	.9866
100	.9822	.9873	.9895
150	.9879	.9913	.9928
200	.9905	.9931	.9942
300	.9935	.9953	.9960

Fuente: Johnson y Wichern (1998, p.193)

entre la muestra (X_j) y los cuantiles de la normal estandarizada y compararlos con los valores críticos obtenidos por Filliben (1975) o los proporcionados por Johnson y Wichern (1998), que se reproducen en el cuadro 2.16. Si calculamos ese coeficiente de correlación, el valor es $0,914$. Si se compara con el valor crítico para $n = 10$, se constata la aceptación de la hipótesis de normalidad para $p < 0,01$ (coeficiente superior al valor crítico).

Más allá de los procedimientos gráficos, y dada la mencionada subjetividad de estos, se han desarrollado otros procedimientos basados en el desarrollo de tests específicos para el contraste de hipótesis.

Los más habituales son el test de **Shapiro-Wilk** (Shapiro y Wilk, 1965), el test de **Kolmogorov-Smirnov** (Chakravarti *et al.*, 1967), el test de **Anderson-Darling** (Stephens, 1974) y el test de **Cramer-von Mises** (Marsaglia y Marsaglia, 2004). Su desarrollo teórico va más allá del alcance de este libro pero, en general, todos ellos plantean la hipótesis nula de que la variable sigue una distribución normal. Es muy importante señalar que, mientras que el test de Shapiro-Wilk es recomendable para muestras inferiores a 2.000 casos, el resto exige muestras superiores a esta cifra.

La ilustración de los resultados sobre los datos del caso 2.3 es, debido a la exigencia de tamaño muestral, meramente ilustrativa para todos los test,

Estadísticos e-Books & Papers

Cuadro 2.17.: Test de contraste de la normalidad univariante

```
        Shapiro-Wilk normality test

data:  Datos_2_3_Caso$peso
W = 0.81173, p-value = 0.02013


        Lilliefors (Kolmogorov-Smirnov) normality test

data:  Datos_2_3_Caso$peso
D = 0.24113, p-value = 0.1018
```

salvo para el ya mencionado de Shapiro-Wilk, que funciona bien para tamaños comprendidos entre 7 y 2.000 casos. Distintos ejercicios de simulación (Wilk *et al.*, 1968) han demostrado que, además, es el más potente en la determinación de la normalidad univariante. Este test compara los datos de la variable analizada con los de una distribución normal con las mismas media y desviación típica. Cuando el test no es significativo $p > 0{,}05$, la muestra no es estadísticamente distinta de la de la distribución normal con la que se compara.

Todos estos test se pueden calcular mediante `R`. El test de Shaphiro-Wilk mediante la función `shapiro.test` del paquete `{stats}`, el test de Kolmogorov con la corrección de Lilliefors mediante `lillie.test` del paquete `{nortest}`, y el paquete `{goftest}` ofrece los test de Anderson-Darling (`ad.test`) y de Cramer-von Mises (`cvm.test`). El cuadro 2.17 recoge la salida de `R`, donde se muestran los dos primeros, dado que los dos segundos no tienen sentido, como se ha señalado, para el tamaño muestral del ejemplo que estamos siguiendo. A la luz del mismo se puede comprobar como, para la variable analizada "peso", no seguiría una distribución normal para $p < 0,05$ según el test de Shapiro-Wilk (el único válido dado el tamaño muestral), mientras que sí que la seguiría según el test de Kolmogorov.

B. Análisis multivariante de la normalidad

Como señala Sharma (1996), existen muy pocos tests para el contraste de la normalidad multivariante. El método gráfico es similar al utilizado para la normalidad univariante (**gráfico ji-cuadrado**), mientras que los tests de **Mardia-curtosis** y **Mardia-apuntamiento** (Mardia, 1980) y el test de **Henze-Zirlke**r (Henze y Zirkler, 1990) o Royston (Royston, 1982) están operativizados en muy pocos paquetes estadísticos y, como indica también Sharma (1996), su distribución no es muy bien conocida, lo que les confiere una utilidad limitada. Ilustraremos inicialmente, por ello, el procedimiento de elaboración del gráfico ji-cuadrado para luego ofrecer los paquetes de `R` que pueden calcular los es-

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

tadísticos reseñados con las limitaciones reiteradas del reducido tamaño de la muestra del caso.

El gráfico ji-cuadrado se construye de una manera muy similar al gráfico Q-Q. Siguiendo a Johnson y Wichern (1998), los pasos que ilustraremos con los datos del caso 2.3 serían los siguientes:

1. Se calculan las distancias de Mahalanobis para todas las variables cuya normalidad multivariante se desee contrastar. En nuestro caso son las variables recogidas en el cuadro 2.12: consumo, motor, potencia, peso y aceleración. El procedimiento de cálculo de esta distancia ya se explicó en el epígrafe 2.3.3 y los resultados de esta distancia se recogen en el cuadro 2.13 (nótese que en dicho cuadro se muestran los valores ordenados de manera creciente, por lo que la asociación caso-distancia no es la misma). Las distancias D^2 de Mahalanobis se ordenan de menor a mayor (tercera columna del cuadro 2.18).

2. Para cada distancia se calcula el percentil $(j - 0,5)/n$, donde n es el número de casos (10).

3. Se calculan los valores χ^2 de los percentiles de una distribución χ^2 con p grados de libertad, donde p es el número de variables implicadas (5 en nuestro ejemplo). Esto puede realizarse, por ejemplo, con la función `qchisq(seq(from=.05,to=.95,by=.1) ,5,lower.tail = T)` de `R`.

4. Se representan en un gráfico de dispersión el cuadrado de la distancia de Mahalanobis y el valor χ^2. La relación debería ser lineal, puesto que, de no serlo, estaríamos ante desviaciones de la normalidad.

Como se observa en la figura 2.10, la relación no es evidentemente lineal, aunque, como se explicó al comentar los gráficos Q-Q, puede calcularse el coeficiente de correlación entre las variables relacionadas en el gráfico (0,9008) y compararlo con los valores críticos del cuadro 2.16. Al ser inferior al valor crítico para $p < 0{,}05$ y superior al de $p < 0{,}01$, se asume la normalidad multivariante de la distribución al no poder rechazar la hipótesis nula ($p > 0{,}01$).

El paquete de `R`, `MVN`, permite calcular los test de Mardia, Henze-Zirkler y Royston de una manera sencilla ofreciendo, además, de manera directa el gráfico ji-cuadrado que hemos derivado de manera manual anteriormente. La sintaxis es inmediata:

```
library(MVN)
mardiaTest(Datos_2_3_Caso, qqplot = TRUE)
hzTest(Datos_2_3_Caso, qqplot = TRUE)
roystonTest(Datos_2_3_Caso, qqplot = TRUE)
```

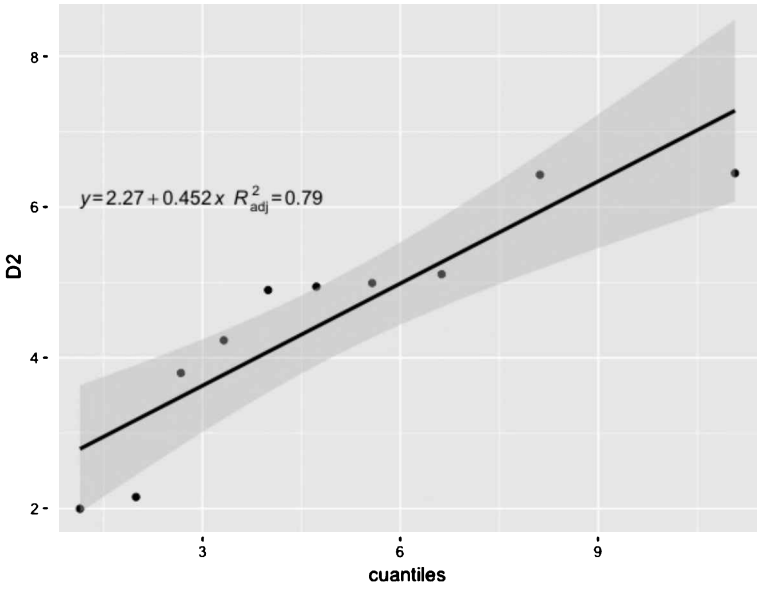
Aunque no aplicables por el tamaño muestral a nuestro ejemplo, el cuadro 2.19 muestra la sencillez de interpretación de la salida donde, de acuerdo con la misma, los datos empleados en el caso 2.3 serían multivariantemente normales.

Estadísticos e-Books & Papers

Cuadro 2.18.: Test de contraste de la normalidad univariante

Observación j	D^2	Nivel de probabilidad $(j-1/2)/n$	Cuantiles χ^2
1	2,00	0,05	1,15
2	2,15	0,15	1,99
3	3,80	0,25	2,67
4	4,23	0,35	3,33
5	4,90	0,45	4,00
6	4,94	0,55	4,73
7	4,99	0,65	5,57
8	5,11	0,75	6,63
9	6,43	0,85	8,12
10	6,45	0,95	11,08

Figura 2.10.: Gráfico ji-cuadrado



© Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 2.19.: Resultados de los test de normalidad multivariante

```
Mardia's Multivariate Normality Test
---------------------------------------
   data : Datos_2_3_Caso

   g1p            : 12.59093
   chi.skew       : 20.98489
   p.value.skew   : 0.9705687

   g2p            : 27.56964
   z.kurtosis     : -1.404206
   p.value.kurt   : 0.1602574

   chi.small.skew : 30.00839
   p.value.small  : 0.7077495

   Result         : Data are multivariate normal.
---------------------------------------

Henze-Zirkler's Multivariate Normality Test
---------------------------------------------
  data : Datos_2_3_Caso

  HZ      : 0.7560497
  p-value : 0.3157237

  Result  : Data are multivariate normal.
---------------------------------------------

Royston's Multivariate Normality Test
---------------------------------------------
  data : Datos_2_3_Caso

  H       : 4.765144
  p-value : 0.05091048

  Result  : Data are multivariate normal.
---------------------------------------------
```


Estadísticos e-Books & Papers

Cuadro 2.20.: Test de normalidad antes y después de la transformación

```
> shapiro.test(Datos_2_2_Caso$SYS)

     Shapiro-Wilk normality test

data:  Datos_2_2_Caso$SYS
W = 0.77262, p-value = 3.879e-11

> logSYS<-log(Datos_2_2_Caso$SYS)
> shapiro.test(logSYS)

     Shapiro-Wilk normality test

data:  logSYS
W = 0.97676, p-value = 0.07418
```

La cuestión que se plantea si no se confirma la normalidad univariante o multivariante es qué hacer. La respuesta es transformar los valores originales. El tipo de transformación depende del problema de asimetría o curtosis que cause la no normalidad. La figura 2.11 ilustra el efecto de posibles transformaciones, aunque hay que tener en cuenta que no siempre es posible realizarlas, puesto que la variable original puede tener una interpretación teórica y no ser fácil de interpretar, por ejemplo, el coeficiente de la variable transformada en una regresión lineal múltiple. También algunas funciones de transformación no permiten valores originales nulos o negativos.

Podemos ilustrar las transformaciones con los datos del caso 2.2, que tiene un tamaño muestral superior y más adaptado a la realidad de este tipo de pruebas. Si aplicamos el test de Shapiro-Wilk a la variable que mide el sueldo de los directivos (SYS), vemos inmediatamente que esta variable no sigue una distribución normal (cuadro 2.20). Como se aprecia en el panel (a) de la figura 2.12, la distribución es fuertemente asimétrica hacia el lado izquierdo. Siguiendo la figura 2.11, parece que la transformación más adecuada es el logaritmo de SYS. Realizamos esta transformación y ahora el test de Shapiro-Wilk no permite descartar la hipótesis nula de normalidad, lo que ya evidenciaba el panel (b) de la figura 2.12.

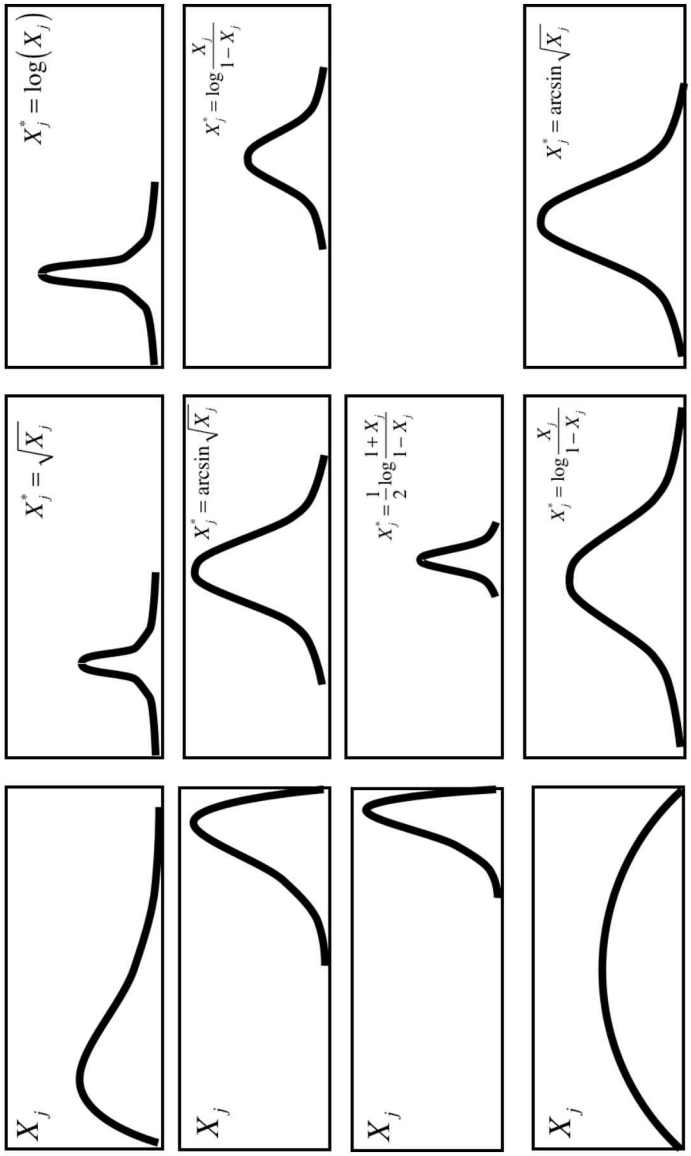
2.4.2. Homoscedasticidad

La homoscedasticidad debe definirse de manera distinta según estemos hablando de datos no agrupados (caso de una regresión lineal múltiple) o datos agrupados (caso de un análisis de la varianza de un factor). En el primer caso, la hipótesis de homoscedasticidad puede definirse como la asunción de que la variabilidad de los valores de una variable continua se mantiene más o menos constante para todos los valores de otra variable continua. En el caso de datos

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

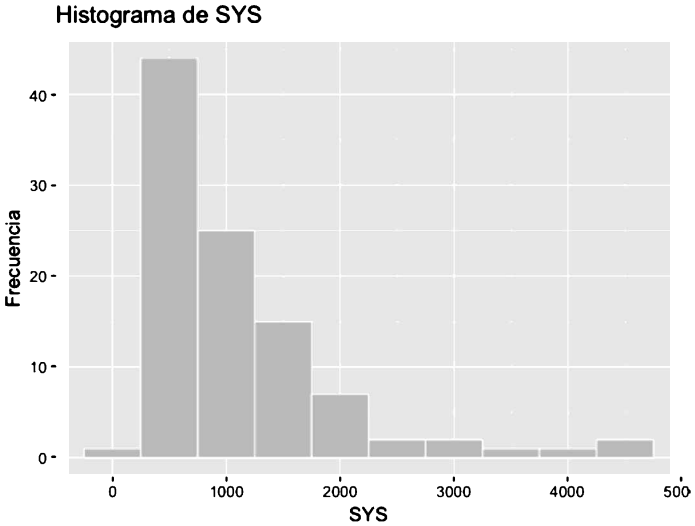
Figura 2.11.: Transformaciones en búsqueda de normalidad



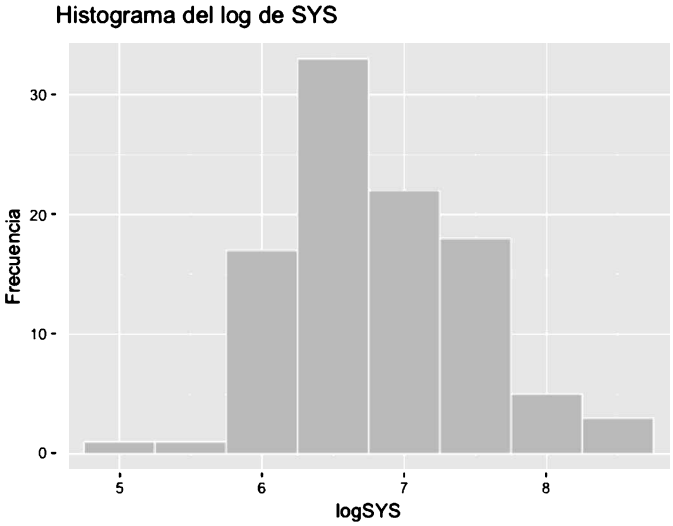
Fuente: Adaptado de Rummel (1970) y Stevens (1996).

Estadísticos e-Books & Papers

Figura 2.12.: Histogramas de la variable SYS y su logaritmo



(a) Variable SYS original



(b) Logaritmo de la variable SYS

© Ediciones Paraninfo

Estadísticos e-Books & Papers

agrupados, la homoscedasticidad implica que la varianza de la variable continua es más o menos la misma en todos los grupos que conforma la variable no métrica que delimita los grupos.

El análisis de la homoscedasticidad en datos no agrupados se abordará con profundidad en el capítulo dedicado al análisis de regresión múltiple. Dedicaremos este epígrafe a evaluarla para datos agrupados utilizando para ello los datos completos del caso 2.1 (los disponibles en la web del libro[3], no la selección simulada del cuadro 2.2). Según el problema planteado, el contraste puede ser univariante (se contrasta si la varianza es la misma) o multivariante (se contrasta si la matriz de covarianzas es o no la misma).

Un investigador puede plantearse si la población en su conjunto es consciente de que "Fumar perjudica la salud" (V1) o si, por el contrario, los fumadores tienen tal hábito porque no son conscientes de este hecho. De ser así, los fumadores estarían significativamente más en desacuerdo con esta afirmación que los no fumadores.

Para contrastar esta hipótesis, como se verá en el capítulo 6, debe realizarse un análisis de la varianza. Pero esta técnica exige homocedasticidad, es decir, que la varianza de la variable V1 (opinión sobre que fumar perjudica la salud) sea más o menos la misma en los grupos que conforma la variable que nos dice si el individuo es o no fumador (C3 "Hábito"). La figura 2.13 ilustra lo que implicaría gráficamente el cumplimiento y la vulneración de este supuesto.

El contraste más habitual para evaluar la homocedasticidad univariante es el **test de Levene** (Levene, 1960). El test de Levene parte de la hipótesis nula de que la varianza de la variable Y (V1 en nuestro ejemplo) es la misma en los k subgrupos (2 en nuestro ejemplo, fumadores y no fumadores) que forma la variable X (C3 en nuestro ejemplo). Esto es:

$$\begin{array}{ll} H_0: \ \sigma_1 = \sigma_2 = \ldots = \sigma_k & \\ H_a: \ \sigma_i \neq \sigma_j & \text{para al menos un par } i, j \end{array} \tag{2.9}$$

siendo σ_i la desviación típica de la variable Y en el subgrupo i. Si N es el tamaño muestral, entonces el estadístico de Levene (W) adopta la siguiente expresión:

$$W = \frac{N-k}{k-1} \frac{\sum_{i=1}^{k} N_i \left(\overline{Z}_{i.} - \overline{Z}_{..}\right)^2}{\sum_{i=1}^{k} \sum_{j=1}^{N} \left(Z_{ij} - \overline{Z}_{i.}\right)} \tag{2.10}$$

donde toda la notación es conocida salvo:

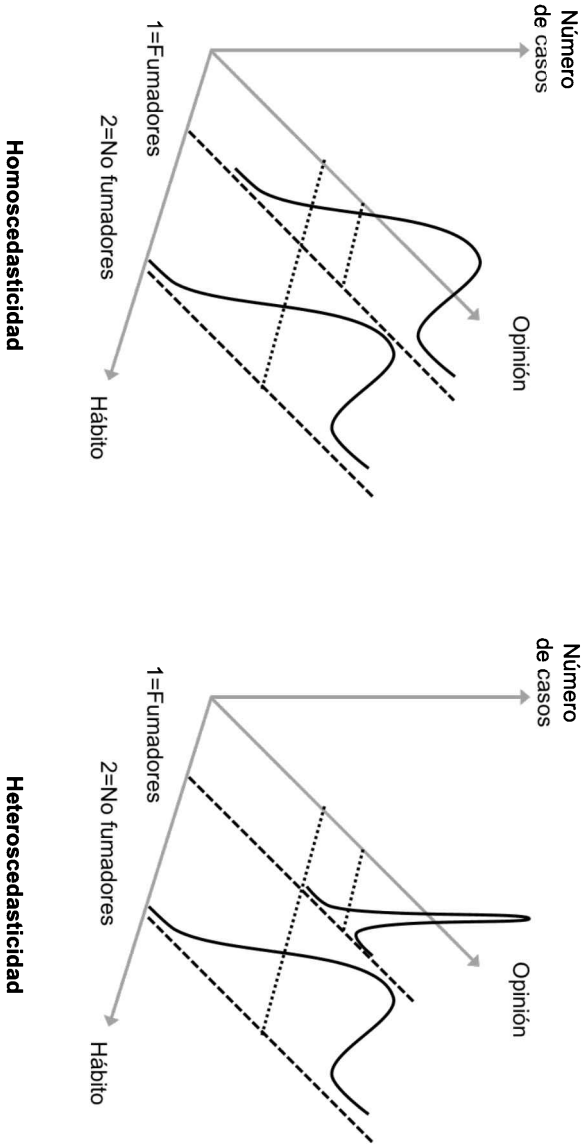
$$Z_{ij} = \left|Y_{ij} - \overline{Y}_{i.}\right| \tag{2.11}$$

siendo $\overline{Y}_{i.}$ la media del subgrupo i, $\overline{Z}_{i.}$ son las medias de los distintos subgrupos de Z_{ij} y $\overline{Z}_{..}$ es la media de Z_{ij} para el conjunto de la muestra sin distinguir grupos. La definición de $\overline{Y}_{i.}$ es la original de Levene. Otros autores, sin embargo,

[3]Para diferenciar, hemos denominado a esa base de datos `Datos_2_1b_Caso`.

Estadísticos e-Books & Papers

Figura 2.13.: Ejemplos de homocedasticidad y heterocedasticidad



© Ediciones Paraninfo

Estadísticos e-Books & Papers

han demostrado que utilizar la mediana o la media recortada al 10 % mejora la robustez de la prueba.

El test de Levene rechaza la hipótesis nula de homocedasticidad si:

$$W > F_{(\alpha,k-1,N-k)} \tag{2.12}$$

es decir, si el estadístico supera el nivel crítico de una distribución F con $k-1$ y $N-k$ grados de libertad para un nivel de significación de α.

Existen distintos paquetes en `R` para calcular el test de Levene, recurriendo a la función `leveneTest` del paquete `car`:

```
library(car)
leveneTest(Datos_2_1b_Caso$v1,Datos_2_1b_Caso$c3,center=mean)
leveneTest(Datos_2_1b_Caso$v2,Datos_2_1b_Caso$c3,center=mean)
leveneTest(Datos_2_1b_Caso$v3,Datos_2_1b_Caso$c3,center=mean)
leveneTest(Datos_2_1b_Caso$v4,Datos_2_1b_Caso$c3,center=mean)
leveneTest(Datos_2_1b_Caso$v5,Datos_2_1b_Caso$c3,center=mean)
```

Como se observa en el cuadro 2.21, se puede comprobar como, para un nivel de significación del 5 %, solo las variables V2 (no debe permitirse fumar en lugares públicos) y V5 (debe informarse sobre los efectos) no seguirían una distribución homocedástica entre fumadores y no fumadores. Llegado el momento, cuando abordemos el análisis de la varianza, analizaremos las alternativas para que este resultado no dificulte las conclusiones de nuestro análisis.

El problema que exige el contraste de la homoscedasticidad multivariante es distinto. Supongamos que deseamos saber si la actitud de los individuos respecto al tabaco (variables V1 a V5) puede explicar razonablemente que el entrevistado sea o no fumador. Como se verá en el capítulo 9, estaríamos ante el típico problema de un análisis discriminante: variable dependiente (ser fumador o no) no métrica con un conjunto de variables independientes (V1 a V5) métricas. Pues bien, como también se verá en ese tema, el análisis discriminante exige que las matrices de varianzas-covarianzas de las variables V1 a V5 sean estadísticamente iguales en el grupo de fumadores y en el de no fumadores. El cuadro 2.22 ofrece las mencionadas matrices muestrales, obtenidas mediante el comando `by` del paquete `{stats}`.

Para contrastar la hipótesis nula de igualdad de matrices de varianzas y covarianzas se utiliza habitualmente el **contraste M de Box**. Dado que en el capítulo 7 se desarrolla detenidamente, aquí solo indicaremos que es un estadístico que necesita de normalidad multivariante y que es muy sensible, por lo que diversos autores recomiendan utilizar para el contraste de la hipótesis nula un nivel de significatividad muy bajo ($p < 0,001$). El cuadro 2.23 muestra los resultados de este test para los datos de nuestro ejemplo obtenidos mediante la función `boxM` del paquete `{biotools}`. Puede comprobarse como la hipótesis nula de igualdad de la matriz de covarianzas puede rechazarse para cualquier nivel de significatividad. Esto provoca que no pueda aplicarse, por ejemplo, el análisis discriminante sin transformaciones previas de las variables.

Estadísticos e-Books & Papers

Cuadro 2.21.: Test de normalidad antes y después de la transformación

```
V1
Levene's Test for Homogeneity of Variance (center = mean)
       Df F value Pr(>F)
group   1  0.0597 0.8072
      239

V2
Levene's Test for Homogeneity of Variance (center = mean)
       Df F value    Pr(>F)
group   1  16.498 6.606e-05 ***
      239

V3
Levene's Test for Homogeneity of Variance (center = mean)
       Df F value Pr(>F)
group   1  2.5608 0.1109
      239

V4
Levene's Test for Homogeneity of Variance (center = mean)
       Df F value  Pr(>F)
group   1  2.9714 0.08604 .
      239

V5
Levene's Test for Homogeneity of Variance (center = mean)
       Df F value    Pr(>F)
group   1  13.525 0.0002908 ***
      239
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 2.22.: Matrices de covarianzas para fumadores (1) y no fumadores (2)

```
library(stats)
by(Datos_2_1b_Caso[,c("v1","v2","v3","v4","v5")],Datos_2_1b_Caso
$c3,cov)

Datos_2_1b_Caso$c3: 1
             v1         v2          v3           v4          v5
v1  0.770939870  0.2514548 -0.01507671  0.003262211 -0.07344384
v2  0.251454770  1.4635867 -0.13463234  0.460941633  0.23426203
v3 -0.015076706 -0.1346323  0.97160995 -0.072914830 -0.07617704
v4  0.003262211  0.4609416 -0.07291483  1.331511197  0.26573797
v5 -0.073443837  0.2342620 -0.07617704  0.265737965  1.03209311
------------------------------------------------------------
Datos_2_1b_Caso$c3: 2
           v1        v2          v3         v4         v5
v1 0.90612726 0.2660195  0.06127258  0.2326338 0.26691729
v2 0.26601953 0.8756032  0.14891707  0.5374257 0.19924812
v3 0.06127258 0.1489171  1.09392885 -0.1473460 0.05263158
v4 0.23263382 0.5374257 -0.14734598  1.5657053 0.20300752
v5 0.26691729 0.1992481  0.05263158  0.2030075 0.55263158
```

Cuadro 2.23.: Test M de Box

```
library(biotools)
boxM(Datos_2_1b_Caso[,c("v1","v2","v3","v4","v5")],Datos_2_
1b_Caso$c3)

   Box's M-test for Homogeneity of Covariance Matrices

data:  Datos_2_1b_Caso[, c("v1", "v2", "v3", "v4", "v5")]
Chi-Sq (approx.) = 50.357, df = 15, p-value = 1.053e-05
```


Estadísticos e-Books & Papers

2.4.3. Linealidad

La asunción de linealidad es fundamental para todas aquellas técnicas que se centran en el análisis de las matrices de correlaciones o de varianzas covarianzas, como el análisis factorial o los modelos de estructuras de covarianza. La razón es sencilla, el coeficiente de correlación de Pearson solo podrá capturar una relación si esta es lineal. Si la relación existe y es intensa pero, por ejemplo, es curvilinea, el coeficiente de correlación de Pearson tomará un valor bajo y el investigador puede interpretarlo como ausencia de relación, cuando, de hecho, esta existe, solo que no es lineal.

Cuando la técnica empleada tiene una variable dependiente, como ocurre en el caso de la regresión lineal múltiple, existen diversos procedimientos para contrastar la linealidad de las relaciones basadas en el análisis de los residuos. Estas técnicas se verán en el capítulo 8. En este epígrafe nos centramos en el análisis previo de los datos antes de aplicar técnica alguna, por lo que el único procedimiento para comprobar la existencia de relaciones lineales es sencillo: el análisis de los **gráficos de dispersión bivariados** entre las variables implicadas.

Aplicando este procedimiento a los datos del caso 2.3, se observa en el gráfico 2.14 como algunas de las relaciones que afloran son claramente no lineales, como es el caso del consumo del automóvil que crece más que proporcionalmente con la cilindrada, la potencia y el peso. Sin embargo, la relación entre las demás variables, dejando fuera el consumo, sí que parecen guardar una relación cercana a la linealidad.

Esto permite interpretar con más precisión la matriz de correlaciones del cuadro que aparece en ese mismo gráfico. Todas las correlaciones son significativas y fuertes, pero, mientras las correlaciones donde está implicada la variable consumo están enmascarando una relación no lineal, el resto corresponde a relaciones lineales.

2.4.4. Independencia de las observaciones

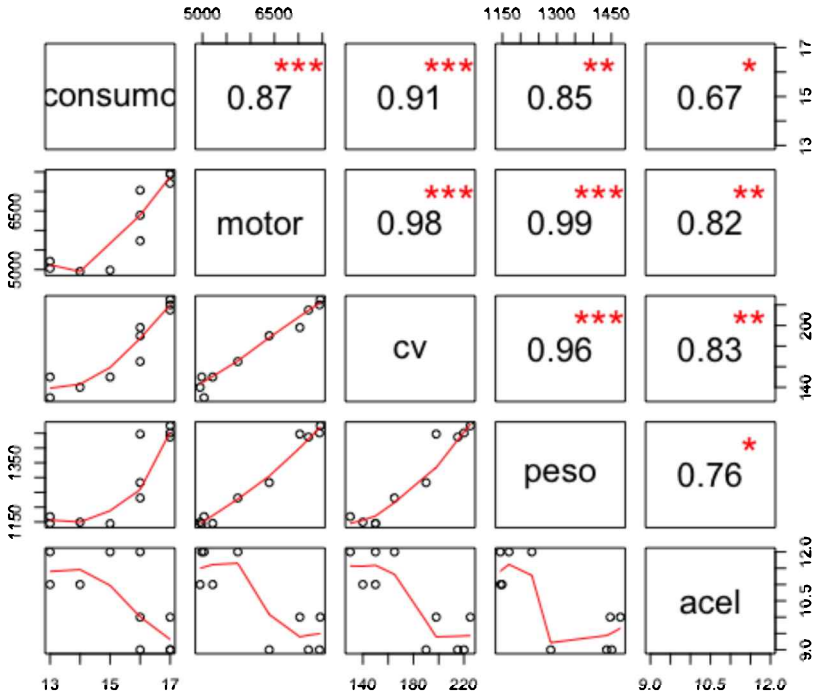
Dos observaciones son independientes cuando los valores que toman las variables de ese caso no se ven influidas por las observaciones que hayan tomado en otro caso. En investigación de mercados podría concretarse en que las respuestas dadas en un cuestionario por un individuo no influyen en las que dará otro, lo que no siempre es sencillo. Si se envía un cuestionario por correo a una familia esperando que lo contesten los dos cónyuges es difícil creer que las respuestas van a ser independientes, que no van a consultar juntos el cuestionario ni comentarlo entre ellos.

Como señala Sharma (1996), la influencia que tiene la violación del supuesto de independencia sobre los niveles de significatividad y la potencia de las pruebas es muy importante. Si las observaciones no son independientes, el nivel de significación de las pruebas debería incrementarse al menos 10 veces (Scariano y Davenport, 1987), esto es, si habitualmente rechazamos una hipótesis

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 2.14.: Gráficos de dispersión bivariados



Estadísticos e-Books & Papers

nula cuando $p < 0,05$, si sospechamos no independencia, deberíamos hacerlo a partir de $p < 0,005$.

La única solución efectiva para preservar el supuesto de independencia pasa por ser cuidadosos en los diseños de las investigaciones. Siguiendo el ejemplo de la entrevista a los dos cónyuges, quizás el investigador debiera plantearse optar por una entrevista personal, con claras instrucciones al entrevistador respecto a su no realización simultánea a los dos sujetos y a evitar la presencia de uno mientras se entrevista al otro.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

3. Análisis de conglomerados

3.1. Introducción

Supóngase que el responsable de marketing de una empresa tiene una base de datos con las características sociodemográficas de sus clientes: edad, nivel educativo, nivel de ingresos, estado civil, tipo de ocupación, número de hijos, etc. Este directivo se plantea si podría dividir a sus clientes en subgrupos que tuvieran características sociodemográficas similares entre sí, pero que unos subgrupos de otros fueran lo más diferentes posibles. Si esto fuera posible, el directivo de marketing podría, por ejemplo, diseñar campañas de publicidad distintas para cada grupo, con creatividades diferentes o utilizando diarios, revistas o cadenas de televisión distintas según el grupo al que fuera dirigida la campaña.

El **análisis de conglomerados**, al que también se denomina comúnmente **análisis *cluster***, es una técnica diseñada para clasificar distintas observaciones en grupos, de tal forma que:

1. Cada grupo (conglomerado o *cluster*) sea homogéneo respecto a las variables utilizadas para caracterizarlo, es decir, que cada observación contenida en él sea parecida a todas las que estén incluidas en ese grupo.

2. Que los grupos sean lo más distintos posible unos de otros respecto a las variables consideradas.

Es importante señalar, para distinguir el análisis de conglomerados de otras técnicas tratadas anteriormente en este libro, que los grupos son desconocidos *a priori* y es necesario derivarlos de las observaciones. En el análisis discriminante o la regresión logística, por ejemplo, las observaciones ya están previamente clasificadas en dos o más grupos, buscándose las razones que explican esa clasificación y no la clasificación en sí.

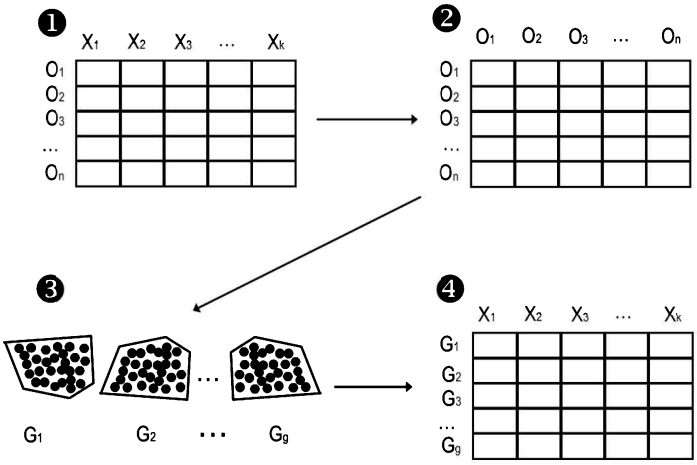
En la figura 3.1 se ilustra la secuencia lógica que se sigue al efectuar un análisis de conglomerados.

1. Inicialmente, el investigador dispone de n observaciones (individuos, empresas, etc.) de las que tiene información sobre k variables (edad, estado civil, número de hijos...).

2. A continuación, se establece un indicador que nos diga en qué medida cada par de observaciones se parece entre sí. A esta medida se la denomina distancia o (di)similaridad.

Estadísticos e-Books & Papers

Figura 3.1.: Proceso de realización de un análisis de conglomerados



3. El paso siguiente consiste en hacer grupos con aquellas observaciones que más se parezcan entre sí, de acuerdo con la medida de similaridad calculada anteriormente. Ello exige elegir entre los dos tipos de análisis de conglomerados: jerárquico y no jerárquico y el método de conglomeración para el tipo de análisis elegido (*v.g.*, centroide o vecino más cercano, entre otros, en el conglomerado jerárquico), como detallaremos a lo largo del tema.

4. Finalmente, el investigador debe describir los grupos que ha obtenido y comparar los unos con los otros. Para ello bastará con ver qué valores promedio toman las k variables utilizadas en el análisis de conglomerados en cada uno de los g grupos obtenidos ($g \leq n$).

Se describirán a continuación las decisiones que es necesario adoptar en cada una de las etapas descritas y que solo se han enunciado someramente.

3.2. Medidas de similaridad

Caso 3.1. Relación entre la publicidad y las ventas
Supongamos que un investigador tiene información del presupuesto que un conjunto de empresas ha destinado a publicidad el último año y de las ventas que han logrado en ese mismo ejercicio (cuadro 3.1). Puede preguntarse si estas empresas pueden agruparse en función de la rentabilidad en términos de ventas que han sido capaces de generar con su inversión publicitaria. Por ejemplo, el investigador puede examinar si existe un grupo de empresas que, invirtiendo

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 3.1.: Inversión en publicidad y ventas de 8 empresas hipotéticas

Nombre de la empresa	Inversión publicitaria	Ventas (millardos)
E1	16	10
E2	12	14
E3	10	22
E4	12	25
E5	45	10
E6	50	15
E7	45	25
E8	50	27

en publicidad relativamente poco, ha logrado una elevada cifra de ventas o, por el contrario, si existe un grupo que, aun invirtiendo mucho en publicidad, no ha sido capaz de vender tanto como sus competidoras. En definitiva, ¿qué tipología de empresas puede establecerse en función de la rentabilidad obtenida de su inversión publicitaria?

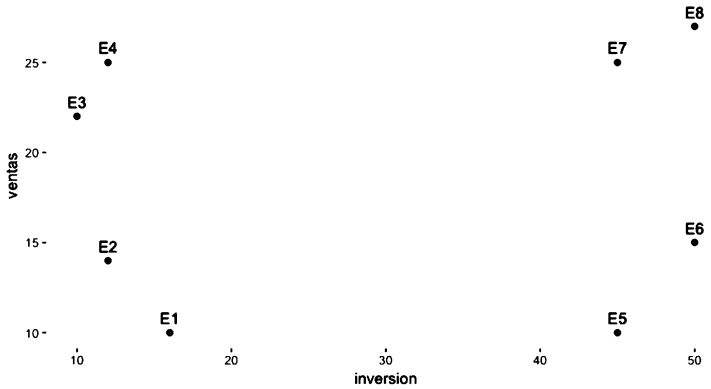
La figura 3.2 ilustra gráficamente los datos anteriores. Al haber utilizado solo dos variables en el ejemplo planteado, este gráfico permite responder de una manera intuitiva a las preguntas que se hace el investigador. A la vista de este gráfico pueden distinguirse cuatro grupos de empresas:

- El grupo formado por las empresas E1 y E2, que, con una pequeña inversión en publicidad, han obtenido también pocas ventas.
- El grupo formado por las empresas E3 y E4, que, pese a haber invertido tan poco como las empresas del grupo anterior, han obtenido una gran rentabilidad en términos de ventas a estas inversiones.
- El grupo formado por las empresas E5 y E6, que, pese a haber efectuado un gran esfuerzo publicitario, no han sido capaces de obtener unas ventas razonables.
- El grupo formado por las empresas E7 y E8, que, con inversiones también elevadas, sí que han logrado, por el contrario, rentabilizar su inversión en términos de ventas.

¿Cómo se han obtenido los grupos anteriores? De una manera intuitiva hemos visto, por ejemplo, que la empresa E1 está a una distancia menor de E2 que de E3 o que de cualquiera de las empresas restantes, y las hemos puesto en un mismo grupo. De manera análoga, e igualmente intuitiva, hemos procedido con las demás empresas, llegando a la solución de cuatro grupos expuesta. Pero ¿qué hubiera ocurrido si en lugar de dos variables pretendiésemos llevar a cabo agrupaciones de observaciones teniendo en cuenta 5, 10 o 50 variables? La intuición debe dejar paso a la formalización. Sin embargo, ilustraremos el

Estadísticos e-Books & Papers

Figura 3.2.: Gráfico de dispersión de los datos hipotéticos



proceso que sigue el análisis de conglomerados con este ejemplo sencillo para, finalmente, aplicarlo a una situación más real en el último epígrafe del capítulo.

Lo primero que se ha hecho de manera intuitiva es ver que E1 está más cerca de E2 que de E3. Este "más cerca" se traduce en el análisis de conglomerados en el cálculo de alguna medida de proximidad o similaridad entre cada par de observaciones. En función del tipo de variables que se estén utilizando para caracterizar a los objetos, las medidas más adecuadas serán diferentes. Comenzaremos ilustrando una de ellas con los datos del ejemplo anterior para señalar posteriormente cómo se calculan otras distancias y en qué casos es recomendable la utilización de unas u otras.

3.2.1. Medidas de similaridad para variables métricas

En el caso en que las variables que se utilizan para caracterizar las observaciones sean métricas, es decir, de intervalo o de razón (véase capítulo 1), se puede recurrir a cualquiera de las siguientes medidas de similaridad.

A. Distancia euclídea

Si consideramos dos observaciones i y j de las n posibles y si llamamos x_{ip} y x_{jp} al valor que toma la variable x_p de las k existentes en dichas observaciones, la distancia euclídea D_{ij} entre ambas se calcularía del siguiente modo:

$$D_{ij} = \sqrt{\sum_{p=1}^{k} (x_{ip} - x_{jp})^2} \tag{3.1}$$

así, por ejemplo, la distancia euclídea entre las empresas E1 y E2 toma el valor siguiente:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 3.2.: Matriz de distancias euclídeas para los datos del ejemplo

	1	2	3	4	5	6	7	8
1	0.00							
2	5.66	0.00						
3	13.42	8.25	0.00					
4	15.52	11.00	3.61	0.00				
5	29.00	33.24	37.00	36.25	0.00			
6	34.37	38.01	40.61	39.29	7.07	0.00		
7	32.65	34.79	35.13	33.00	15.00	11.18	0.00	
8	38.01	40.16	40.31	38.05	17.72	12.00	5.39	0.00

$$D_{12} = \sqrt{(16-12)^2 + (10-14)^2} = \sqrt{32} = 5,66$$

que, correspondiéndose con la intuición derivada del análisis de la figura 3.2, es menor que la distancia existente entre las empresas E1 y E3:

$$D_{13} = \sqrt{(16-10)^2 + (10-22)^2} = \sqrt{180} = 13,42$$

La mayoría de algoritmos calculan las distancias entre todos los pares de observaciones, como paso inicial del análisis de conglomerados, ofreciendo la matriz que se recoge en el cuadro 3.2 para los datos del ejemplo utilizado. Se ha obtenido con la función `dist` del paquete `{stats}`.

```
matriz.dis.euclid<-dist(DatosCaso3.1[,c("inversion","ventas")],
method="euclidean",diag=TRUE)round(print(matriz.dis.euclid),2)
```

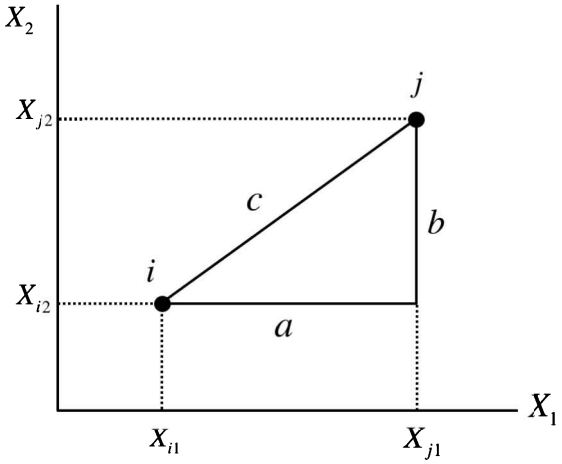
B. Distancia euclídea al cuadrado

El cálculo de la raíz cuadrada al que obliga la aplicación de la distancia euclídea, tal y como se aprecia en la expresión (3.1), puede ser demasiado exigente en términos de capacidad de cómputo del ordenador, sobre todo si se utiliza en combinación de métodos de agrupación como los de Ward o el del centroide, que serán descritos posteriormente, y el número de casos y/o variables es también muy elevado. Una forma sencilla de reducir los cálculos que hay que realizar consiste en tomar como medida de similaridad el cuadrado de la distancia euclídea:

$$D_{ij} = \sum_{p=1}^{k} (x_{ip} - x_{jp})^2 \tag{3.2}$$

Estadísticos e-Books & Papers

Figura 3.3.: Ilustración de la distancia *city block*



C. Distancia de Minkowski

Las dos distancias descritas anteriormente son un caso particular de la distancia de Minkowski, que viene dada por la expresión:

$$D_{ij} = \left[\sum_{p=1}^{k} |x_{ip} - x_{jp}|^n \right]^{1/n} \tag{3.3}$$

Puede comprobarse que haciendo $n = 2$ en 3.3 se obtiene la expresión correspondiente a la distancia euclídea.

D. Distancia *city block* o "Manhattan"

Si en la expresión de la distancia de Minkowski tomáramos $n = 1$, obtendríamos la denominada distancia *city block,* en alusión al hecho de que la distancia entre dos observaciones se calcula siguiendo el camino que un transeúnte utilizaría en una ciudad para trasladarse entre dos puntos i y j donde la línea recta c es imposible por la existencia de una manzana de edificios ($a + b$ en la figura 3.3). La expresión de esta distancia viene dada por:

$$D_{ij} = \sum_{p=1}^{k} |x_{ip} - x_{jp}|^n \tag{3.4}$$

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 3.3.: Base de datos hipotética de variables binarias

Observaciones	Variables			
	X1	X2	X3	X4
E1	1	1	0	0
E2	0	1	1	1
E3	1	1	0	1
E4	0	0	0	1
E5	1	1	1	0

Cuadro 3.4.: Cálculo de similitudes

		E1	
		1	0
E2	1	1	2
	0	1	0

		E1	
		1	0
E2	1	a	b
	0	c	d

3.2.2. Medidas de similaridad para datos binarios

En algunas ocasiones, las variables utilizadas para caracterizar a las observaciones están codificadas como ficticias, es decir, únicamente contemplan la presencia (1) o ausencia (0) del atributo considerado. Estas variables suelen aflorar en el proceso de codificación de atributos medidos en escalas nominales u ordinales.

Consideraremos, para ilustrar el cálculo de algunas de estas medidas, una hipotética base de datos formada por 5 observaciones de 4 variables, tal y como se recoge en el cuadro 3.3.

Pues bien, para calcular las medidas de similaridad se construye en primer lugar una matriz 2×2 para cada par de observaciones que se están comparando. En esta matriz se recogen las coincidencias y las divergencias entre las distintas variables correspondientes a las dos observaciones comparadas, tal y como se ilustra para el caso de las observaciones E1 y E2 en el cuadro 3.4.

Dado que, como puede comprobarse en el cuadro 3.3, la observación E1 presenta un 1 a la vez que la E2 en una sola ocasión (para la variable X2), la celda a que recoge este hecho aparece como 1. Como para las variables X3 y X4, el atributo está presente en E2 y ausente en E1, en la casilla b aparece un 2. Análogamente, en este ejemplo, las casillas c y d toman los valores 1 y 0, respectivamente.

De este modo, se calculan distintas medidas de similitud, las más utilizadas

Estadísticos e-Books & Papers

de las cuales son las siguientes. Para dos observaciones i y j cualquiera:

- Índice de Jaccard (1901):
  $\sqrt{1-[a/(a+b+c)]}$
- Coeficiente *simple matching* de Sokal y Michener (1958):
  $\sqrt{1-[(a+d)/(a+b+c+d)]}$
- Sokal y Sneath (1963):
  $\sqrt{1-[a/(a+2(b+c)]}$
- Rogers y Tanimoto (1960):
  $\sqrt{1-[(a+d)/(a+2(b+c)+d)]}$
- Dice (1945) o Sorenson (1948):
  $\sqrt{1-[2a/(2a+b+c)]}$
- Coeficiente de Hamann (Gower y Legendre, 1986)
  $\sqrt{1-[(a-(b+c)+d)/(a+b+c+d)]}$
- Ochiai (1957):
  $\sqrt{1-\left[a/\sqrt{(a+b)(a+c)}\right]}$
- Sokal y Sneath (1963):
  $\sqrt{1-\left[ad/\sqrt{(a+b)(a+c)(d+b)(d+c)}\right]}$
- Phi de Pearson g (Gower y Legendre, 1986):
  $\sqrt{1-\left[(ad-bc)/\sqrt{(a+b)(a+c)(d+b)(d+c)}\right]}$
- Coeficiente S2 de Gower y Legendre (1986):
  $\sqrt{1-[a/(a+b+c+d)]}$

El cuadro 3.5 ilustra para los hipotéticos datos anteriores, la salida obtenida mediante la función `dist.binary {ade4}` de `R` del índice de Jaccard `(method=1)` y el coeficiente de *simple matching* `(method=2)`, aunque el mencionado paquete las calcula todas. Así, por ejemplo, la distancia entre las observaciones E1 y E2 respondería al siguiente cálculo:

$$D_{Jaccard} = \sqrt{1-\frac{a}{a+b+c}} = \sqrt{1-\frac{1}{1+2+1}} = 0,8660$$

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 3.5.: Matriz de distancias euclídeas para los datos del ejemplo

```
> dist.binary(DatosCuadro3.5[,c("X1","X2","X3","X4")], method
= 1, diag = TRUE, upper = FALSE)
          1         2         3         4         5
1 0.0000000
2 0.8660254 0.0000000
3 0.5773503 0.7071068 0.0000000
4 1.0000000 0.8164966 0.8164966 0.0000000
5 0.5773503 0.7071068 0.7071068 1.0000000 0.0000000

> dist.binary(DatosCuadro3.5[,c("X1","X2","X3","X4")], method
= 2, diag = TRUE, upper = FALSE)
          1         2         3         4         5
1 0.0000000
2 0.8660254 0.0000000
3 0.5000000 0.7071068 0.0000000
4 0.8660254 0.7071068 0.7071068 0.0000000
5 0.5000000 0.7071068 0.7071068 1.0000000 0.0000000
```

$$D_{simple\ matching} = \sqrt{1 - \frac{a+d}{a+b+c+d}} = \sqrt{1 - \frac{1+0}{1+2+1+0}} = 0,8660$$

3.2.3. Estandarización de los datos

Si se analizan con detenimiento las medidas de distancia presentadas en apartados anteriores, se puede comprobar que todas ellas están basadas en la sustracción, para cada par de observaciones, de los valores de las variables utilizadas en su caracterización. Por ello, se puede esperar que las medidas de disimilaridad sean muy sensibles a las unidades en que estén medidas dichas variables. Si pretendemos agrupar empresas en función de dos variables, como el tamaño de su activos medido en pesetas[1] y el número de trabajadores, la primera variable contribuirá mucho más a establecer los grupos que la segunda. Y esto no se debe a que conceptualmente una sea mucho más importante que la otra, sino a que, con esas unidades, su valor absoluto será siempre muy superior.

Veamos este problema con el siguiente ejemplo. El cuadro 3.6 recoge el tamaño de los activos y el número de trabajadores de ocho empresas hipotéticas. Si efectuamos un análisis de conglomerados con las unidades originales, la matriz de distancias que se obtiene (cuadro 3.7) muestra que los dos grupos obtenidos responden exclusivamente a la variable "activos de la empresa", puesto que sitúa en un mismo grupo a aquellas con cifras que rondan los 10.000 millones de pesetas (E1, E2, E3 y E4) y en otro grupo a las que tienen activos en torno

[1] El uso de esta moneda solo pretende que, para un mismo valor monetario, el número que lo representa sea mucho más grande y, de esta manera, visualizar mejor el efecto de las unidades de medida.

Estadísticos e-Books & Papers

Cuadro 3.6.: Activo y número de trabajadores de 8 empresas hipotéticas

Nombre de la empresa	Activos (pesetas)	Trabajadores
E1	10.000.000.000	100
E2	10.050.000.000	90
E3	10.000.000.000	200
E4	10.050.000.000	190
E5	20.000.000.000	200
E6	20.050.000.000	190
E7	20.000.000.000	100
E8	20.050.000.000	90

Cuadro 3.7.: Matriz de distancias euclídeas para los datos del ejemplo

Grupo 1

```
          1         2         3         4         5         6         7         8
1 0.00e+00
2 5.00e+08 0.00e+00
3 1.00e+02 5.00e+08 0.00e+00
4 5.00e+08 1.00e+02 5.00e+08 0.00e+00
5 1.00e+10 9.50e+09 1.00e+10 9.50e+09 0.00e+00
6 1.05e+10 1.00e+10 1.05e+10 1.00e+10 5.00e+08 0.00e+00
7 1.00e+10 9.50e+09 1.00e+10 9.50e+09 1.00e+02 5.00e+08 0.00e+00
8 1.05e+10 1.00e+10 1.05e+10 1.00e+10 5.00e+08 1.00e+02 5.00e+08 0.00e+00
```

Grupo 2

a los 20.000 millones (E5, E6, E7 y E8). Es decir, la influencia del número de trabajadores en la obtención de estos conglomerados es prácticamente nula.

Para evitar esta influencia no deseable de una variable debida exclusivamente a la unidad en que viene medida es necesario corregir el efecto de los datos recurriendo a un proceso de estandarización. R ofrece distintas posibilidades, de las que se detallan las de uso más frecuente.

Puntuaciones Z Los datos son estandarizados, restando al valor de cada observación de una variable determinada, la media de esa variable para el conjunto de las observaciones y dividiendo el resultado por su desviación típica. De esta forma la variable estandarizada tiene media 0, y desviación típica, 1.

Rango 1 El valor de una variable dada en cada observación es dividido por el rango de esa variable para el conjunto de observaciones. De esta forma el rango de variación de la variable así estandarizada queda reducido a un intervalo de valor 1.

Rango 0 a 1 El valor de una variable determinada para cada observación es estandarizado sustrayéndole el valor mínimo que toma esa variable en el

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 3.8.: Activo y número de trabajadores de 8 empresas hipotéticas

Nombre de la empresa	Activos (pesetas)	Trabajadores	Valores estandarizados	
			Activos	Trabajadores
E1	10.000.000.000	100	-0,98	-0,84
E2	10.050.000.000	90	-0,89	-1,02
E3	10.000.000.000	200	-0,98	1,02
E4	10.050.000.000	190	-0,89	0,84
E5	20.000.000.000	200	0,89	1,02
E6	20.050.000.000	190	0,98	0,84
E7	20.000.000.000	100	0,89	-0,84
E8	20.050.000.000	90	0,98	-1,02
Media	15.025.000.000	145	0,00	0,00
Desv. Típica	5.000.062.499	50,24	1,00	1,00

conjunto de las observaciones y a continuación dividiendo por el rango. De esta forma el valor mínimo de las variables será 0, y el máximo, 1.

Veamos si estandarizando los datos del cuadro 3.6, por ejemplo mediante el procedimiento de las puntuaciones Z, se logra corregir la influencia desproporcionada de la variable activos de la empresa en la formación de los grupos. Usaremos para ello la función `scale{base}`. Los datos se encuentran en el cuadro 3.8.

```
DatosCuadro3.7.norm<-scale(DatosCuadro3.7
[,c("activos","trabajadores")])
matriz.dis.euclid.norm<-dist(DatosCuadro3.7.norm
[,c("activos","trabajadores")],
method="euclidean",diag=TRUE)
```

A continuación volvemos a calcular la matriz de distancias con los datos tipificados del cuadro 3.8, la matriz de distancias que se obtiene (cuadro 3.9) muestra como ahora aparecen cuatro grupos formados por dos empresas que se parecen mucho entre sí. Así, el formado por E1 y E2 tienen activos en torno a los 10.000 millones, pero los separa del grupo formado por E3 y E4 el hecho de que estas últimas empresas les doblan en términos de número de trabajadores. Se observa como, estandarizando los datos, se elimina el efecto de las unidades de medida y las dos variables que caracterizan las observaciones tienen el mismo peso a la hora de formar los grupos.

Estadísticos e-Books & Papers

Cuadro 3.9.: Matriz de distancias euclídeas para los datos del ejemplo tras la normalización

	1	2	3	4	5	6	7	8
1	0.00							
2	0.21	0.00						
3	1.86	2.05	0.00					
4	1.68	1.86	0.21	0.00				
5	2.64	2.71	1.87	1.78	0.00			
6	2.58	2.64	1.97	1.87	0.21	0.00		
7	1.87	1.78	2.64	2.44	1.86	1.68	0.00	
8	1.97	1.87	2.84	2.64	2.05	1.86	0.21	0.00

Grupo 1: E1 y E2 **Grupo 2: E3 y E4** **Grupo 3: E5 y E6** **Grupo 3: E7 y E8**

3.3. Formación de los grupos: análisis jerárquico de conglomerados

Una vez que, mediante el cálculo de la matriz de distancias, se sabe qué observaciones están más próximas entre sí, y más distantes de otras, es necesario formar los grupos, lo que implica tomar dos decisiones: selección del algoritmo de agrupación que se quiere seguir y determinación de cuál es el número de grupos razonable dados los datos.

Adoptar estas decisiones no es sencillo desde el momento en que existen decenas de algoritmos de agrupación. La mayoría de autores, de hecho, recomiendan utilizar diversos procedimientos y comparar sus resultados (Sharma, 1996; Johnson, 1998). Si distintos métodos aportan agrupaciones similares, será razonable suponer que existe una agrupación natural objetiva. Si no fuera así, habría que examinar las distintas agrupaciones a la luz de un marco teórico o de trabajos precedentes para elegir el resultado más razonable.

Los algoritmos de agrupación existentes responden a dos grandes enfoques:

1. **Métodos jerárquicos**. Suponen la toma de $n-1$ decisiones de agrupación (donde n es el tamaño muestral). Existen dos enfoques, los métodos jerárquicos *aglomerativos,* en los que, inicialmente, cada individuo es un grupo en sí mismo. Sucesivamente se van formando grupos de mayor tamaño fusionando grupos cercanos entre sí. Finalmente, todos los individuos confluyen en un solo grupo. En los métodos jerárquicos *desagregativos*, el proceso es equivalente, solo que, inicialmente, todos los individuos forman un único grupo y se van sucesivamente desgajando de él, formando dos grupos, tres grupos y así hasta que al final del proceso cada caso forma un único grupo. La mayoría de paquetes estadísticos

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

usan el primer enfoque y en él nos centraremos en la presentación de los métodos jerárquicos.

2. **Métodos no jerárquicos**. Los grupos no se forman en un proceso secuencial de fusión de grupos de menor tamaño. En estos métodos se establece inicialmente un número de grupos a priori y los individuos se van clasificando en cada uno de esos grupos. Es decir, una solución de cinco grupos no proviene de agregar dos grupos de la solución de seis, sino que es aquella solución de cinco grupos en la que existe una mayor homogeneidad entre los miembros que pertenecen a cada uno de ellos y, además, cada grupo es lo más distinto posible de los demás.

Comenzaremos desarrollando los principales algoritmos de agrupamiento jerárquico para, en el epígrafe siguiente, centrarnos en los métodos no jerárquicos. En ese momento se señalará bajo qué condiciones es más adecuado optar por uno u otro tipo de procedimiento.

3.3.1. Método del centroide

El método del centroide (Sokal y Michener, 1958) está implementado en la función de `R, hclust{stats}`. Ilustraremos su funcionamiento con los datos de las ocho empresas del cuadro 3.1. En primer lugar, como se indicó con anterioridad, se calcula la matriz de distancias, en este caso euclídea al cuadrado, entre las ocho empresas, según se refleja en el cuadro 3.10.

```
#calculo de la distancia euclídea
matriz.dis.euclid<-dist(DatosCaso3.1[,c("inversion","ventas")],
method="euclidean",diag=TRUE)

#calculo de la distancia euclidea al cuadrado
matriz.dis.euclid2<-(matriz.dis.euclid)^2

#efectuamos el cluster con método centroide
hclust.centroide<-hclust(matriz.dis.euclid2,method="centroid")
plot(hclust.centroide,labels=DatosCaso3.1$nombre.empresa)

#Saca el historial de aglomeración del objeto
hclust.centroide data.frame(hclust.centroide[2:1])
```

Pues bien, el método del centroide comienza uniendo aquellas dos observaciones que están más cercanas, en este caso las empresas E3 y E4 (la distancia es 13). A continuación el grupo formado es sustituido por una observación que lo representa y en la que las variables toman los valores medios de todas las observaciones que constituyen el grupo representado (centroide). En nuestro ejemplo, las empresas E3 y E4 son sustituidas por una empresa promedio, que

Estadísticos e-Books & Papers

Cuadro 3.10.: Matriz de distancias euclídeas al cuadrado para los datos del caso 3.1

	1	2	3	4	5	6	7	8
1	0							
2	32	0						
3	180	68	0					
4	241	121	13	0				
5	841	1105	1369	1314	0			
6	1181	1445	1649	1544	50	0		
7	1066	1210	1234	1089	225	125	0	
8	1445	1613	1625	1448	314	144	29	0

llamaremos E3-4, para la que el gasto en publicidad y las ventas toman los siguientes valores:

$$\textit{Publicidad} \text{ de E3-4} = \frac{10+12}{2} = 11$$

$$\textit{Ventas} \text{ de E3-4} = \frac{22+25}{2} = 23,5$$

En ese momento se recalcula la matriz de distancias, solo que, en lugar de estar presentes las empresas E3 y E4, está su centroide E3-4. Se unen entonces aquellas dos observaciones que están de nuevo más cerca. `hclust` muestra esas distancias sucesivas en lo que denominamos el historial de conglomeración, tal y como se ilustra en el cuadro 3.11. Vemos como, efectivamente, en el primer paso se fusionarion los casos E3 y E4 y lo hicieron a una distancia (*height*) de 13. En el segundo paso, son ahora las empresas E7 y E8, que están a una distancia de 29, las que se fusionarán. Ahora las empresas E7 y E8 serán sustituidas por su centroide E7-8, se recalculará la matriz de distancias y se repetirá el proceso. Este termina cuando todas las empresas están en un solo grupo. En el historial de conglomeración se observa que en la etapa siguiente (etapa 3) se juntan las empresas E1 y E2 y, en la etapa 4, las E5 y E6. Solo en la etapa 5 dejan de fusionarse empresas individuales para fusionarse dos grupos (lo que `hclust` identifica en el cese en el uso del signo $-$). Se fusionan las empresas que lo hicieron en el paso 1 (E3-4) con las que lo hicieron en el paso 3 (E1-2), y ese es el indicador (el número del paso) que se muestra en el cuadro.

Es importante señalar que en la columna *height* que aparece en el cuadro 3.11 se reflejan, como hemos indicado, las distancias a la que estaban los grupos que se van fusionando en cada etapa. Así, el coeficiente de la primera etapa es 13 porque, como ya se indicó, las empresas que se fusionan, E3 y E4, están a esa distancia.

El **historial de conglomeración** tiene una traducción gráfica que es de gran utilidad para determinar el número razonable de grupos que debe retenerse. A

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 3.11.: Datos en el paso 2 del proceso de conglomeración e historial de conglomeración

Nombre de la empresa	Inversión publicitaria	Ventas (millardos)
E1	16	10
E2	12	14
E3-4	11	23,5
E5	45	10
E6	50	15
E7	45	25
E8	50	27

```
   height merge.1 merge.2
1   13.00      -3      -4
2   29.00      -7      -8
3   32.00      -1      -2
4   50.00      -5      -6
5  141.25       1       3
6  182.25       2       4
7 1227.25       5       6
```

este gráfico se le denomina **dendograma** y viene recogido en la figura 3.4.

En este dendograma hemos señalado con un número rodeado de un círculo la etapa del historial de conglomeración que veíamos en el cuadro 3.11. ¿Cómo sirve el dendograma para determinar cuál es el número razonable de grupos que debe retenerse? Como hemos señalado, el análisis de conglomerados jerárquico comienza considerando a cada individuo como un grupo independiente y sucesivamente va fusionando a los más cercanos hasta que todos forman un solo grupo. Pero cada etapa une individuos más distantes, es decir, más diferentes, menos susceptibles de formar un grupo. Obsérvese en el cuadro 3.11 que, mientras la primera etapa fusiona observaciones que distan 13 unidades, en la etapa 5 se unen individuos que distan 141 unidades. ¿Dónde cortar y dejar de fusionar? Con independencia de los indicadores que propondremos posteriormente, la respuesta es que en aquel momento en que la fusión siguiente va a unir individuos muy distintos, es decir, donde el dendograma da un gran salto (marcado con rectángulos en la figura 3.4). Obsérvese como los grupos que se formaron en la etapa 5 (empresas 3, 4, 1 y 2) y los que se formaron en la 6 (7, 8, 5 y 6) están a tal distancia que no es razonable fusionarlos. Esos dos grupos son los que el analista debería retener.

Estadísticos e-Books & Papers

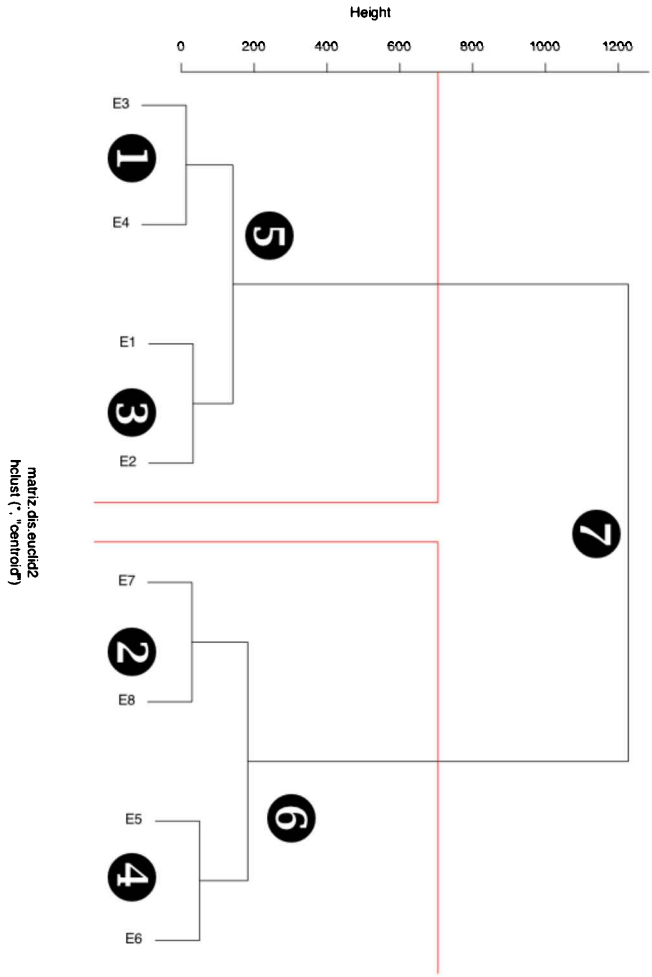


Figura 3.4.: Dendograma

© Ediciones Paraninfo

Estadísticos e-Books & Papers

3.3.2. Método del vecino más cercano

En el método anterior, la distancia entre los grupos se obtenía calculando las distancias entre sus centroides. En el método del vecino más cercano (Florek *et al.*, 1951; Sokal y Michener, 1958), que en algunos textos aparece también bajo la etiqueta de vinculación simple *single linkage*, la distancia entre dos grupos es aquella que se da entre los dos miembros más cercanos de esos grupos. Así, retomando el ejemplo anterior (véase cuadro 3.6), la distancia entre los grupos E1-2 y E3-4 estará representada por la distancia entre E2 y E3, que son los más cercanos (68 unidades, puesto que E1 dista 180 unidades de E3, y de E4, 241 unidades, mientras que E2 dista 121 unidades de E4, como se veía en la matriz de distancias del cuadro 3.10).

El historial de conglomeración que proporciona el `hclust` para este método aparece recogido también en la figura 3.5. Puede comprobarse, por ejemplo, que el coeficiente de la etapa 6 es 125, que se corresponde con la distancia entre E6 y E7 (véase cuadro 3.10), que son los "vecinos más cercanos" de sus respectivos grupos.

3.3.3. Método del vecino más lejano

El método del vecino más lejano, al que también se refieren algunos textos como vinculación completa, *complete linkage* (Sorenson, 1948), es análogo al anterior, pero con la diferencia de que la distancia entre dos grupos se mide por la distancia entre sus miembros más alejados. La figura 3.6 ilustra, como en el caso anterior, los historiales de conglomeración. Puede comprobarse ahora que el coeficiente de la etapa 5 es 241, que se corresponde con la distancia entre las empresas E1 y E4 (cuadro 3.10), o que en la etapa 7 el coeficiente es 1649, correspondiéndose con la distancia entre las empresas 3 y 6.

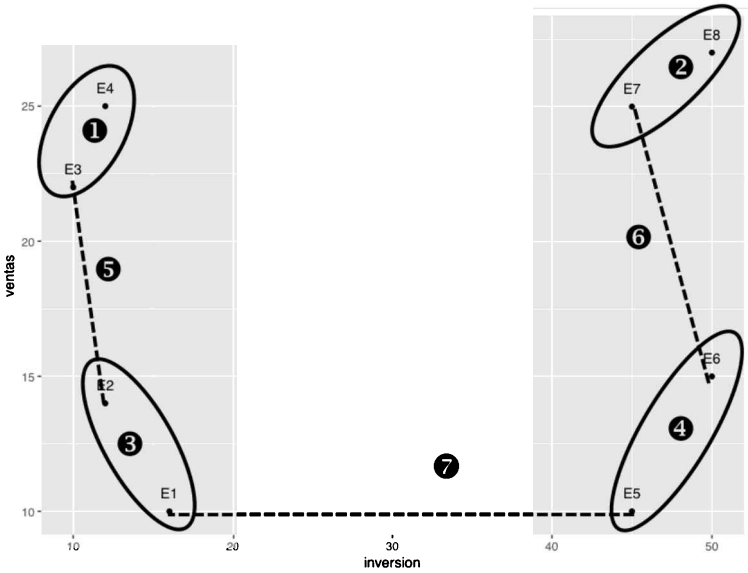
3.3.4. Método de la vinculación promedio

En este procedimiento, conocido también como *average linkage* (Sokal y Michener, 1958), la distancia entre dos grupos se obtiene calculando la distancia promedio entre todos los pares de observaciones que pueden formarse tomando un miembro de un grupo y otro miembro del otro grupo. Ilustremos este procedimiento con los datos del ejemplo que venimos utilizando. En el cuadro 3.14, se observa como en la etapa 5 se fusiona el grupo formado por las empresas E1 y E2 (etiquetado como 3 en el cuadro) con el formado por las empresas E3 y E4 (etiquetado como 1 en el cuadro). El coeficiente, es decir, la distancia entre ambos grupos, es de 152,5. ¿Cómo se ha obtenido esa distancia con el procedimiento de vinculación promedio?

El cuadro 3.13 recoge todas las posibles combinaciones entre pares de puntos de los grupos E1-2 y E3-4 en las que hay un miembro de cada grupo. Las distancias entre esos pares de observaciones las hemos tomado del cuadro 3.10. Pues bien, la distancia entre los mencionados grupos es el promedio de todas ellas, es decir:

Estadísticos e-Books & Papers

Figura 3.5.: Historial de conglomeración "vecino más cercano"



	height	merge.1	merge.2
1	13	-3	-4
2	29	-7	-8
3	32	-1	-2
4	50	-5	-6
5	68	1	3
6	125	2	4
7	841	5	6

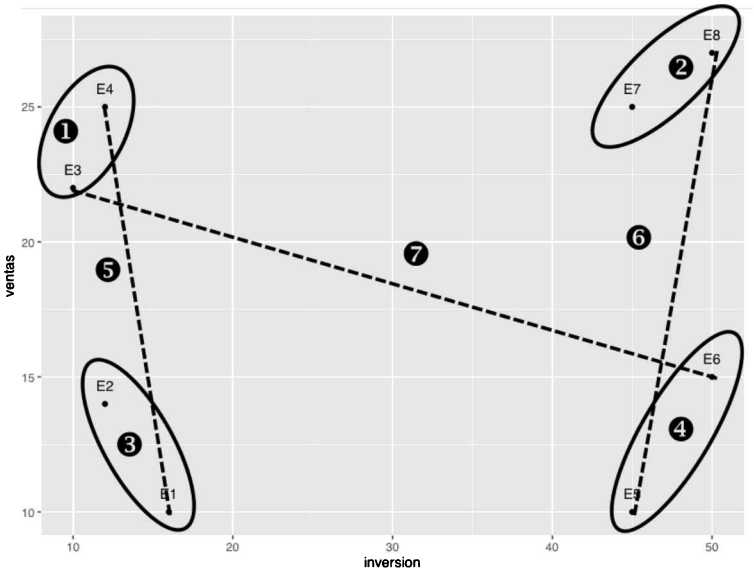
Cuadro 3.12.: Historial de conglomeración método "vinculación promedio"

	height	merge.1	merge.2
1	13.000	-3	-4
2	29.000	-7	-8
3	32.000	-1	-2
4	50.000	-5	-6
5	152.500	1	3
6	202.000	2	4
7	1323.625	5	6

© Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 3.6.: Historial de conglomeración "vecino más lejano"



	height	merge.1	merge.2
1	13	-3	-4
2	29	-7	-8
3	32	-1	-2
4	50	-5	-6
5	241	1	3
6	314	2	4
7	1649	5	6

Estadísticos e-Books & Papers

Cuadro 3.13.: Distancias entre pares de observaciones en la etapa 6 del método "vinculación promedio"

Pares de observaciones	Distancia	Promedio
E1,E3	180	152,5
E1,E4	241	
E2,E3	68	
E2,E4	121	

$$D_{E1-2,E3-4} = \frac{180 + 241 + 68 + 121}{4} = 152,5$$

3.3.5. Método de Ward

El método de Ward (Ward, 1963) no calcula la distancia entre los distintos conglomerados para decidir cuáles se deben fusionar, ya que su objetivo es maximizar la homogeneidad dentro de cada conglomerado. Para ello plantea todas las posibles combinaciones de observaciones para el número de grupos que se esté considerando en cada etapa concreta. Por ejemplo, para las ocho empresas del cuadro 3.1, se parte de que cada una es un grupo distinto. Por ello, en la etapa 1 se agrupan dos de ellas para lograr una solución de 7 grupos, concretamente las empresas E3 y E4. Después de concluir la etapa 5 se han obtenido los grupos siguientes: [E1,E2,E3,E4], [E5,E6] y [E7,E8]. Ahora, en la etapa 6, para obtener una solución de 2 grupos, se pueden realizar las fusiones que se plantean en el cuadro 3.14.

¿Qué criterio utiliza el método de Ward para decidir con qué solución se obtienen grupos más homogéneos entre sí? Calcula los centroides de los grupos resultantes de las posibles fusiones y a continuación obtiene la distancia euclídea al cuadrado al centroide de todas las observaciones del grupo ("suma de cuadrados total" en el cuadro 3.14). Aquella solución en que se obtiene una menor suma de cuadrados es la que garantiza la máxima homogeneidad y es, por tanto, la elegida. Veámoslo con los datos del ejemplo.

La tercera de las soluciones posibles recogidas en el cuadro 3.14 consiste en fusionar los grupos [E5,E6] y [E7,E8] y dejar sin fusionar el grupo formado por las empresas [E1,E2,E3,E4], con lo que quedarían dos grupos: [E5,E6,E7,E8] y [E1,E2,E3,E4]. ¿Qué grado de homogeneidad tendrían estos grupos?

Las empresas del grupo [E1,E2,E3,E4] distarían de su centroide:

$$\begin{gathered} D_{[E1,E2,E3,E4],C} = (16 - 12{,}5)^2 + (12 - 12{,}5)^2 + (10 - 12{,}5)^2 + \\ + (12 - 12,5)^2 + (10 - 17{,}75)^2 + (14 - 17{,}75)^2 + (22 - 17{,}75)^2 + \\ + (15 - 17{,}75)^2 = 163,75 \end{gathered}$$

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 3.14.: Decisión de los grupos a fusionar en la etapa 6 mediante el método de "Ward"

Alternativas para fusionar	Se deja sin fusionar	Suma de cuadrados total SCT
[E1,E2,E3,E4] + [E5,E6]	[E7,E8]	1.873,33
[E1,E2,E3,E4] + [E7,E8]	[E5,E6]	1.927,33
[E5,E6]+[E7,E8]	[E1,E2,E3,E4]	385,50

```
     height merge.1 merge.2
1     6.500      -3      -4
2    21.000      -7      -8
3    32.000      -1      -2
4    37.000      -5      -6
5   203.250       1       3
6   385.500       2       4
7  2840.000       5       6
```

mientras que las empresas del grupo [E5,E6,E7,E8] distarían del suyo:

$$D_{[E5,E6,E7,E8],C} = (45-47{,}5)^2 + (50-47{,}5)^2 + (45-47{,}5)^2 + \\ + (50-47,5)^2 + (10-19,25)^2 + (15-19,25)^2 \\ + (25-19,25)^2 + (27-19,25)^2 = 221,75$$

En resumen, la heterogeneidad total medida del modo descrito sería:

$$SCT = D_{[E1,E2,E3,E4],C} + D_{[E5,E6,E7,E8],C} = 163,75 + 221,75 = 385,5$$

que, como el lector podría calcular con los datos del cuadro 3.14, es la menor de todas las posibles fusiones y por eso es la que el análisis de conglomerados elige, como se observa en el historial de conglomeración de ese mismo cuadro.

3.4. Selección del número de conglomerados de la solución

Como se ha mostrado, el análisis de conglomerados jerárquico ofrece al investigador la posibilidad de elegir entre muchas soluciones que difieren en cuanto al número de conglomerados finales que las conforman: desde un grupo por cada observación, hasta un único grupo que integraría todas las observaciones. Ha de decidirse, pues, cuál es el número de conglomerados que conforman una

Estadísticos e-Books & Papers

solución razonable.

Algunos programas estadísticos, como el SPSS, ofrecen solamente el dendograma como herramienta de apoyo para tomar esta decisión. Su uso ya se ilustró en el apartado 3.3.1. Debe detenerse el proceso de fusión cuando los grupos que se han de unir están a una distancia significativamente mayor de los que previamente se han fusionado.

Algunos autores como Hair *et al.* (2014a), tras apuntar que no existe un criterio objetivo, en la medida en que no se ha construido un estadístico que ofrezca un criterio de decisión basado en la inferencia (Bock, 1985; Hartigan, 1985), proponen realizar el cálculo de las **tasas de variación** entre los coeficientes de conglomeración obtenidos entre etapas sucesivas. Así, cuando una tasa de variación sea drásticamente superior a la anterior será el momento de detener las fusiones. Esta tasa es fácil de obtener a partir de la información del *software*. El cuadro 3.5, que utiliza los coeficientes de conglomeración que se obtenían en la aplicación a los datos de nuestro ejemplo del método del vecino más cercano (Figura 3.5), muestra que lo razonable es no ejecutar la etapa 7, donde el coeficiente da un salto del 573 %. Por lo tanto, convendrá retener la solución de dos grupos que se resalta en el cuadro.

Sin embargo, en fechas recientes, algunos autores han revisado e implementado en `R` un conjunto de índices que, si bien de manera individual pueden tener mayor o menor eficacia en la detección del número óptimo de conglomerados, tomados en su conjunto pueden ser de gran ayuda en la medida en que una mayoría de ellos apunten a una solución determinada. Concretamente Charrad *et al.* (2014) han incorporado en su paquete de `R`, `NbClust`[2] un total de 30 índices que permiten tomar una decisión bien fundamentada.

Desarrollar cada uno de los 30 índices iría más allá del alcance de este manual y referimos al lector al mencionado trabajo de Charrad *et al.* (2014) del que este apartado es deudor. Nos limitaremos a presentar los tres que están implementados en *software* comercial (SAS), esto es, el índice CH, (Calinski y Harabasz, 1974), CCC (Sarle, 1983) y Pseudot2 (Duda y Hart, 1973) y aquellos que son compartidos por más paquetes de `R`, en concreto el índice DB (Davies y Bouldin, 1979)y el índice de Dunn (Dunn, 1974), junto con dos índices con soporte gráfico, el índice de Hubert (Hubert y Arabie, 1985) y el Dindex (Lebart *et al.*, 2000). En cualquier caso, el mencionado paquete `{NbClust}` los desarrolla con detalle y, en la aplicación que haremos al caso 3.2 las salidas los tendrán todos en cuenta.

3.4.1. Índice CH

El índice de Calinski y Harabasz (CH, Calinski y Harabasz, 1974) está definido por la expresión:

$$CH(q) = \frac{\text{trace}\,(B_q)\,/\,(q-1)}{\text{trace}\,(W_q)\,/\,(n-q)} \tag{3.5}$$

[2] `https://cran.r-project.org/web/packages/NbClust/`

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 3.15.: Aplicación de las tasas de variación de los coeficientes de conglomeración

Etapa	Observaciones que se fusiona	Grupos resultantes	Grupos	Coeficiente	Tasa de variación
1	[E3,E4]	[E3,E4], E1, E2, E5, E6, E7, E8	7	13	1,23
2	[E7,E8]	[E3,E4][E7,E8], E1, E2, E5, E6	6	29	0,10
3	[E1,E2]	[E1,E2][E3,E4][E7,E8], E5, E6	5	32	0,56
4	[E5,E6]	[E1,E2][E3,E4][E5,E6][E7,E8]	4	50	0,36
5	[E1,E2][E3,E4]	[E1,E2,E3,E4][E5,E6][E7,E8]	3	68	0,84
6	**[E5,E6][E7,E8]**	**[E1,E2,E3,E4][E5,E6,E7,E8]**	**2**	**125**	**5,73**
7	[E1,E2][E3,E4][E5,E6][E7,E8]	[E1,E2,E3,E4,E5,E6,E7,E8]	1	841	–

© Ediciones Paraninfo

Estadísticos e-Books & Papers

donde q es el número de conglomerados, n es el número de observaciones, B_q es la matriz de dispersión entre grupos para una solución de q clusters, esto es, $B_q = \sum_{k=1}^{q} n_k \left(c_k - \bar{x}\left(c_k - \bar{x}\right)^\top\right)$, donde n_k es el número de casos en el conglomerado k y $\bar{x}$ el centroide de la matriz de datos. Pues bien, el valor q que maximiza la expresión $CH\left(q\right)$ se considera el número más adecuado de conglomerados.

3.4.2. Índice CCC

El índice CCC (Cubic Clustering Criterion; Sarle, 1983) se calcula mediante la expresión:

$$CCC = \ln\left[\frac{1 - E\left(R^2\right)}{1 - R^2}\frac{\sqrt{\frac{np*}{2}}}{\left(0{,}001 + E\left(R^2\right)\right)^{1,2}}\right] \tag{3.6}$$

donde

$$R^2 = 1 - \frac{\text{trace}\left(X^\top X - \bar{X}^\top Z^\top Z\bar{X}\right)}{\text{trace}\left(X^\top X\right)}$$

siendo:

- $X^\top X$ es la matriz con la suma de cuadrados y productos cruzados (SSCP).
- $\bar{X} = \left(Z^\top Z\right)^{-1} Z^\top X$.
- Z es una matriz que caracteriza a los cluster cuyo elemento $z_{ik} = 1$ si la observación i-ésima pertenece al conglomerado k-ésimo y 0 en caso contrario.

$$E\left(R^2\right) = 1 - \left[\frac{\sum_{j=1}^{p^*}\frac{1}{n+u_j} + \sum_{j=p^*+1}^{p}\frac{u_j^2}{n+u_j}}{\sum_{j=1}^{p} u_j^2}\right]\left[\frac{\left(n-q\right)^2}{n}\right]\left[1 + \frac{4}{n}\right]$$

- $u_j = \frac{s_j}{c}$.
- s_j es la raíz cuadrada del autovalor j-ésimo de $X^\top X/\left(n-1\right)$.
- $v^* = \prod_{j=1}^{p^*} s_j$.
- p^* se elige para ser el entero más grande que siendo menor que q hace que u_{p^*} no sea menor que 1.

Pues bien, de acuerdo con Milligan y Cooper (1985), el máximo del índice permite determinar el número óptimo de conglomerados.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

3.4.3. Índice Pseudo t^2

El índice Pseudo t^2 propuesto por Duda y Hart (1973) solo es aplicable a métodos jerárquicos. Se obtiene mediante la expresión:

$$\text{Pseudo } t^2 = \frac{V_{kl}}{\frac{W_k+W_l}{n_k+n_l-2}} \tag{3.7}$$

donde $V_{kl} = W_m - W_k - W_l$ si $C_m = C_k \cup C_l$. La notación ya ha sido presentada en expresiones anteriores. Gordon (1999) demuestra que el número óptimo de conlgomerados es aquel que cumpla que:

$$\text{Pseudo } t^2 \leq \left(\frac{1 - critValueDuda}{critValueDuda}\right) \times (n_k + n_l - 2)$$

donde el criterio que se aplica al índice de Duda (Duda y Hart, 1973) puede consultarse en Charrad *et al.* (2014).

3.4.4. Índice DB

El índice DB (Davies y Bouldin, 1979) es una función de la ratio entre la dispersión interna de cada cluster y la separación entre ellos. Se calcula mediante la expresión:

$$DB\,(q) = \frac{1}{q}\sum_{k=1}^{q} \operatorname*{máx}_{k\neq l}\left(\frac{\delta_k + \delta_l}{d_{kl}}\right) \tag{3.8}$$

donde

- $k, l = 1, \ldots, q =$ número del conglomerado,
- $d_{kl} = \sqrt[v]{\sum_{j=1}^{p} |c_{kj} - c_{lj}|^v}$ es la distancia entre los centroides de los conglomerados C_k y C_l.
- $\delta_k = \sqrt[u]{\frac{1}{n_k}\sum_{i\in C_k}\sum_{j=1}^{p} |x_{ij} - c_{kj}|^u}$ es la medida de la dispersión del cluster C_k.

Pues bien, el valor q que minimiza DB(q) se considera el número adecuado de conglomerados (Milligan y Cooper, 1985; Davies y Bouldin, 1979).

3.4.5. Índice de Dunn

El índice de Dunn (Dunn, 1974) define la ratio entre la mínima distancia intraconglomerado y la máxima distancia interconglomerado. Lo hace del siguiente modo:

Estadísticos e-Books & Papers

$$\text{Dunn} = \frac{\min\limits_{1 \leq i < j \leq q} d\left(C_i, C_j\right)}{\max\limits_{1 \leq k \leq q} \operatorname{diam}\left(C_k\right)} \tag{3.9}$$

donde $d(C_i, C_j)$ es la función de disimilaridad entre dos conglomerados C_i y C_j definida como $d(C_i, C_j) = \min\limits_{x \in C_i, y \in C_j} d\left(x, y\right)$, y $\operatorname{diam}\left(C\right)$ es el diámetro de un conglomerado que puede considerarse como una medida de dispersión. Si los datos contienen conglomerados compactos y bien separados, el diámetro de los mismos debería ser pequeño y la distancia entre los conglomerados se espera que sea grande, de esta forma se maximizaría el índice de Dunn.

3.4.6. Estadístico de Hubert

El estadístico Γ de Hubert (Hubert y Arabie, 1985) es el coeficiente de correlación serial entre dos matrices que, cuando son simétricas, puede escribirse del siguiente modo:

$$\Gamma\left(P, Q\right)) \frac{1}{N_t} \sum_{\substack{i=1 \\ i<j}}^{n-1} P_{ij} Q_{ij} \tag{3.10}$$

donde P es la matriz de proximidad del conjunto de datos y Q es una matriz $n \times n$ cuyo elemento (i, j) es igual a la distancia entre dos puntos representativos (v_{ci}, v_{cj}) de los conglomerados a los que pertenecen x_i y x_j. Valores altos del estadístico Γ normalizado indican la existencia de conglomerados compactos, por lo que en el gráfico en el que se relaciona Γ y q (número de conglomerados) buscaremos un punto de inflexión que se corresponda con un incremento significativo de Γ (Halkidi *et al.*, 2001). El número de conglomerado donde esto ocurra será el óptimo. En el paquete `NbClust` se ofrecen también las segundas diferencias en el valor normalizado de Γ para distinguir un pico de este valor ocasionado por el óptimo de otras situaciones anómalas.

3.4.7. Índice Dindex

El índice Dindex (Lebart *et al.*, 2000) se basa en la ganancia de inercia dentro del conglomerado, que mide el grado de homogeneidad de los datos asociados con ese grupo. Calcula las distancias y las compara con un punto de referencia del perfil, normalmente el centroide. Se define:

$$w\left(P^q\right) = \frac{1}{q} \sum_{k=1}^{q} \frac{1}{n_k} \sum_{x_i \in C_k} d\left(x_i, c_k\right) \tag{3.11}$$

Dadas dos particiones P^{k-1} formadas por $k-1$ conglomerados y P^k particiones formadas por k conglomerados, la ganancia de inercia intraconglomerado se define por:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

$$\text{Gain} = w\left(P^{q-1}\right) - w\left(P^{q}\right) \tag{3.12}$$

debiendo minimizarse el valor de Gain, lo que se identifica con un recodo en el gráfico que se corresponde con una caída significativa de las primeras diferencias de Gain cuando se representa frente al número de conglomerados a cuya identificación ayuda el equivalente crecimiento en las segundas diferencias que debe acompañarle.

Caso 3.2 Aplicación de los índices a la identificación del número adecuado de conglomerados

Seguiremos en este caso el planteamiento seguido por Charrad *et al.* (2014) que generan una base de datos simulada de dos variables y 200 casos donde los conglomerados están perfectamente separados, en la medida en que son datos aleatorios que siguen una normal donde las medias de cada conglomerado son, respectivamente, 1, 3, 6 y 9 con distintas varianzas. El objetivo de esta simulación de datos es didáctico. Si el lector visualiza claramente que existen cuatro conglomerados muy definidos que el conjunto de índices es capaz de detectar, la confianza en la eficacia del sistema de índices para detectar el número adecuado de grupos en situaciones no evidentes se refuerza.

La figura 3.7 recoge los datos simulados mediante la siguiente sentencia:

```
set.seed(1) x<-rbind(
matrix(rnorm(100,sd=0.1),ncol=2),
matrix(rnorm(100,mean=1,sd=0.2),ncol=2),
matrix(rnorm(100,mean=5,sd=0.1),ncol=2),
matrix(rnorm(100,mean=7,sd=0.2),ncol=2))
DatosCaso3.2<-data.frame(x)
```

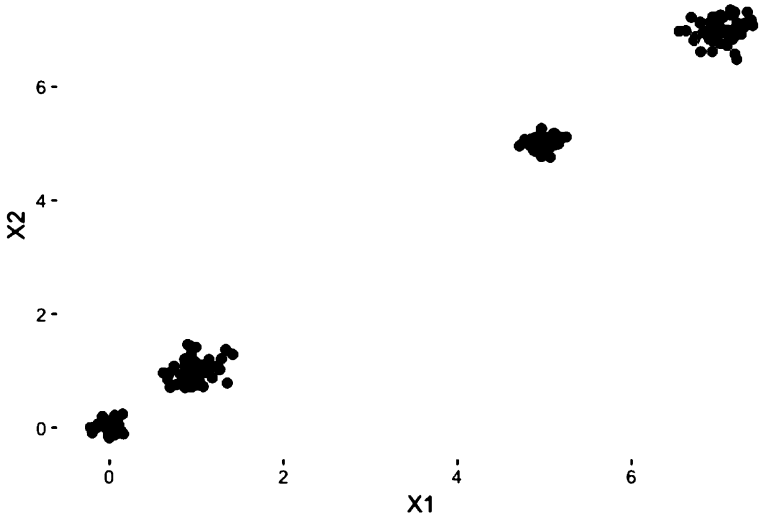
La sentencia que solicita el cálculo de los indicadores es la siguiente:

```
library(NbClust)
res<-NbClust(DatosCaso3.2, distance = "euclidean", min.nc=2,
max.nc=8, method = "ward.D2", index = "alllong")
res$All.index
res$Best.nc
res$Best.partition
```

El cuadro 3.16 muestra la propuesta que realiza cada índice respecto al número adecuado de conglomerados. Como dos de estos índices, Hubert y Dindex, tienen un formato gráfico —nótese que aparecen como 0.00 en el cuadro 3.16 por esta razón—, los recogemos en la figura 3.8. La conclusión más significativa que podemos sacar es que no todos los indicadores, incluso en un ejemplo de separación tan clara entre los conglomerados, realizan la misma propuesta (por ejemplo, en la primera columna el índice KL recomienda una solución de 4 grupos mientras que el TraceW recomienda la opción de 3). Esto nos debe

Estadísticos e-Books & Papers

Figura 3.7.: Datos simulados de cuatro conglomerados



Fuente: Charrad *et al.* (2014, p. 23)

llevar a seguir un criterio para decidir. Dos son las alternativas. La opción por defecto del paquete `NbClust` es lo que denomina *majority rule*, es decir elegir la solución que más indicadores recomiendan. Otra alternativa es analizar la propuesta que realizan aquellos indicadores que mejor desempeño han demostrado en los ejercicios de simulación. De acuerdo con Milligan y Cooper (1985), los mejores cinco indicadores serían los índices: CH, Duda, Cindex, Gamma y Beale.

El cuadro 3.17 muestra la salida de donde sintetiza su denominada regla de la mayoría. De acuerdo con esta salida, la opción elegida sería la de 4 conglomerados en la medida en que 14 de los indicadores la proponen. Si analizamos los cinco indicadores que se supone que tienen un mejor desempeño, la propuesta de los mismos sería: CH (4), Duda (4), Cindex (6) y Beale (4), lo que parece corroborar la regla de la mayoría.

Aunque el apoyo de estos índices es tremendamente valioso y supone una gran mejora sobre la mera consulta del dendograma, el investigador no debe olvidar que la mejor validación de los conglomerados es que estos tengan sentido y puedan interpretarse en el contexto del problema de investigación.

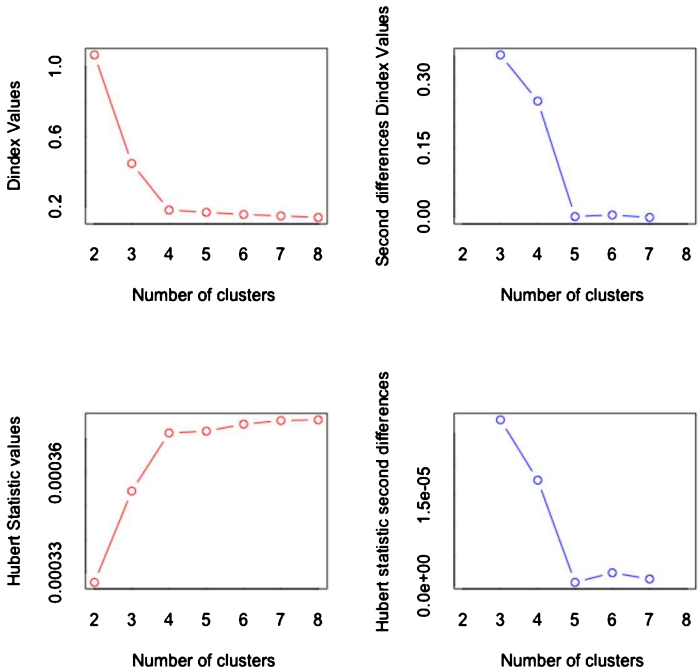
 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 3.16.: Decisión de los grupos a retener mediante distintos indicadores

```
                     KL       CH Hartigan     CCC    Scott  Marriot TrCovW
Number_clusters  4.0000     4.00   4.0000  4.0000   4.0000    4.000  7.000
Value_Index     64.3738 23004.82 983.2942 41.8671 460.5888 1614.143  5.477
                  TraceW Friedman     Rubin Cindex     DB Silhouette   Duda
Number_clusters   3.0000   4.0000    4.0000 6.0000 3.0000     3.0000 4.0000
Value_Index     153.2925 640.3886 -528.9154 0.2346 0.1451     0.8856 0.5971
                PseudoT2  Beale Ratkowsky     Ball PtBiserial    Gap Frey
Number_clusters   4.0000 4.0000    2.0000   3.0000     2.0000 4.0000   NA
Value_Index      32.3941 0.6611    0.6787 110.3298     0.9237 2.5891   NA
                McClain Gamma Gplus      Tau   Dunn Hubert SDindex Dindex
Number_clusters   4.000     2     2    2.000 2.0000      0  2.0000      0
Value_Index       0.135     1     0 4974.874 1.3974      0  0.8034      0
                  SDbw
Number_clusters 4.0000
Value_Index     0.0028
```

Figura 3.8.: Índices gráficos para la elección del número de conglomerados



Fuente: Charrad *et al.* (2014, p. 23)

Estadísticos e-Books & Papers

Cuadro 3.17.: Decisión de los grupos a fusionar en la etapa 6 mediante el método de Ward

```
*******************************************************************
* Among all indices:
* 7 proposed 2 as the best number of clusters
* 4 proposed 3 as the best number of clusters
* 14 proposed 4 as the best number of clusters
* 1 proposed 6 as the best number of clusters
* 1 proposed 7 as the best number of clusters

                    ***** Conclusion *****

* According to the majority rule, the best number of clusters is  4


*******************************************************************
```

3.5. Formación de los grupos: análisis no jerárquico de conglomerados

Como se indicó en el epígrafe 3.3, el análisis de conglomerados no jerárquico se caracteriza porque, a diferencia del jerárquico, se conoce a priori el número k de grupos que se desea, y las observaciones son entonces asignadas a cada uno de esos k conglomerados, de tal forma que se maximiza la homogeneidad de los sujetos asignados a un mismo grupo y la heterogeneidad entre los distintos conglomerados.

El análisis de conglomerados no jerárquico pasa por la realización de las siguientes tareas:

1. Se han de determinar los centroides iniciales de los k grupos, esto es, los valores medios de las variables que caracterizan las observaciones en cada uno de esos grupos. Estos centroides iniciales, que se conocen como *semillas*, pueden ser fijados por el investigador de acuerdo con información previa (el resultado de un conglomerado jerárquico, por ejemplo) o dejar que sea el algoritmo de conglomeración quien decida sus valores mediante el procedimiento que luego se describirá.

2. Una vez establecidas las semillas, cada observación se asigna a aquel conglomerado, de entre los k existentes, cuyo centroide esté más cercano a esa observación en términos de distancia euclídea.

3. Se recalculan entonces los centroides de los k grupos de acuerdo con las observaciones que han sido clasificadas en cada uno de ellos. Si el cambio en los centroides (distancia entre nuevos y viejos centroides) es mayor que un valor criterio de convergencia preestablecido, entonces se vuelve al paso

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

2, finalizando el proceso cuando se cumpla el criterio de convergencia o se supere un número prefijado de iteraciones.

Ilustraremos, utilizando los datos de ejemplo del cuadro 3.1, cada uno de los pasos señalados.

Selección de los centroides iniciales R recurre al siguiente procedimiento para escoger las semillas o centroides iniciales de los k conglomerados basado en las siguientes dos pruebas (cuando se utiliza el término distancia se hace referencia a distancia euclídea al cuadrado):

1. Se utilizan las k primeras observaciones del fichero de datos como centroide de partida. Se calcula la distancia entre las k observaciones y se retiene la correspondiente a las dos más cercanas (O_1 y O_2 en la figura 3.8). A continuación se determina si alguna de esas dos observaciones puede ser sustituida en el centroide por la observación $k+1$. Si la distancia de la observación $k+1$ a la observación del centroide (de las k) que tenga más cerca (O_k en la figura 3.8) es mayor que la distancia entre el par de observaciones O_1 y O_2, la observación $k+1$ sustituirá a una de las dos. La sustituida será la observación que esté más cerca (O_2 en la figura 3.9).

2. Si la observación $k+1$ no supera la prueba anterior, todavía puede entrar en el centroide si cumple la siguiente regla: sustituirá a la observación más cercana de las k existentes en el centroide si su distancia a cualquiera de las observaciones del centroide (exceptuando la más cercana) es más grande que la menor distancia de la más cercana a todas las que integran el centroide. Si tampoco supera esta prueba se probará con la siguiente observación del fichero de datos.

Ilustremos el proceso de elección de la semilla inicial con los datos de ejemplo del cuadro 3.1. Para seguir la ilustración es necesario también considerar el cuadro 3.10 donde aparecen las distancias euclídeas al cuadrado entre todas las observaciones del fichero.

En primer lugar se considera como centroide inicial las k primeras observaciones. Dado que, como vimos en el conglomerado jerárquico, la solución natural parece ser la de dos grupos, plantearemos un análisis de conglomerados no jerárquico para este mismo número de grupos. De esta forma, en el paso 1, las dos primeras observaciones del fichero, E1 y E2, están en el centroide. A continuación se determina qué par de observaciones del centroide están más cercanas. Como solo hay dos observaciones, se considera ese par. Inmediatamente, en ese mismo paso 1, se comprueba si la observación siguiente en el fichero (E3) es o no candidata a sustituir a la del par elegido del centroide que esté más cercana a ella (E2). La regla 1 dice que sustituirá a E2 si su distancia a ella (68) es mayor que lo que distan entre sí el par de observaciones del centroide más cercanas entre sí (32). Como este es el caso, E3 sustituye a E2

Estadísticos e-Books & Papers

Figura 3.9.: Reglas para la determinación del centroide inicial

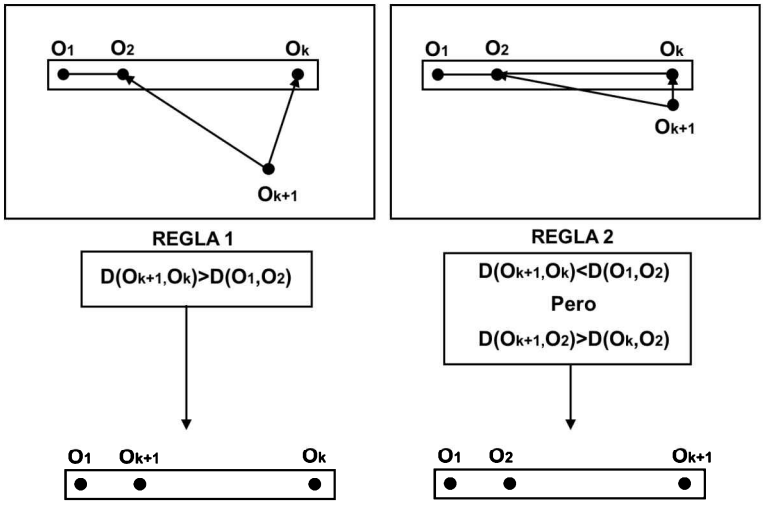
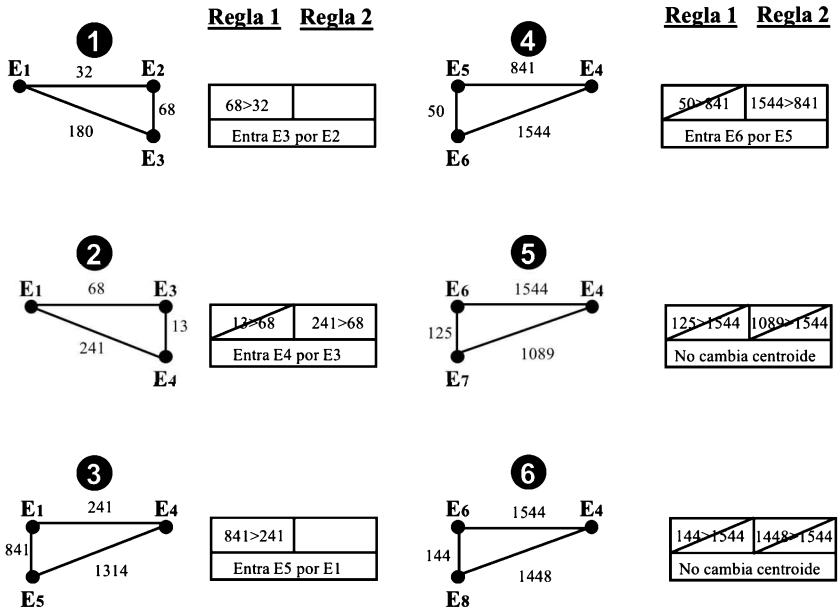


Figura 3.10.: Proceso de determinación del centroide inicial



© Ediciones Paraninfo

Estadísticos e-Books & Papers

y el centroide queda constituido por E1 y E3. Es el momento de probar con la siguiente observación del fichero (E4).

En el paso 2, la regla 1 establece que E4 sustituirá a la más cercana del par del centroide (E3) si la distancia a ella (13) es mayor que lo que distan entre sí el par más cercano de las observaciones que constituyen el centroide (68). Como esta regla no se cumple, se intenta comprobar si se cumple, la segunda y E4 puede entrar. Esta segunda regla indica que E4 sustituirá a la observación del centroide que tenga más cerca si la distancia más pequeña entre E4 y el resto de observaciones del centroide (sin contar la más cercana, esto es, respecto a E1: 241) es mayor que la distancia más pequeña entre la observación del centroide más cercana (E3) y las demás integrantes del mismo (E1: 68). Como es así, entonces E4 entra en el centroide, que quedará formado por E1 y E4.

El proceso se repite con todas las observaciones del fichero. Obsérvese cómo, a partir del paso 4, el centroide no cambia, quedando integrado por las observaciones E4 y E6. Esta es la semilla inicial que toma el análisis de conglomerados no jerárquico para comenzar a constituir los grupos.

Formación de los grupos `R`, mediante la función `kmeans{stats}`, a diferencia de SPSS y SAS, puede utilizar distintos algoritmos para forma los grupos. El que coincide con los mencionados programas es algoritmo de Lloyd (Lloyd, 1982), por este motivo, centraremos nuestra exposición en su desarrollo frente a procedimientos alternativos, pero también pueden aplicarse los algoritmos de Hartigan y Wong (1979) —que es el que se aplica por defecto—, y los de MacQueen (1967) o Forgy (1965).

En primer lugar, se calcula la distancia de cada observación a los k centroides iniciales calculados en la fase anterior. Cada observación se asigna al conglomerado al que esté más cercana. Para los datos del ejemplo, el cuadro 3.18 recoge la asignación efectuada con este criterio (las distancias son euclídeas).

Una vez efectuada la asignación de observaciones a conglomerados, se recalculan los centroides (centroides finales en el cuadro 3.18) y se repite el paso anterior, clasificando cada observación en el conglomerado del que dista menos.

En este caso, no se produce reasignación alguna de observaciones a conglomerados y el proceso se detiene. El conglomerado 1 estará formado por las observaciones [E5,E6,E7,E8] y el segundo por [E1,E2,E3,E4]. De haber habido reasignaciones, el proceso se hubiera continuado hasta que ninguna observación cambiara de conglomerado o hasta que se alcanzase un determinado número de iteraciones que se puede establecer como opción al ejecutar el análisis.

La sintaxis para pedir la estimación es la siguiente:

```
#efectuamos el cluster con metodo centroide
kmeans.caso3.1<-kmeans(DatosCaso3.1b, 2)
#obtenemos las medias
aggregate(DatosCaso3.1b,by=
list(kmeans.caso3.1$cluster),FUN=mean)
#adicionamos la pertenencia al cluster
```


Estadísticos e-Books & Papers

Cuadro 3.18.: Asignación de observaciones en el primer paso

Observación	Publicidad	Ventas	Distancia centroide		Conglo-merado
			1	2	
E1	16	10	34,37	**15,52**	2
E2	12	14	38,01	**11,00**	2
E3	10	22	40,61	**3,61**	2
E4	12	25	39,29	**0,00**	2
E5	45	10	**7,07**	36,25	1
E6	50	15	**0,00**	39,29	1
E7	45	25	**11,18**	33,00	1
E8	50	27	**12,00**	38,05	1

Conglomerado	Centroides iniciales		Centroides finales	
	Publicidad	Ventas	Publicidad	Ventas
1	50	15	47,5	19,25
2	12	25	12,5	17,75

Cuadro 3.19.: Asignación de observaciones en el segundo paso

Observación	Publicidad	Ventas	Distancia centroide		Conglo-merado
			1	2	
E1	16	10	32,83	**8,50**	2
E2	12	14	35,89	**3,78**	2
E3	10	22	37,60	**4,93**	2
E4	12	25	35,96	**7,27**	2
E5	45	10	**9,58**	33,41	1
E6	50	15	**4,93**	37,60	1
E7	45	25	**6,27**	33,30	1
E8	50	27	**8,14**	38,62	1

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 3.20.: Centroides finales

```
K-means clustering with 2 clusters of sizes 4, 4

Cluster means:
  inversion ventas
1      47.5  19.25
2      12.5  17.75

Clustering vector:
[1] 2 2 2 2 1 1 1 1

Within cluster sum of squares by cluster:
[1] 221.75 163.75
 (between_SS / total_SS =  86.4 %)
```

```
DatosCaso3.1b <-
data.frame(DatosCaso3.1b, kmeans.caso3.1$cluster)
```

La salida del programa, que se observa en el cuadro 3.20, indica a qué conglomerado se ha asignado cada observación y los centroides finales que, como puede verse, coinciden con los que ya dedujimos en el cuadro 3.18. Esta información es fundamental para caracterizar a los conglomerados obtenidos. La misión del analista no es solo determinar qué observaciones van a cada conglomerado, sino obtener las características de los mismos. El cuadro 3.20 nos indica que hay dos tipos de empresas que se diferencian porque unas (conglomerado 1) necesitan mucha más inversión publicitaria para alcanzar niveles similares de ventas, esto es, obtienen mucha menor rentabilidad de su inversión que las del conglomerado 2.

En el ejemplo que hemos venido utilizando, el número de observaciones en cada conglomerado es pequeño y la media de cada variable en los dos conglomerados es información suficiente para caracterizarlos. Sin embargo, si contásemos con muchas más observaciones tendría interés tratar de determinar qué variables toman valores medios claramente distintos en los distintos conglomerados y utilizar solo esas variables para efectuar la caracterización. R permite obtener el resultado de efectuar una serie de análisis de varianza donde el factor es la pertenencia al conglomerado y las variables dependientes son, sucesivamente, cada una de las utilizadas para caracterizar a los grupos.

```
library(psych)
fit.inversion<-aov(inversion ~ kmeans.caso3.1.cluster,
```


Estadísticos e-Books & Papers

Cuadro 3.21.: Análisis de varianza sobre los conglomerados finales

```
> summary(fit.inversion)
                        Df Sum Sq Mean Sq F value   Pr(>F)
kmeans.caso3.1.cluster  1   2450  2450.0   334.1 1.73e-06 ***
Residuals               6     44     7.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(fit.ventas)
                        Df Sum Sq Mean Sq F value Pr(>F)
kmeans.caso3.1.cluster  1    4.5    4.50   0.079  0.788
Residuals               6  341.5   56.92
```

```
data = DatosCaso3.1b)
fit.ventas<-aov(ventas ~ kmeans.caso3.1.cluster,
data = DatosCaso3.1b)
summary(fit.inversion)
summary(fit.ventas)
```

Del resultado que se ofrece en el cuadro 3.21 se observa que las diferencias entre las inversiones publicitarias de los dos grupos son muy grandes, pero no así las ventas. Este resultado confirmaría la interpretación de los conglomerados que expusimos anteriormente.

3.6. Elección entre los distintos tipos de análisis de conglomerados

Lo expuesto hasta el momento puede generar una duda en el lector. Existen dos grandes enfoques en el análisis de conglomerados (jerárquicos y no jerárquicos) y, dentro de los jerárquicos existen distintos métodos de conglomeración, pero ¿cuál ofrece mejores resultados? ¿cuál es más adecuado para los objetivos de una investigación determinada?

Responder a estas preguntas no es sencillo y, por ello, el lector no puede esperar una respuesta categórica, por cuanto la misma depende de los objetivos del estudio y de las propiedades de los distintos métodos (Hair *et al.*, 2014a). Intentaremos, sin embargo, dar algunas indicaciones.

3.6.1. Elección entre análisis de conglomerados jerárquico y no jerárquico

Bajo nuestro punto de vista esta decisión no debe tomarse en términos disyuntivos, pues consideramos que un enfoque es un buen complemento del otro. Si el investigador tiene una presunción razonable de cuál puede ser el número de

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

grupos naturales en que se agregan sus observaciones, el análisis no jerárquico sería una buena opción. Sin embargo, este enfoque requiere que se suministren los centroides iniciales de esos grupos y estos rara vez están disponibles. Existen diversos trabajos (Milligan, 1980) que demuestran que el resultado final de un análisis de conglomerados no jerárquico depende de lo cercano a la realidad que sea la semilla inicial, no siendo siempre recomendable que el ordenador la elija aleatoriamente.

Pues bien, la mejor forma de obtener una buena aproximación de cuál es el número razonable de conglomerados (si el investigador no tiene ninguna opinión a priori) y de conseguir simultáneamente una semilla fiable, pasa por efectuar, en primer lugar, un análisis de conglomerados jerárquico, utilizar las herramientas que se han ofrecido para seleccionar el número de grupos, y alimentar con esta información la realización de un análisis de conglomerados no jerárquico que nos permitirá maximizar la homogeneidad dentro de cada grupo y la heterogeneidad entre unos conglomerados y otros.

Sin embargo, como señalan Hair *et al.* (2014b), los métodos jerárquicos no siempre son de posible aplicación cuando los tamaños muestrales son muy altos y/o con numerosas variables y eso independientemente del incremento en la potencia de los ordenadores. Por ejemplo, una muestra de 400 casos necesita del cálculo de 80.000 distancias que se incrementan a 125.000 cuando son 500.

3.6.2. Elección entre los distintos métodos de agrupación en el análisis de conglomerados jerárquico

Aunque se han realizado numerosos estudios comparando los distintos procedimientos de agrupación, los resultados a los que se ha llegado no son, en modo alguno, categóricos. Ello nos lleva a ser partidarios de ensayar varios de estos procedimientos en un mismo estudio. Si los resultados son coherentes, habremos dado con agrupaciones naturales, si no es así, habrá que elegir entre los distintos resultados reteniendo aquel que le parezca más razonable al investigador o esté de acuerdo con trabajos previos efectuados.

Pese a lo expuesto, conviene sintetizar algunos de los resultados que se han obtenido en estos estudios. Sharma (1996) resume el trabajo de Punj y Stewart (1983) en los siguientes términos:

1. El método del vecino más cercano es más sensible a la presencia de observaciones anómalas (*outliers*) que el método del vecino más lejano.

2. El método del vecino más lejano identifica habitualmente grupos muy homogéneos, en los que las observaciones son muy parecidas unas a otras.

3. El método de Ward tiende a encontrar conglomerados no solo muy compactos, sino también de tamaño similar.

4. El método del vecino más cercano tiene tendencia a crear menos grupos que el del vecino más lejano (Johnson, 1998).

Estadísticos e-Books & Papers

3.7. Un ejemplo de aplicación del análisis de conglomerados

El objetivo de este epígrafe es ofrecer una visión integrada de los pasos que requiere la aplicación de un análisis de conglomerados, desde el establecimiento de los objetivos hasta la validación de los resultados.

Caso 3.3. Diseño de un plan de incentivos para vendedores
El director de ventas de una cadena de tiendas de electrodomésticos con implantación nacional está estudiando el plan de incentivos de sus vendedores. Considera que los incentivos deben estar ajustados a las dificultades de las distintas zonas de ventas, siendo necesario fijar incentivos más altos en aquellas zonas geográficas en que las condiciones de vida de sus habitantes hacen más difícil las ventas. Por este motivo quiere determinar si las comunidades autónomas se pueden segmentar en grupos homogéneos respecto al equipamiento de los hogares.

Para ello dispone de los datos que aparecen en el cuadro 3.22 y el objetivo es establecer cuántos grupos de comunidades autónomas con niveles de equipamiento similar pueden establecerse y en qué radican las diferencias entre esos grupos. El procedimiento que aplicaremos es el descrito en el tema, a saber:

1. Análisis de la existencia de *outliers* en la medida en que pueden generar importantes distorsiones en la detección del número de grupos.

2. Realización de un análisis de conglomerados jerárquicos, evaluando la solución de distintos métodos de conglomeración, aplicando los criterios presentados para identificar el número adecuado de grupos y obtención de los centroides que han de servir de partida para el paso siguiente.

3. Realización de un análisis de conglomerados no jerárquico mediante el método de k-medias para la obtención de una solución óptima en términos de homogeneidad intrasegmentos y heterogeneidad intersegmentos.

En el capítulo 2 describimos el procedimiento para detectar los *outliers* mediante la distancia de Mahalanobis. Aplicando este procedimiento mediante `R` a nuestros datos obtendríamos las siguientes distancias y, respectivas significatividades (cuadro 3.23). A la luz de esta información no cabría considerar a ninguna comunidad autónoma como un valor atípico.

Estimaríamos ahora los conglomerados por el procedimiento jerárquico. Como debemos evaluar la coherencia de los resultados para decantarnos o descartar alguno de ellos, lo hacemos por todos los métodos de conglomeración que hemos visto —entre paréntesis el nombre que tiene la función de la función `hclust{stats}`— centroide (*centroid*), vecino más cercano, (*single*), vecino más lejano (*complete*), promedio (*average*), y Ward (*ward.D2*). La sintaxis, para todos ellos, con la modificación del método sería:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 3.22.: Equipamiento de los hogares en distintas comunidades autónomas

CC.AA.	Porcentaje de hogares que poseen					
	Auto-móvil	TV color	Vídeo	Micro-ondas	Lava-vajillas	Teléfono
España	69,0	97,6	62,4	32,3	17,0	85,2
Andalucía	66,7	98,0	82,7	24,1	12,7	74,7
Aragón	67,2	97,5	56,8	43,4	20,6	88,4
Asturias	63,7	95,2	52,1	24,4	13,3	88,1
Balerares	71,9	98,8	62,4	29,8	10,1	87,9
Canarias	72,7	96,8	68,4	27,9	5,80	75,4
Cantabria	63,4	94,9	48,9	36,5	11,2	80,5
Castilla y León	65,8	97,1	47,7	28,1	14,0	85,0
Cast.-La Mancha	61,5	97,3	53,6	21,7	7,10	72,9
Cataluña	70,4	98,1	71,1	36,8	19,8	92,2
Com. Valenciana	72,7	98,4	68,2	26,6	12,1	84,4
Extremadura	60,5	97,7	43.7	20,7	11,7	67,1
Galicia	65,5	91,3	42,7	13,5	14,6	85,9
Madrid	74,0	99,4	76,3	53,9	32,3	95,7
Murcia	69,0	98,7	59,3	19,5	12,1	81,4
Navarra	76,4	99,3	60,6	44,0	20,6	87,4
País Vasco	71,3	98,3	61,6	45,7	23,7	94,3
La Rioja	64,9	98,6	54,4	44,4	17,6	83,4

Fuente: Panel de Hogares de la Unión Europea. INE.

Estadísticos e-Books & Papers

Cuadro 3.23.: Resultados de la detección de *outliers*

CC.AA.	D^2	p-value $\chi^2(df = 6)$
España	0,40	0,99
Andalucía	3,93	0,68
Aragón	1,94	0,92
Asturias	4,46	9,61
Balerares	6,02	0,42
Canarias	10,47	0,10
Cantabria	7,27	0,29
Castilla y León	3,25	0,77
Castilla-La Mancha	4,12	0,66
Cataluña	4,21	0,64
Com. Valenciana	2,85	0,82
Extremadura	0,29	0,15
Galicia	13,30	0,03
Madrid	9,49	0,14
Murcia	4,61	0,59
Navarra	9,58	0,14
País Vasco	2,55	0,86
La Rioja	4,25	0,64

© Ediciones Paraninfo

Estadísticos e-Books & Papers

```
hclust.average.caso3<-hclust(matriz.dis.euclid.caso3,
method="average")
data.frame(hclust.average.caso3[2:1])
```

La figura 3.11 recoge los dendogramas de todas estas estimaciones. Vemos claramente dos patrones: el correspondiente a los métodos *centroid* y *single,* que agregan a todas las comunidades —salvo a Madrid— en un mismo grupo pero con claras distorsiones en el dendograma, y los otros tres procedimientos —Ward, *complete* y *average*— que generan una solución muy parecida. Hemos de establecer cuántos grupos, pero los dendogramas nos permiten intuir que las comunidades que acabarán en cada grupo van a ser las mismas independientemente del método de conglomeración empleado. Pasamos a la siguiente fase —determinación del número de conglomerados— solo para las tres técnicas que no nos generan dudas respecto a la claridad de los dendogramas: Ward, *complete* y *average.*

Aplicamos para ello el procedimiento descrito de generación de índices propuestos por **{NbClust}** y que se describieron en el apartado 3.4 mediante la sintaxis:

```
Datos.NbClust<-Datos_3_3_Caso[,c("automovi","tvcolor","video",
"microond","lavavaji","telefono")]

res.wardD2<-NbClust(Datos.NbClust, distance = "euclidean",
min.nc=2, max.nc=15, method = "ward.D2", index = "alllong")
```

Solo cabría sustituir el método de conglomeración para obtener la mejor solución para cada uno de ellos. La aplicación de los criterios es consistente en sus resultados y para cualquier método de conglomeración la solución adecuada es la de 2 conglomerados. Ilustramos la consistencia de los resultados con la evolución para distintos conglomerados del criterio CCC (Sarle, 1983), que, recordemos, alcanzaba su valor máximo para el número óptimo de grupos, que siempre es 2, tal y como se aprecia en la figura 3.12.

Decidido trabajar con una solución de dos grupos que, como mostramos en la figura 3.13, agrupa a las mismas comunidades en los mismos grupos independientemente del método de conglomeración, el paso siguiente es obtener los centroides (valores medios de las 6 variables en cada uno de los dos grupos) con el fin de alimentar con ello el método no jerárquico. Es recomendable dejar que el método no jerárquico los obtenga de manera aleatoria, en la medida en que un punto de partida sensato —y el resultado del jerárquico lo es— aumenta las probabilidades de que el proceso de optimización no se estanque en un mínimo local (Milligan, 1980). A partir de este momento ilustraremos el resultado solo con los datos obtenidos del método de conglomeración de Ward.

Generamos una variable que contiene la pertenencia al grupo y la añadimos a la base de datos:

Estadísticos e-Books & Papers

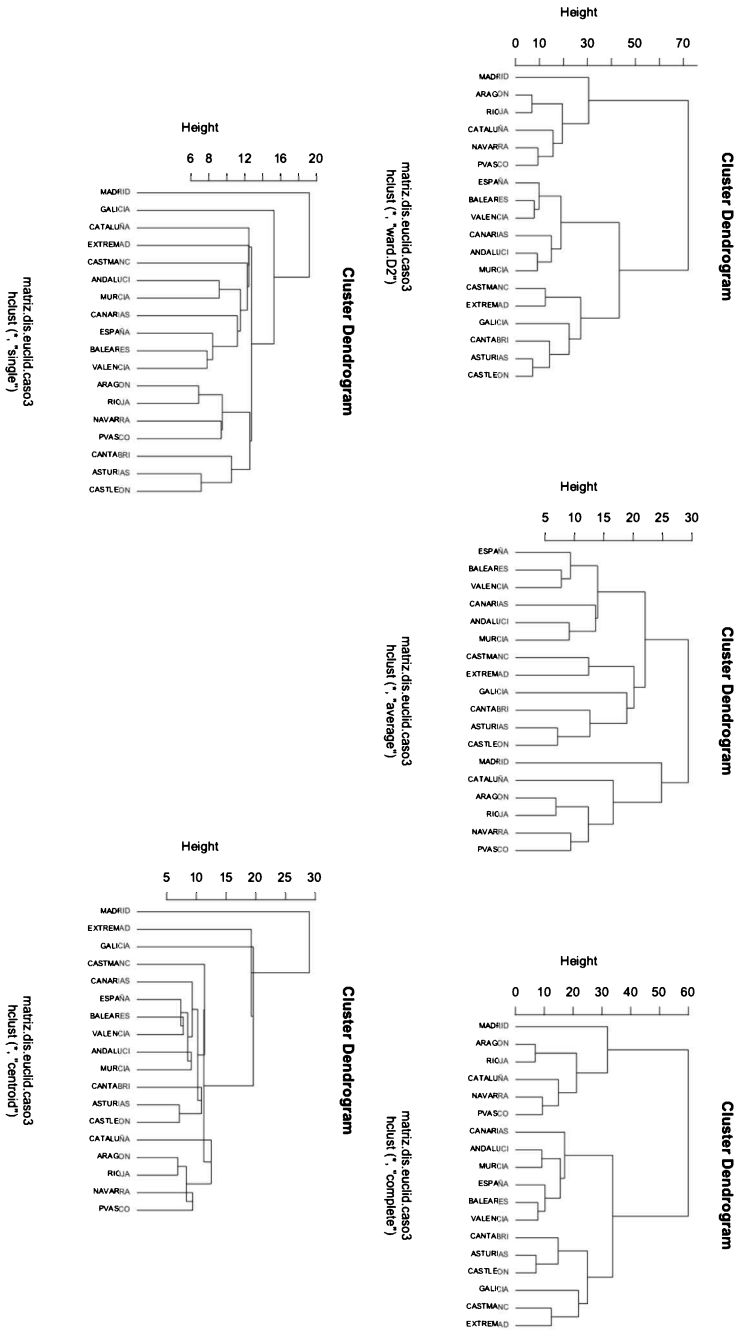
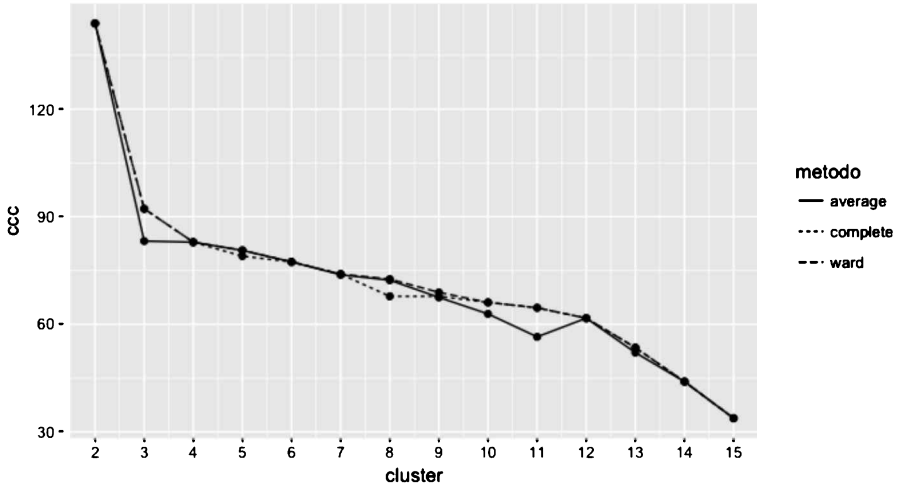


Figura 3.11.: Dendogramas mediante resultantes de aplicar distintos métodos de conglomeración

© Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 3.12.: Proceso de determinación del centroide inicial



Cuadro 3.24.: Centroides resultantes del método jerárquico

	Group.1	automovi	tvcolor	video	microond	lavavaji	telefono	grupo.ward
1	1	66.87	96.82	56.01	25.43	11.81	80.71	1
2	2	70.70	98.53	63.47	44.70	22.43	90.23	2

```
grupo.ward<-cutree(hclust.ward.caso3, k = 2, h = NULL)
datos.caso3.grupos<-cbind(Datos_3_3_Caso,grupo.ward)
datos.caso3.grupos$id<-NULL
```

A continuación obtenemos los centroides, que no es otra cosa que la media de las seis variables analizadas en cada uno de los dos grupos obtenidos. El resultado se ofrece en el cuadro 3.24:

```
round(aggregate(datos.caso3.grupos,list(grupo.ward), mean ),2)
```

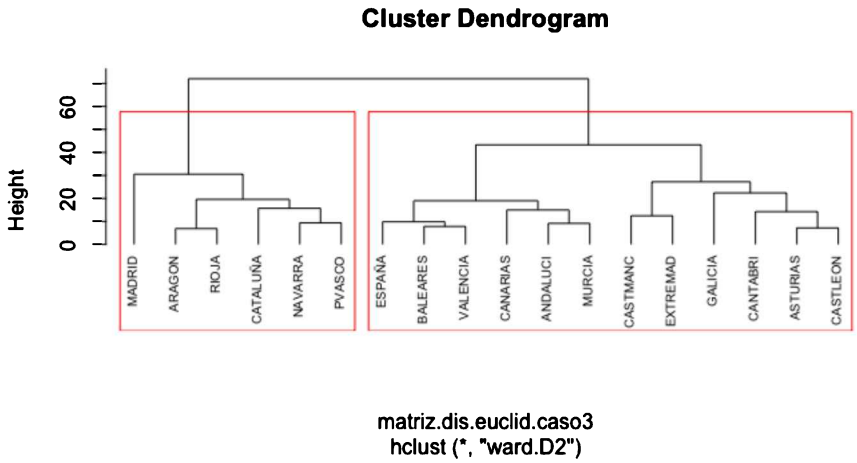
Solo resta estimar el análisis de conglomerados no jerárquico tomando como centroides iniciales los anteriores mediante la función **kmeans{stats}**.

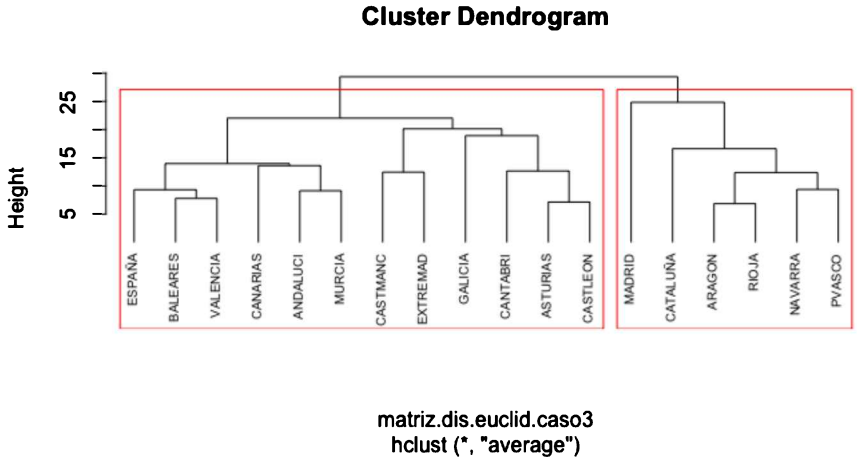
```
c1<-c(66.87,96.82,56.01,25.43,11.81,80.71)
c2<-c(70.70,98.53,63.47,44.70,22.43,90.23)
solucion<-kmeans(datos.caso3.grupos.kmeans,rbind(c1, c2))
```

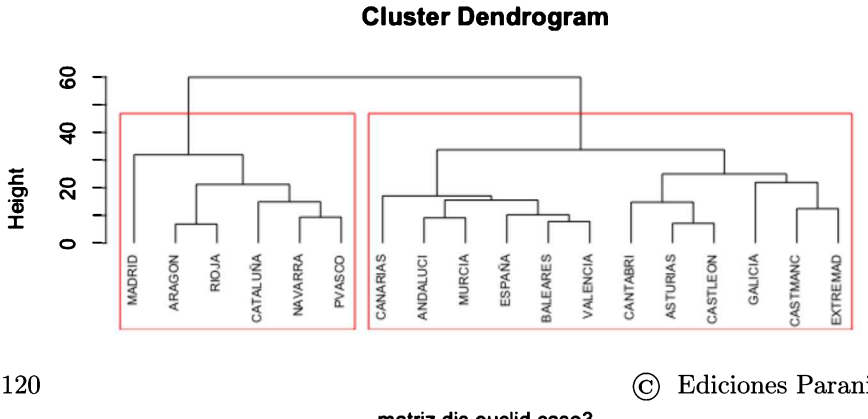
El cuadro 3.25 recoge la salida, donde, como se puede comprobar, al ser un caso con pocos datos, la solución del jerárquico coincide con la del no jerárquico.

Estadísticos e-Books & Papers

Figura 3.13.: Dendogramas mediante resultantes de aplicar distintos métodos de conglomeración







 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 3.25.: Centroides resultantes del método no jerárquico

```
K-means clustering with 2 clusters of sizes 12, 6

Cluster means:
  automovi  tvcolor    video microond lavavaji telefono
1 66.86667 96.81667 56.00833   25.425 11.80833 80.70833
2 70.70000 98.53333 63.46667   44.700 22.43333 90.23333

Clustering vector:
 [1] 1 1 2 1 1 1 1 1 1 2 1 1 1 2 1 2 2 2

Within cluster sum of squares by cluster:
[1] 2176.3133  848.3533
 (between_SS / total_SS =  46.2 %)

      Welch Two Sample t-test

  data:  automovi by solucion.cluster
  t = -1.8106, df = 10.091, p-value = 0.1
  alternative hypothesis: true difference in means is not equal to 0
  95 percent confidence interval:
   -8.5449256  0.8782589
  sample estimates:
  mean in group 1 mean in group 2
         66.86667        70.70000
```

Antes de interpretar las medias, conviene saber cuáles son significativamente diferentes entre los grupos. En temas posteriores se presentará el análisis de la varianza de un factor, que sería la herramienta que conviene aplicar para este fin si tuviéramos más de dos grupos. Al tener solo dos grupos, podemos aplicar una prueba t, con la siguiente sintaxis para cada variable dependiente:

```
t.test(automovi~solucion.cluster,
data=datos.caso3.grupos.kmeans)
```

En el ejemplo ilustrado vemos como el porcentaje de la población que tiene automóvil en el primer grupo de comunidades autónomas es (66,86 %) inferior al del segundo grupo (70,70 %), aunque esta diferencia no es significativa ($t = -1{,}81; p > 0{,}05$). Para facilitar el análisis general, llevamos todos los resultados al cuadro 3.26. A la luz del mismo podemos concluir que el grupo 2 se corresponde con comunidades autónomas donde los equipamientos son significativamente superiores, probablemente debido a una mayor renta per cápita.

Aunque veremos la herramienta con mayor detalle en el tema correspondiente, una manera de intentar visualizar los resultados de un análisis de conglomerados, más allá de los dendogramas, es sintetizar todas las variables en dos componentes principales y proyectar sobre ellos los objetos. La figura 3.14 muestra el claro nivel de separación entre los dos grupos que ya hemos interpretado.

Estadísticos e-Books & Papers

Cuadro 3.26.: Significatividad de las diferencias entre los perfiles de los conglomerados

Variable	Grupo de CC.AA.		Prueba t
	Grupo 1	Grupo 2	
Automóvil	66,86	70,70	1,81
TV color	96,82	98,53	2,51*
Vídeo	56,01	63,46	1,71
Microondas	25,43	44,70	6,73**
Lavavajillas	11,81	22,43	4,61**
Teléfono	80,71	90,23	3,50**

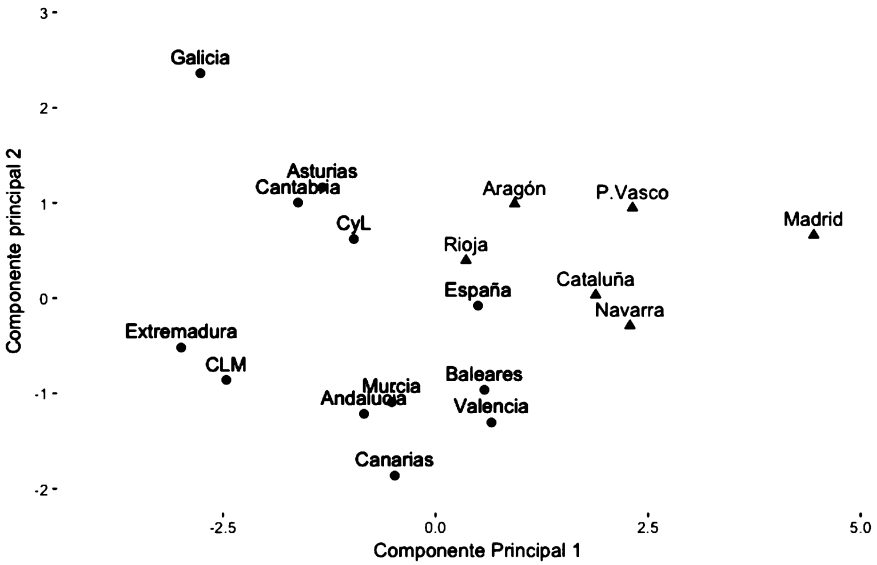
$^{**}p < 0,01;^{*}\, p < 0,05$

En el tema correspondiente veremos cómo interpretar los ejes.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 3.14.: Visualización de los resultados de un análisis de conglomerados



Estadísticos e-Books & Papers

Estadísticos e-Books & Papers

4. Escalamiento multidimensional

4.1. Introducción

El análisis de escalamiento multidimensional (MDS) —*multidimensional scaling*— es una técnica de reducción de datos como otras que se presentarán más adelante: análisis factorial o análisis de componentes principales, por ejemplo. El objetivo principal del MDS es representar N objetos en un espacio dimensional reducido (q dimensiones, siendo $q < N$), de tal forma que la distorsión causada por la reducción de la dimensionalidad sea la menor posible, es decir, que las distancias entre los objetos representados en el espacio q dimensional sean lo más parecidas posible a las distancias en el espacio N dimensional.

Dado que será difícil que las distancias coincidan, el objetivo del MDS es conseguir que ambas configuraciones dimensionales sean lo más parecidas posible. Para ello será necesario construir un indicador de esa proximidad que, como se detallará más adelante, denominaremos *stress* o *sstress*.

Intentemos ofrecer una visión más intuitiva de las posibilidades que tiene la técnica MDS. Utilizaremos para ello un ejemplo que, por su claridad, es utilizado con frecuencia en los textos que abordan de manera introductoria esta técnica (Dillon y Goldstein, 1984; Johnson y Wichern, 1998; Kruskal y Wish, 1978). Supongamos que se nos da un mapa con la localización de distintas ciudades españolas y se nos pide que construyamos una matriz con las distancia entre todas ellas. Es bastante sencillo rellenar cada elemento de esta matriz, simplemente midiendo con una regla y aplicando la correspondiente escala del mapa.

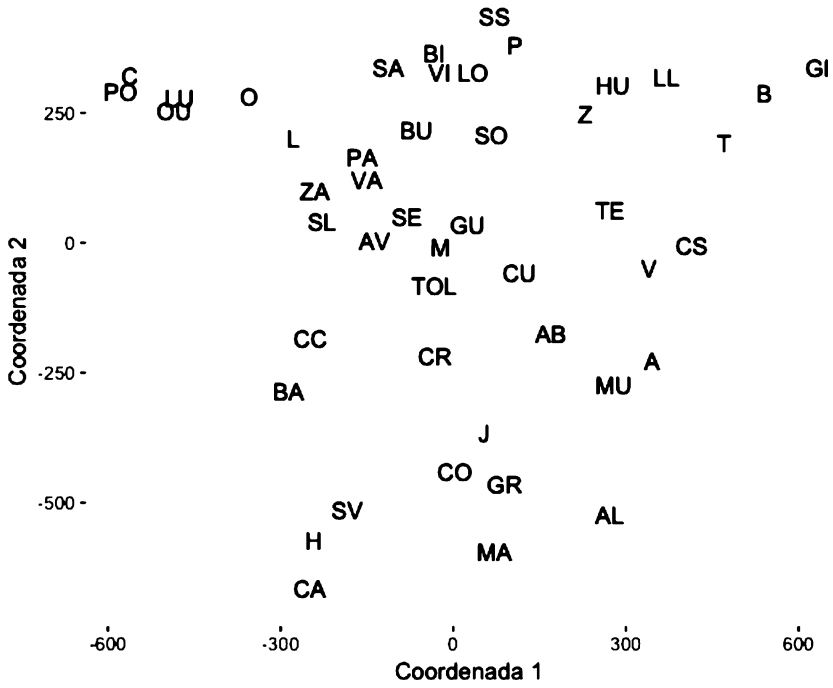
Consideremos ahora, sin embargo, el problema inverso. Se nos da la matriz de distancias entre las ciudades y se nos pide que reproduzcamos el mapa. La solución ahora es mucho más difícil. Podremos situar una ciudad y colocar otra a la distancia que nos indique la matriz. Podremos, incluso, colocar una tercera a la distancia correcta de las otras dos, pero colocar la cuarta ya será mucho más complicado porque habrá que recolocar las otras tres... y así hasta 40 o 50 ciudades. En esencia, MDS es una técnica que permite resolver este tipo de problema inverso.

Ilustración 4.1

La figura 4.1 es el resultado de aplicar un MDS a la matriz de distancias entre las capitales de provincia españolas. En este momento todavía es pronto para ofrecer la sintaxis del paquete `mds{smacof}` que ha efectuado el análisis

Estadísticos e-Books & Papers

Figura 4.1.: Ilustración de la aplicación del MDS a la distancia entre las capitales españolas de provincia



—aunque la misma y la base de datos acompañan al manual—, lo importante es que el lector compruebe como, a partir de la matriz de distancia entre capitales de provincias, la herramienta es capaz de reproducir con bastante nitidez y un mínimo de distorsión el mapa de España. Bien cierto es que la solución natural de este problema es una solución de dos dimensiones, que son las mismas que ofrece el mapa y que, cuando un mapa de dos dimensiones intente sintetizar soluciones naturales de un número mayor de ellas, la distorsión será menor, pero para eso tendremos indicadores como los mencionados *stress* y *sstress*.

Caso 4.1. Valoración de la imagen de superficies comerciales

Ilustremos ahora, con un ejemplo numérico y gráfico, las ideas básicas en las que se fundamenta el MDS. Supongamos que hemos pedido a 100 consumidores que valoren la imagen que tienen de 5 superficies comerciales, atendiendo a la similitud con que las perciben. Para ello se utiliza una escala de 0 (idénticas) a 7 (totalmente diferentes). La siguiente matriz de *disparidades originales*[1] —o

[1] A veces se utiliza también la denominación de distancias. No obstante, con objeto de evitar confusiones al lector, el término distancias será utilizado exclusivamente al describir los algoritmos (sección 4.2 y apéndice) para aquellas medidas obtenidas mediante un algoritmo matemático (distancia ecuclídea, por ejemplo). Posteriormente, en la sección 4.4 veremos

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

proximidades— nos muestra las medias de las puntuaciones ofrecidas por los 100 consumidores.

$$S = \begin{bmatrix} & X_1 & X_2 & X_3 & X_4 & X_5 \\ X_1 & 0,0 & & & & \\ X_2 & 1,0 & 0,0 & & & \\ X_3 & 2,1 & 2,4 & 0,0 & & \\ X_4 & 6,1 & 6,9 & 5,1 & 0,0 & \\ X_5 & 5,2 & 5,3 & 4,1 & 3,1 & 0,0 \end{bmatrix} \tag{4.1}$$

Nótese que en la diagonal de la matriz aparecen ceros porque la imagen de una superficie comercial siempre ha de ser idéntica a sí misma. También se observa que solo se representa el triángulo inferior de la matriz al ser esta simétrica por construcción. Si creamos un mapa en dos dimensiones para ilustrar mejor la percepción de los consumidores, este mapa debería representar como puntos cercanos a las superficies X1 y X2 porque la disparidad entre ellas es pequeña (1,0), tal y como refleja la matriz. Asimismo, las superficies X2 y X4 deberían aparecer representadas muy distantes una de la otra, por cuanto su disparidad en la matriz es elevada (6,9).

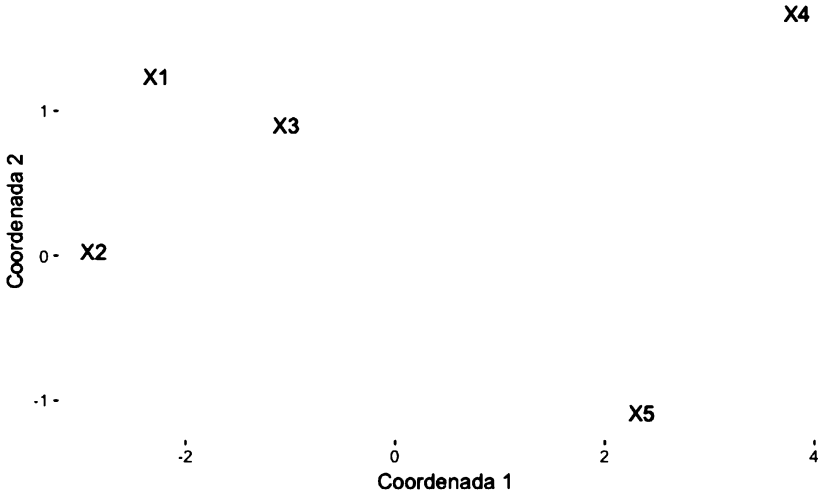
La figura 4.2 ofrece el mapa que se obtiene al representar las coordenadas bidimensionales resultantes de aplicar a la matriz anterior uno de los algoritmos que existen para efectuar un MDS, el implementado en `mds{smacof}`, en el que luego profundizaremos. Puede comprobarse como se constata la cercanía de X1 y X2 y la lejanía de X2 y X4 que esperábamos.

```
# Introduccion de la matriz de datos
datos<-matrix(c(
0.0,1.0,2.1,6.1,5.2,
1.0,0.0,2.4,6.9,5.3,
2.1,2.4,0.0,5.1,4.1,
6.1,6.9,5.1,0.0,3.1,
5.2,5.3,4.1,3.1,0.0),
ncol=5,nrow=5,byrow=T,
dimnames=list(c("X1","X2","X3","X4","X5")))

library(smacof)
fit2<-mds(delta=datos,ndim=2,type="interval")

# Coordenadas  print
(fit2$conf)
# Disparidades
print(fit2$dhat)
# Distancias entre configuraciones
```

algunos tipos de MDS en los que los datos iniciales se transforman de forma inmediata en distancias.

Estadísticos e-Books & Papers

Figura 4.2.: Ilustración de la aplicación del MDS a los datos de la imagen de cadenas de electrodomésticos




```
print(fit2$confdiss)
# Stress
print(fit2$stress)
# Stress por punto
print(fit2$spp)
#RSQ
print(1-fit2$rss)
```

Las dos configuraciones dimensionales, sin embargo, no son perfectamente equiparables, entre otras cosas porque, probablemente, la solución de dos dimensiones no sea la más adecuada para este problema. Esto hará que el indicador de calidad del MDS (*stress* o *sstress*) refleje este hecho.

4.2. El algoritmo básico del MDS

Cuando nos referimos al MDS no estamos hablando de una técnica sino de un conjunto de ellas. Como se verá a continuación, la resolución del problema que se plantea en el MDS se realiza a través de un algoritmo determinado. Distintos algoritmos suponen distintas soluciones al problema. En el presente epígrafe se planteará una visión general de la solución, mientras que en el apéndice a este capítulo ofreceremos el detalle de uno de los algoritmos más utilizados, el implementado en `mds{smacof}`.

En cualquier caso, si se desea una visión histórica de la evolución del MDS, desde el trabajo seminal de Torgeson (1952), el desarrollo de Shepard (1962)

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

y Kruskal (1964a; 1964b) hasta las contribuciones más recientes de Ramsay (1977; 1982) o Takane (1982), el lector puede remitirse al libro de Young y Hamer (1987). Si lo que se desea es una revisión de los distintos algoritmos y ejemplos de uso de los diferentes programas que los implementan, entonces el texto más adecuado es el de Green *et al.* (1989).

Ahora vamos a ilustrar el desarrollo formal de la técnica MDS con el ejemplo de las superficies comerciales expuesto con anterioridad. Si partimos de N objetos (superficies comerciales), tendremos entonces $M = N(N–1)/2$ disparidades originales entre pares de objetos (10 en nuestro ejemplo). Asumiendo que no haya empates (los distintos algoritmos resuelven los empates de distintos modos), las disparidades pueden escribirse en un orden estrictamente ascendente:

$$s_{i_1k_1} < s_{i_2k_2} < \ldots < s_{i_Mk_M}$$

donde $s_{i_1k_1}$ es la menor de las disparidades. El subíndice i_1k_1 indica el par de objetos que son más parecidos. En nuestro ejemplo esta ordenación sería la siguiente:

$$1,0 < 2,1 < 2,4 < 3,1 < 4,1 < 5,1 < 5,2 < 5,3 < 6,1 < 6,9$$
$$s_{12} < s_{13} < s_{23} < s_{45} < s_{35} < s_{34} < s_{15} < s_{25} < s_{14} < s_{24}$$

Nuestro objetivo es encontrar una nueva configuración q dimensional de los N objetos (2 dimensiones y 5 objetos en el ejemplo), de tal forma que las distancias calculadas entre ellos en ese espacio q dimensional mantengan la ordenación anterior. En el caso ideal de que se mantuviera el orden y las proporciones entre disparidades y distancias, el gráfico de dispersión entre ambas se representaría mediante una línea recta.

Pues bien, en el MDS se van ensayando distintas configuraciones q dimensionales hasta que las distancias en ese espacio y las disparidades originales guarden una relación lo más próxima posible a esta recta ideal.

El cuadro 4.1 ofrece la solución bidimensional final resultante de la aplicación del MDS a los datos de nuestro ejemplo. En este cuadro aparecen las coordenadas de cada objeto (superficie comercial) en ese espacio bidimensional, coordenadas cuya representación gráfica recogimos en la figura 4.2. A partir de esas coordenadas es sencillo derivar la matriz de distancia entre los distintos objetos. Así, tomando distancias euclídeas, la distancia entre, por ejemplo, X1 y X2 tomaría el valor:

$$d\left(X_1, X_2\right) = \sqrt{\left(-0{,}5504 - [-0{,}6431]\right)^2 + \left(0{,}0197 - [-0{,}0451]\right)^2} = 0{,}1131$$

Repitiendo los cálculos para todos los objetos (estímulos), obtendríamos la matriz de distancias **D** entre las configuraciones que muestra el cuadro 4.2.

En el algoritmo del MDS se obtiene, como hemos señalado, una transformación monótona de la matriz de distancias originales en el espacio N dimensional y es respecto a esa transformación (matriz de disparidades Δ) obtenida en ca-

Estadísticos e-Books & Papers

Cuadro 4.1.: Solución bidimensional

```
Configurations:
        D1      D2
X1 -0.5504  0.0197
X2 -0.6431 -0.0451
X3 -0.2234  0.1432
X4  0.8855  0.2016
X5  0.5314 -0.3194
```

Cuadro 4.2.: Distancias entre las configuraciones

```
          X1        X2        X3        X4
X2 0.1131206
X3 0.3496955 0.4601708
X4 1.4479391 1.5489607 1.1108364
X5 1.1341299 1.2065465 0.8855614 0.6302148
```

da iteración con la que se va comparando la matriz **D**. En nuestro ejemplo, la solución final bidimensional alcanzada en la última iteración proporciona la matriz de disparidades Δ que aparece en el cuadro 4.3.

Así pues, la matriz de disparidades Δ es simplemente una transformación monótona de la matriz de distancias originales **S**. Esto se puede comprobar representando, simplemente, en un gráfico de dispersión las distancias que aparecen en ambas (figura 4.3). Sobre este gráfico, conocido como diagrama de Shepard, volveremos posteriormente. En este momento nos interesa centrarnos en los puntos sobre la diagonal que muestran la transformación monótona y no en las discrepancias que se producen (puntos grises), que serán un indicador de la bondad del ajuste del modelo que veremos posteriormente. Este gráfico se ha solicitado añadiendo a la sintaxis mostrada con anterioridad mediante la instrucción:

```
# Grafico de Shepard

plot(fit2, plot.type = "Shepard", plot.dim = c(1,2),
```

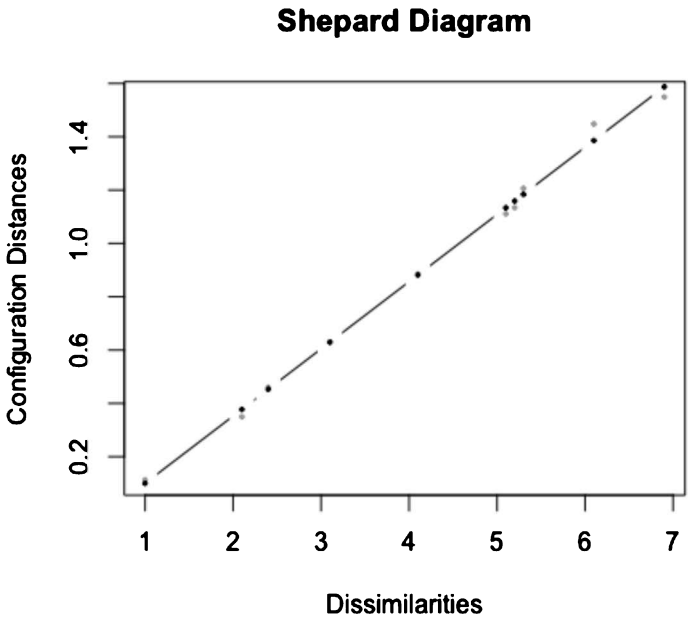
Cuadro 4.3.: Matriz de disparidades

```
          X1        X2        X3        X4
X2 0.1001109
X3 0.3773818 0.4530011
X4 1.3856397 1.5872913 1.1335752
X5 1.1587817 1.1839881 0.8815108 0.6294463
```

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 4.3.: Diagrama de Shepard




```
sphere = TRUE, bubscale = 0.1, col = 1,
label.conf = list(label = TRUE,
pos = 3, col = 1, cex = 0.8),
shepard.x = NULL, identify = FALSE,
type = "p", pch = 20, asp = 1, col.hist = NULL)
```

El cuadro 4.1 muestra la solución final que se alcanza con la técnica MDS. Sin embargo, para llegar a esta solución final, el algoritmo habrá ido ensayando distintas configuraciones bidimensionales hasta dar con aquella que reduce en mayor grado las diferencias entre las matrices de distancias $\mathbf{D}$ y disparidades Δ. Para ello necesitamos una función objetivo que se minimizará en cada iteración. Kruskal (1964a) propuso la siguiente función, que denominó *stress*:

$$Stress = \sqrt{\frac{\sum_{i \neq j} (d_{ij} - \delta_{ij})^2}{\sum_{i \neq j} d_{ij}^2}} \tag{4.2}$$

donde d_{ij} son los elementos de la matriz de distancias resultante de la solución

Estadísticos e-Books & Papers

Cuadro 4.4.: Interpretación del *stress* en términos de bondad de ajuste del MDS

Stress	Bondad de ajuste
0.200	Malo
0.100	Mínimo razonable
0.050	Bueno
0.025	Excelente
0.000	Perfecto

Fuente: Kruskal (1964a).

q dimensional en la interacción que se esté realizando y δ_{ij} son los elementos de la matriz de disparidades que, recordemos, no son sino una transformación monótona de los elementos de la matriz de disparidades originales entre los distintos objetos (estímulos). En síntesis, el *stress* no es sino un indicador de cuánto difieren en promedio la matriz con las distancias de la solución dimensional reducida respecto a la matriz con las disparidades originales. El cuadrado del numerador pretende, únicamente, que no se compensen diferencias positivas con negativas.

El valor del *stress* deberá ser tan pequeño como sea posible y, en todo caso, reducirse en cada iteración. De no ser así, el algoritmo se detendrá. Kruskal (1964a) sugiere que el *stress* de la solución final debería ser interpretado en los términos del cuadro 4.4.

Una segunda medida de las discrepancias entre las matrices de disparidades y distancias y, por ello, de calidad de la representación lograda por el MDS, es el estadístico denominado *s-stress* que fue propuesto por Takane *et al.* (1977), autores del algoritmo ALSCAL y que, por ello, es la función que se minimiza en ese algoritmo:

$$S - stress = \sqrt{\frac{\sum_{i \neq j} (d_{ij} - \delta_{ij})^2}{\sum_{i \neq j} d_{ij}^4}} \tag{4.3}$$

El valor del *s-stress* está siempre comprendido entre 0 y 1 y cualquier valor inferior a 0,1 indica que la solución obtenida es una buena representación de los objetos de la solución N dimensional inicial.

El cálculo de los valores del *stress* y del *s-stress* para la solución final de los datos de nuestro ejemplo es sencillo, por cuanto en el cuadro 4.2 tenemos la matriz de disparidades Δ que proporciona los elementos δ_{ij} de las expresiones (4.2) y (4.3) y en el cuadro 4.3 tenemos la matriz $\mathbf{D}$ que proporciona las distancias d_{ij}. El cuadro 4.5 muestra la salida de `smacof`, donde se proporciona el valor final del *stress*, dado que este paquete no optimiza, y, por ello, no calcula el *s-stress*. Se comprueba que, el *stress* (0,0282 <0,05) obtenido alcanza el valor de "bueno", de acuerdo con la valoración de Kruskal presentada en el

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 4.5.: Valor del *stress* y otros indicadores de ajuste para los datos del ejemplo

```
> print(fit2$stress)
[1] 0.02826073

> print(1-fit2$rss)
[1] 0.9920117

> print(fit2$spp)
       X1        X2        X3        X4        X5
33.954189 13.762478  8.458632 36.729371  7.095329
```

cuadro 4.4, estando muy cerca de ser "excelente".

El paquete `smacof` permite también analizar la contribución de cada punto representado al *stress*, es decir, del total del desajuste, qué parte es debida a una mala representación de un punto determinado. En el caso del ejemplo, vemos en el cuadro 4.5 que el punto peor representado sería el X4 que supone el 36,7 % del total del *stress*. Para visualizar esta información `smacof` puede representar sobre el mapa un círculo con un radio proporcional al tamaño de la proporción de *stress* de cada punto (`plot.type="bubbleplot"`) o, directamente, una barra para cada punto con el porcentaje de *stress* del cuadro 4.5 (`plot.type="stressplot"`), como se aprecia en la figura 4.4. En ambos casos, cuanto mayor es la barra o la burbuja, peor es el ajuste o representación de ese punto.

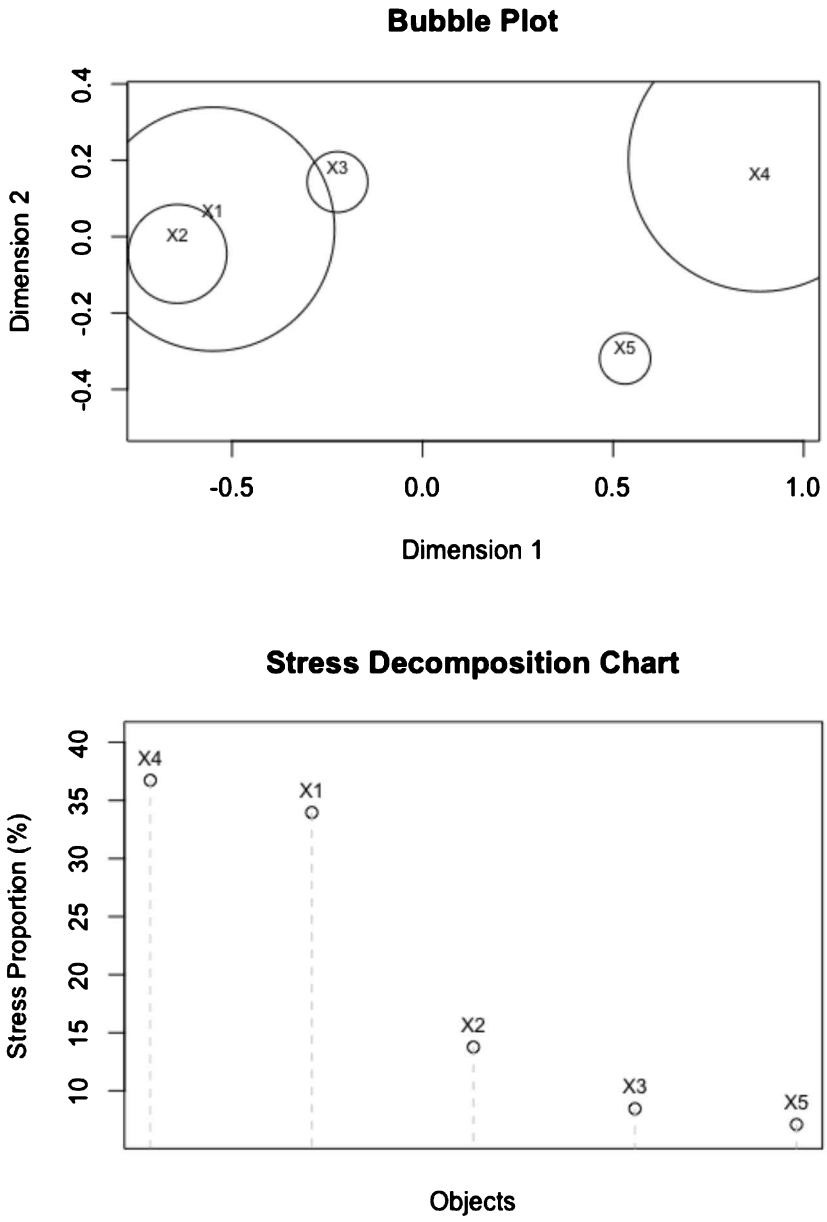
El último indicador de bondad de ajuste que aparece en el cuadro 4.4 y que vamos a analizar es el llamado **RSQ o coeficiente de determinación**. Como ya se indicó, la solución dimensional reducida es una buena representación de la solución N dimensional si la ordenación de las distancias entre los objetos de la primera mantiene la ordenación de las disparidades originales de la segunda. En el caso ideal, el gráfico de dispersión que representara distancias y disparidades debería ser una línea recta.

Ya vimos que el paquete `smacof` proporciona el gráfico de dispersión que aparecía en la figura 4.3 y que ya vimos que se conoce como diagrama de Shepard. De su análisis se desprende que la ordenación lograda con las distancias coincide de manera prácticamente perfecta con las disparidades. Tanto es así que hemos tenido que pedir a `smacof` que nos ofrezca el gráfico solo con los residuos, sin que esté trazada la línea, para ver dónde se producen las discrepancias (figura 4.5).

Ahora cabe preguntarse ¿en qué medida se aleja la relación que se ha representado de la relación ideal? Si efectuáramos una regresión simple tomando las disparidades como variables independientes y las distancias como dependientes, el coeficiente de determinación de esta regresión sería un buen indicador de lo cercano al ideal de la solución obtenida. Pues bien, ese coeficiente de determi-

Estadísticos e-Books & Papers

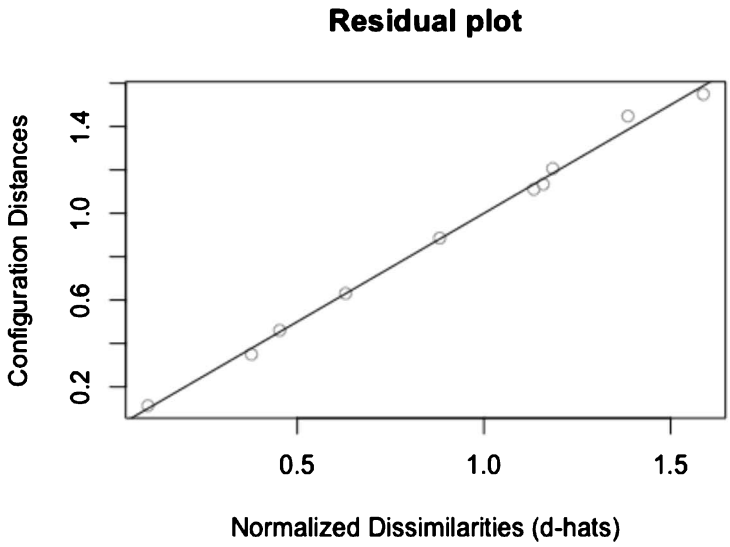
Figura 4.4.: Contribución relativa de cada punto al *stress*



 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 4.5.: Diagrama de residuos



nación es el indicador RSQ del cuadro 4.5 que, como vemos, toma un valor muy cercano a la unidad (0,9920). En el cuadro 4.6 aparecen los datos que permiten al lector realizar por sí mismo una representación como la que aparece en la figura 4.5 y efectuar como ejercicio la regresión simple para confirmar el valor del RSQ. Las dos columnas de valores no son sino los elementos de las matrices $\mathbf{D}$ y Δ.

Con esta información se deriva fácilmente que:

$$Stress = \sqrt{\frac{0{,}008}{10{,}000}} = 0{,}0282$$

$$S - stress = \sqrt{\frac{0{,}055}{16{,}281}} = 0{,}0580$$

y que el resultado de la regresión simple de d sobre δ es:

$$d = 0{,}003378 + 0{,}996598 \times \delta$$

$$R^2 = RSQ = 0{,}9962$$

Estadísticos e-Books & Papers

Cuadro 4.6.: Información para el cálculo de los indicadores de bondad del MDS

Estímulos	d	δ	$(d-\delta)^2$	d^2	δ^2	d^4	$(d^2-\delta^2)^2$
1-2	0.113	0.100	0.000	0.013	0.010	0.000	0.000
1-3	0.350	0.377	0.001	0.122	0.142	0.015	0.000
1-4	1.448	1.386	0.004	2.097	1.920	4.395	0.031
1-5	1.134	1.159	0.001	1.286	1.343	1.654	0.003
2-3	0.460	0.453	0.000	0.212	0.205	0.045	0.000
2-4	1.549	1.587	0.001	2.399	2.519	5.757	0.014
2-5	1.207	1.184	0.001	1.456	1.402	2.119	0.003
3-4	1.111	1.134	0.001	1.234	1.285	1.523	0.003
3-5	0.886	0.882	0.000	0.785	0.777	0.615	0.000
4-5	0.630	0.629	0.000	0.397	0.396	0.158	0.000
Suma	8.887	8.891	0.008	10.000	10.000	16.281	0.055

4.3. Recogida de datos para un escalamiento multidimensional

El input básico del MDS es, como ya hemos señalado, la *similaridad* entre cada par de los N objetos que se están analizando. A esta medida, se la suele denominar también *proximidad*, tal como se ha apuntado. Estas medidas pueden obtenerse de muy diversas formas. Las dos más habituales son (Dillon y Goldstein, 1984): (a) pedir a los individuos que emitan un juicio de similaridad entre cada par de estímulos o (b) que puntúen en qué grado un atributo determinado está presente en el estímulo. Al primer tipo de medidas se las conoce como *similaridades directas* y al segundo como *similaridades derivadas*. A continuación describimos los procedimientos más habituales de recogida de uno y otro tipo de medidas. Para una descripción mucho más detallada, el lector puede remitirse a Coombs (1964) y Shepard (1972).

4.3.1. Similaridades directas

A los individuos se les presentan los pares de estímulos y estos deben juzgar su similaridad. Esto puede hacerse de diversas formas:

1. *Puntuando en una escala.* Ante cada par de estímulos el individuo puntuará su similaridad en una escala de 5 o 7 puntos, donde el 0 querrá decir que ambos estímulos son percibidos como idénticos, y el 5, (o 7) que ambos son totalmente diferentes. Una matriz cuadrada (los estímulos determinan el número de filas y columnas) con las medias de las puntuaciones dadas por el conjunto de los individuos.

2. *Datos de confusión.* Al individuo se le da cada estímulo escrito en una tarjeta y se le pide que forme grupos con ellos de tal forma que en cada

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 4.7.: Descriptores de destinos turísticos

Descriptor	Benidorm	Benicàssim	Toledo	Marbella
Playas limpias y cuidadas				
Diversión nocturna				
Puede hacerse turismo rural				
Buenas infraestructuras deportivas				
Entornos naturales muy cuidados				
Tranquilidad, se puede descansar				

grupo estén los estímulos que se parecen entre sí. Los grupos deben ser exclusivos y exhaustivos y el número puede ser prefijado por el investigador o puede dejarse su elección al arbitrio del entrevistado. A partir de esta información se construye una matriz cuadrada en la cual se pondrá un 1 si los estímulos de la fila y columna de referencia están en un grupo y un 0 en caso contrario (figura 4.6). Estas matrices se agregan para el conjunto de los individuos analizados.

3. *Datos que no son proximidades.* En muchas ocasiones nos podemos encontrar con la necesidad de establecer distancias relativas entre estímulos, a partir de datos que no son proximidades, por ejemplo, entre países a partir de sus datos del PIB, tasa de inflación, tasa de desempleo, etc. La forma más recomendable de derivar una matriz de proximidades de los datos anteriores es obtener una matriz de distancias euclídeas entre los estímulos Wish y Carroll (1974).

4. *Ordenación condicionada.* Este método toma cada estímulo como elemento de comparación y se pide al individuo que ordene el resto de estímulos en función de su similitud con el estándar. El resultado es una matriz que tiene tantas filas y columnas como estímulos. El 0 indica en cada fila cuál es el estímulo de referencia, y el resto de elementos de la fila, la ordenación hecha por el individuo.

4.3.2. Similaridades derivadas

El término "derivadas" tiene su origen en el hecho de que las proximidades se obtienen o "derivan" a partir de las puntuaciones que los individuos dan a los estímulos sobre la base de determinadas afirmaciones que se hacen sobre ellos. Así, por ejemplo, si deseáramos aplicar la técnica MDS para determinar qué destinos turísticos son percibidos como más similares por un conjunto de individuos, podríamos pedirles que evaluaran en qué medida cada uno de los atributos del cuadro 4.7 está presente en cada uno de ellos en una escala, digamos, de 0 a 10, donde el 0 indicaría que el atributo no está presente en el destino de referencia y 10 significaría que lo describe perfectamente.

Estadísticos e-Books & Papers

Figura 4.6.: Obtención de la matriz individual de proximidades

Matriz binaria individual

	1	2	3	4	5	6	7	8	9	10
1	1	1	0	0	0	1	0	0	1	0
2	1	1	0	0	0	1	0	0	1	0
3	0	0	1	1	0	0	0	0	0	0
4	0	0	1	1	0	0	0	0	0	0
5	0	0	0	0	1	0	0	1	0	0
6	1	1	0	0	0	1	0	0	1	0
7	0	0	0	0	0	0	1	0	0	0
8	0	0	0	0	1	0	0	1	0	0
9	1	1	0	0	0	1	0	0	1	0
10	0	0	0	0	0	0	0	0	0	1

Fuente: Trochim (1989).

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Una vez se han obtenido las puntuaciones anteriores, la matriz de proximidades se derivaría calculando la distancia entre los distintos estímulos mediante una medida adecuada. Por ejemplo, si optáramos por la distancia euclídea, el elemento ij de la matriz de proximidades (distancia entre los estímulos i y j, por ejemplo Benidorm y Marbella) se obtendría de la siguiente forma:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (X_{ik} - X_{jk})^2}$$

donde el subíndice k indica el descriptor que se está considerando de los p posibles (6 en el ejemplo del cuadro 4.7).

4.3.3. Consideraciones respecto a la recogida de los datos

El MDS, al igual que ocurre con el análisis de conglomerados, no es una técnica exigente respecto a las condiciones que deben cumplir los datos. Dado que no tiene pretensiones inferenciales, sino meramente descriptivas de la muestra empleada, no le es exigible ninguna de las propiedades tradicionales de normalidad, homocedasticidad, etc.

De todas formas, el hecho de que no le sean exigibles propiedades estadísticas a los datos no quiere decir, ni muchísimo menos, que no haya que tener en cuenta una serie de condiciones en su recogida de cuyo cumplimiento dependerá la calidad del análisis efectuado. Algunas se convierten en limitaciones de la técnica. Veamos las más importantes:

1. La facilidad o dificultad de la recolección de los datos viene determinada por el número de estímulos. Cuanto mayor sea este número, más comparaciones deberán efectuar los individuos y más probabilidad hay de que el cansancio sesgue los resultados. Así, el número de pares para comparar depende del número n de estímulos, del siguiente modo: $C_2^n = n(n-1)/2$. Así, 10 estímulos provocan 45 pares, y 20 estímulos, 190 pares. Puede pensarse que es conveniente que existan pocos estímulos, pero esto puede provocar que afloren soluciones poco estables (Dillon y Goldstein, 1984). Además, el número de dimensiones que se pueden considerar aumenta con el número de estímulos y, por ello, si hay pocos estímulos, relaciones que solo aparecerían con un número elevado de dimensiones quedarían ocultas. A modo de guía, Schiffman *et al.* (1981) recomiendan 12 estímulos para una solución bidimensional y 18 para una tridimensional. Kruskal y Wish (1978) recomiendan 9 estímulos para una solución de dos dimensiones, 13 para una de tres y 17 para una de cuatro.

2. Hay que tener en cuenta que no todos los individuos tendrán en su mente las mismas dimensiones a la hora de juzgar la proximidad entre dos estímulos. Así, si se les pide que comparen modelos de coches, unos pueden tener en su mente la potencia y el diseño para juzgar la similitud, mien-

Estadísticos e-Books & Papers

tras que otros pueden estar considerando el coste y la comodidad (Hair *et al.*, 2014a).

3. Aunque todos los entrevistados estuvieran considerando las mismas dimensiones, no todos tendrían por qué estar atribuyendo el mismo peso a cada una de ellas.

4. Los juicios emitidos respecto a un par de estímulos, sus dimensiones o sus pesos no tienen por qué ser estables en el tiempo (Hair *et al.*, 2014a).

4.4. Tipos de escalamiento multidimensional

En esta sección describiremos e ilustraremos con un ejemplo los principales tipos de escalamiento multidimensional. Como ya se ha indicado, el MDS no es una única técnica, sino un conjunto de ellas. Los elementos que permiten diferenciarlas son: el número de matrices de proximidades, la forma de las mismas, cuadradas o rectangulares, y si el algoritmo contempla o no ponderaciones. La combinación de estos elementos da lugar a la tipología que detallaremos a continuación.

4.4.1. Escalamiento multidimensional clásico

Aparece en la literatura bajo las siglas CMDS (*Classic Multidimensional Scaling*) y es el tipo que hemos utilizado para ilustrar la técnica en la sección 4.2., es decir, está formado por una única matriz de proximidades y esta es cuadrada: el número de filas y columnas es el mismo e igual al número de estímulos que se comparan. Existen dos tipos de CMDS en función de cómo sean las medidas de similaridad. Así el **CMDS métrico**, debido al trabajo seminal de Torgeson (1952), asume que las medidas de similaridad son de intervalo o de razón. Este sería el caso, por ejemplo, de una matriz donde los estímulos son ciudades, y la medida de proximidad, la distancia en kilómetros entre ellas.

Este enfoque, muy limitado, fue mejorado por los trabajos de Shepard (1962) y Kruskal (1964b) y dio lugar al **CMDS no métrico**, donde el nivel de medida de las variables es ordinal. En la función `mds{smacof}` que es la que estima el CMDS, la opción por uno u otro instrumento de medida se elige con el modificador `type = c("ratio", "interval", "ordinal", "mspline")`.

Caso 4.2. Desarrollo educativo de distintas zonas del mundo

Para ilustrar la aplicación del CMDS, dado que en el epígrafe 4.2 se pusieron ejemplos con los procedimientos más habituales de recogida de la información de los que hemos destacado en el epígrafe 4.3, examinaremos ahora un ejemplo que se basa en datos que no son proximidades y que deben ser transformados previamente para poder someterse a un CMDS.

Supongamos que un investigador en economía de la educación desea saber cuál es la posición relativa de distintas zonas del mundo respecto al nivel de

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

desarrollo educativo. El cuadro 4.8 recoge los principales indicadores que este investigador maneja para esta finalidad. Es obvio que podría recurrirse a un análisis de conglomerados para conseguir un número de grupos territoriales homogéneos, pero estos grupos estarán próximos o alejados entre sí. Estas distancias, para el investigador, son relevantes porque le muestran las diferencias regionales y el camino que queda por recorrer.

Los indicadores se corresponden con la siguiente leyenda:

- I1. Tirada de diarios. Número de ejemplares por mil habitantes.
- I2. Tasa bruta de escolarización en la enseñanza preprimaria.
- I3. Tasa bruta de escolarización en la enseñanza secundaria.
- I4. Tasa bruta de escolarización en la enseñanza superior.
- I5. Gasto público en educación (% del PIB).
- I6. Número de docentes de todos los niveles educativos (miles) por cada mil habitantes de 15 a 64 años.
- I7. Consumo de papel de imprenta (kg por habitante).

Las zonas geográficas que requieren alguna aclaración respecto a los países que las componen serán las siguientes:

- Asia / Oceanía. Exclusivamente Australia, Israel, Japón y Nueva Zelanda.
- Europa: no comprende la Europa del Este, que se incluye en países en transición.
- Países en transición: Albania, Armenia, Azerbaiyán, Eslovaquia, Federación Rusa, República Checa, Bulgaria, Hungría, Yugoslavia, entre otros.
- África subsahariana: Angola, Benín, Botsuana, Burkina Faso, Gambia, Guinea, Kenia, Sudáfrica, Sudán, entre otros.
- Estados árabes: Arabia Saudí, Argelia, Bahréin, Yibuti, Egipto, Emiratos Árabes, Irak, Jordania, Kuwait, Líbano, Marruecos, Siria, entre otros.
- Asia Oriental y Oceanía: Brunéi, Camboya, Fiyi, Filipinas, Hong Kong, Indonesia, Macao, Malasia, Corea del Norte, Corea del Sur, Singapur, Tailandia, entre otros. Por su importancia relativa, China se considera aparte.
- Asia Meridional: Afganistán, Bangladés, Bután, Irán, Maldivas, Nepal, Pakistán y Sri Lanka. Por su importancia relativa, India se considera aparte.

Estadísticos e-Books & Papers

Cuadro 4.8.: Indicadores de desarrollo educativo

Etiqueta	Descripción	I1	I2	I3	I4	I5	I6	I7
Z1	América del Norte	213	68,1	97,0	84,0	5,5	23	148,0
Z2	Asia/Oceanía	520	53,3	107,8	45,3	4,0	21	108,2
Z3	Europa	250	77,3	111,4	47,8	5,4	24	82,2
Z4	Países en transición	114	54,0	86,9	34,2	5,2	27	6,1
Z5	África subsahariana	12	9,2	24,3	3,5	5,6	10	1,5
Z6	Estados árabes	37	15,4	53,7	12,5	5,2	20	2,7
Z7	América Latina y Caribe	80	51,1	56,6	17,3	4,5	22	10,7
Z8	Asia Oriental y Oceanía	57	28,9	61,5	8,9	3,0	14	7,3
Z9	China	43	28,7	66,6	5,3	2,3	13	5,5
Z10	Asia Meridional	27	9,8	44,5	6,5	4,3	9	1,9
Z11	India	32	4,8	48,7	6,4	3,5	9	1,9
Z12	Países menos adelantados	7	10,8	18,4	3,2	2,5	8	0,4

© Ediciones Paraninfo

Estadísticos e-Books & Papers

- Países menos adelantados. Se consideran, de todos los anteriores, aquellos países con un atraso mayor. Este punto servirá en el MDS de referencia de nivel. Incluye Afganistán, Angola, Bangladés, Benín, Bután, Etiopía, Gambia, Guinea, Haití, Nepal, Níger, Congo, Tanzania, Yemen, Zambia, entre otros.

Como puede comprobarse, los datos del cuadro 4.8 no son una medida de proximidad entre los estímulos (las 12 zonas). Para obtener esta matriz calcularíamos las distancias euclídeas entre los estímulos teniendo en cuenta las variables que los caracterizan (I1 a I7), de acuerdo con la sintaxis del cuadro 4.8. Lógicamente las variables son previamente estandarizadas, como se observa en la sintaxis.

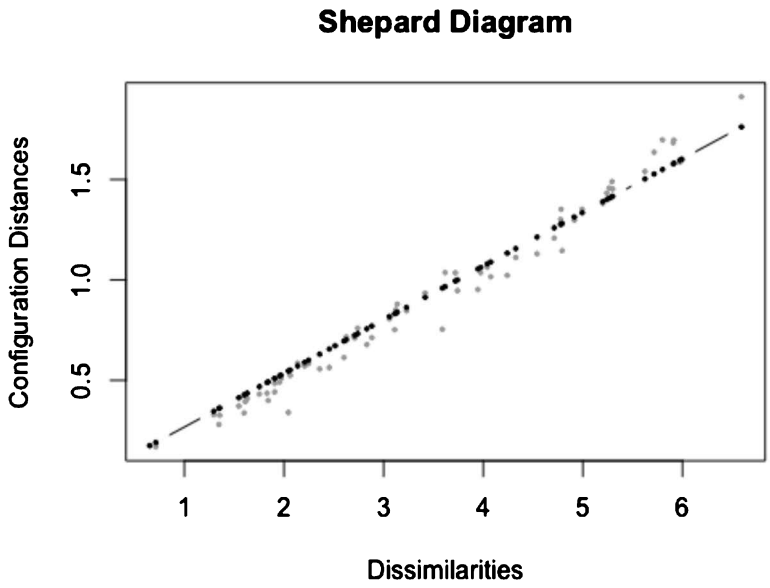
```
#Quitamos zona de la base de datos
Datos_4_2_Caso$zona <- NULL
#Normalizamos los indicadores
datos.normalizados <- scale(Datos_4_2_Caso)
#Calculamos la matriz de distancias
datos<-dist(datos.normalizados, method = "euclidean",
diag = TRUE, upper = TRUE)

m <- as.matrix(datos)
rownames(m) <- paste("Z", 1:12)
colnames(m) <- paste("Z", 1:12)
datos<-as.dist(m)
```

Como se tienen 12 estímulos, estaríamos en la cifra recomendada por Schiffman *et al.* (1981) para una solución de dos dimensiones y por encima de la recomendada por Kruskal y Wish (1978). Efectuamos a continuación la aplicación del MDS para una solución bidimensional.

```
library(smacof)
fit<-mds(delta=datos,ndim=2,type="ratio")
# Coordenadas
print(fit$conf)
# Disparidades
print(fit$dhat)
# Distancias entre configuraciones
print(fit$confdiss)
# Stress
print(fit$stress)
# Stress por punto
print(fit$spp)
#RSQ
```


Estadísticos e-Books & Papers

Figura 4.7.: Gráfico de Shepard




```
dist<-cbind(c(fit$dhat))
dism<-cbind(c(fit$confdiss))
summary(lm(dist~dism))

# gr<U+00E1>fico: "confplot", "resplot", "Shepard",
#"stressplot", "bubbleplot"
plot(fit, plot.type="confplot", plot.dim = c(1,2),
sphere = TRUE, bubscale = 0.1, col = 1,
label.conf = list(label = TRUE, pos = 3,
col = 1, cex = 0.8),
shepard.x= NULL, identify = FALSE, type = "p", pch = 20,
asp = 1, col.hist = NULL)
```

El cuadro 4.9 nos ofrece los indicadores de bondad de dicha solución. Puede verse que el algoritmo convergió tras 64 iteraciones ofreciendo un *stress* de 0,071, cifra que puede ser considerada entre "bueno" y "mínimo razonable" de acuerdo con Kruskal (1964a) (véase cuadro 4.4). Asimismo el coeficiente de determinación (RSQ) está muy cercano a 1 (0,9818), demostrando así que el ajuste entre disparidades y distancias es casi perfecto. Este hecho se constata gráficamente en la figura 4.7, que recoge el gráfico de Shepard.

 © Ediciones Paraninfo

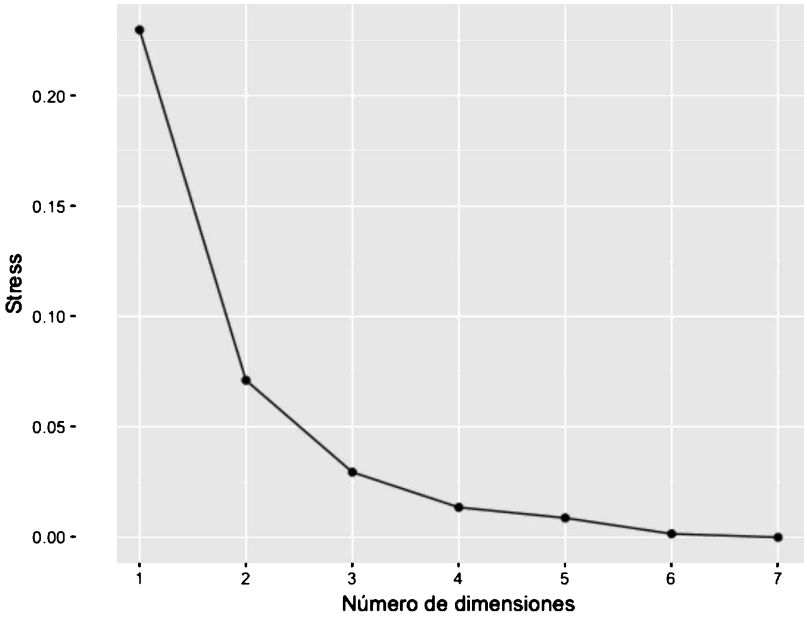
Estadísticos e-Books & Papers

Cuadro 4.9.: Valor del *stress* y otros indicadores de ajuste para los datos del caso

```
Call:
mds(delta = datos, ndim = 2, type = "ratio")

Model: Symmetric SMACOF
Number of objects: 12
Stress-1 value: 0.071
Number of iterations: 64

Call:
lm(formula = dist ~ dism)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.09872    0.01546   6.385  2.2e-08 ***
dism         0.90927    0.01546  58.811  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 0.05644 on 64 degrees of freedom
Multiple R-squared:  0.9818, Adjusted R-squared:  0.9815
F-statistic:  3459 on 1 and 64 DF,  p-value: < 2.2e-16
```


Estadísticos e-Books & Papers

Figura 4.8.: Valor del *stress* para distinto número de dimensiones



Hemos impuesto una solución de dos dimensiones porque la mayor aportación del MDS reside en la posibilidad de visualizar las posiciones relativas de las regiones en un mapa de estas características. Sin embargo, es necesario corroborar que esta solución bidimensional es razonable. De no ser así, si fuera necesario recurrir a cuatro o cinco dimensiones, el MDS no ofrecería demasiada ayuda respecto al análisis de conglomerados, puesto que el análisis gráfico de la solución sería complicado.

Los indicadores de ajuste de la solución bidimensional parecen asegurar la adecuación de la misma. Dillon y Goldstein (1984) y Johnson y Wichern (1998) recomiendan, sin embargo, una prueba adicional. Esta consiste en ensayar distintas soluciones dimensionales (digamos de 1 a 7) y representar en un gráfico las dimensiones en abscisas y el *stress* en ordenadas. Si la solución bidimensional es válida, el *stress* debe caer rápidamente hasta alcanzar esa dimensión, disminuyendo mucho menos en dimensiones adicionales. La figura 4.8, obtenida de este modo, confirma que la solución bidimensional es la adecuada.

Para facilitar la interpretación de los resultados, `mds{smacof}` deriva las coordenadas de los estímulos en el nuevo espacio bidimensional (cuadro 4.10) y los representa gráficamente (figura 4.9). El análisis de la representación gráfica es la principal herramienta para deducir conclusiones.

La figura 4.9 nos muestra que existen, al menos, tres grupos de regiones con un desarrollo educativo muy diferenciado. Un grupo lo formarían América del

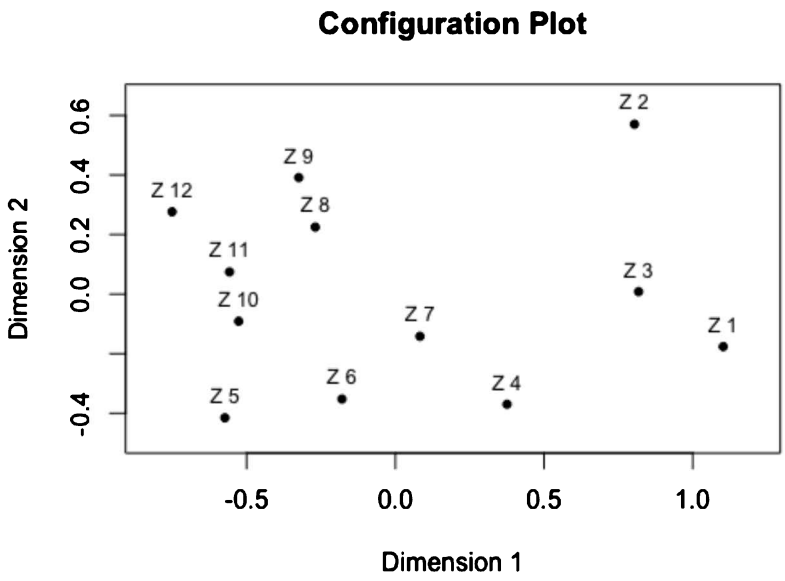
 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 4.10.: Coordenadas de los estímulos

```
Configurations:
          D1      D2
Z 1   1.1024 -0.1762
Z 2   0.8037  0.5703
Z 3   0.8179  0.0085
Z 4   0.3755 -0.3695
Z 5  -0.5730 -0.4152
Z 6  -0.1796 -0.3523
Z 7   0.0829 -0.1416
Z 8  -0.2694  0.2249
Z 9  -0.3249  0.3914
Z 10 -0.5270 -0.0910
Z 11 -0.5577  0.0745
Z 12 -0.7506  0.2764
```

Figura 4.9.: Representación bidimensional de las regiones producida por el CMDS



Estadísticos e-Books & Papers

Norte (Z1), Asia / Oceanía (Z2) y Europa (Z3). Este grupo vendría caracterizado por contener a los países con mayor desarrollo de sus sistemas educativos. Téngase en cuenta que Z2 incluye solo a aquellos países de Asia y Oceanía más avanzados, como son Japón, Nueva Zelanda o Australia. Interpretamos que estos países son los más avanzados no solo por el sentido común, difícil de aplicar en otro tipo de análisis, sino porque son los más alejados de Z12, que, recordemos, incluía como referencia a aquellos países menos adelantados.

Un grupo intermedio vendría formado por los países euroasiáticos en transición (Z4), América Latina y el Caribe (Z7) y los Estados árabes (Z6), que tendrían un desarrollo educativo intermedio entre los más avanzados y los de menor desarrollo. Estos últimos serían los recogidos en el mapa más cerca de Z12, es decir, África subsahariana, Asia Oriental y Oceanía, Asia Meridional e India.

El lector puede preguntarse el motivo de algunas decisiones que se han tomado en el análisis, como, por ejemplo, asignar Z5 al grupo de países menos desarrollados y, en cambio, asignar Z6 al grupo intermedio. O por qué considerar tres agregados y no más (o menos) en la interpretación. Para ayudar en este tipo de análisis de los gráficos de MDS, Kruskal y Wish (1978) recomiendan complementar esta técnica con otras ya expuestas, como el análisis de conglomerados. Así, si sometiéramos las coordenadas de los estímulos en el espacio bidimensional a un análisis de conglomerados jerárquico, el dendograma nos ayudaría a tomar la decisión acerca de qué estímulos son más cercanos o pueden formar grupos de interpretación. Atendiendo a ese dendograma (figura 4.9), se han formado los grupos de la figura 4.10 que, como se puede comprobar, sirvió de base para la interpretación que expusimos con anterioridad.

4.4.2. Escalamiento multidimensional ponderado

Parece lógico plantearse la cuestión de si es posible analizar varias matrices de proximidades simultáneamente cuando tenemos la sospecha de que pueden proceder de individuos o colectivos con esquemas perceptuales distintos. Por ejemplo, podemos efectuar una encuesta para determinar el posicionamiento en la mente del consumidor de un conjunto de superficies comerciales. Bastaría para ello que los entrevistados hicieran grupos de superficies que para ellos fueran similares y derivar de ahí una matriz de proximidad. Pero el investigador puede sospechar que el posicionamiento de estas superficies puede ser distinto entre hombres y mujeres. Si obtiene dos matrices de proximidad para estos colectivos, ¿puede analizarlas simultáneamente aun sabiendo que las estructuras perceptuales son distintas?

Carroll y Chang (1970) generalizaron el MDS de tal forma que pudieran analizarse diversas matrices de proximidades cuando se presupone que existen diferencias sistemáticas entre ellas, originadas por diferencias individuales en los procesos perceptuales o cognitivos que los han generado. Esta técnica se conoce como **escalamiento multidimensional ponderado** o WMDS (*Weighted Multidimensional Scaling*). Por introducir diferencias individuales también se

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 4.10.: Dendograma del análisis de conglomerados sobre las coordenadas de los estímulos

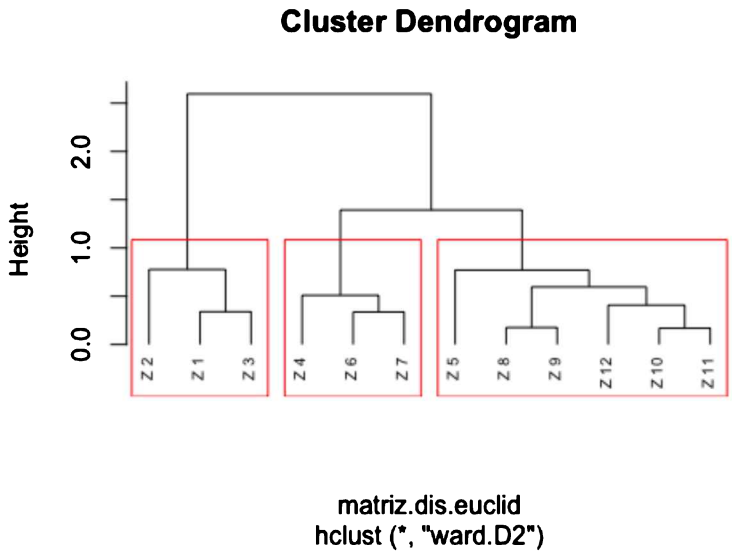
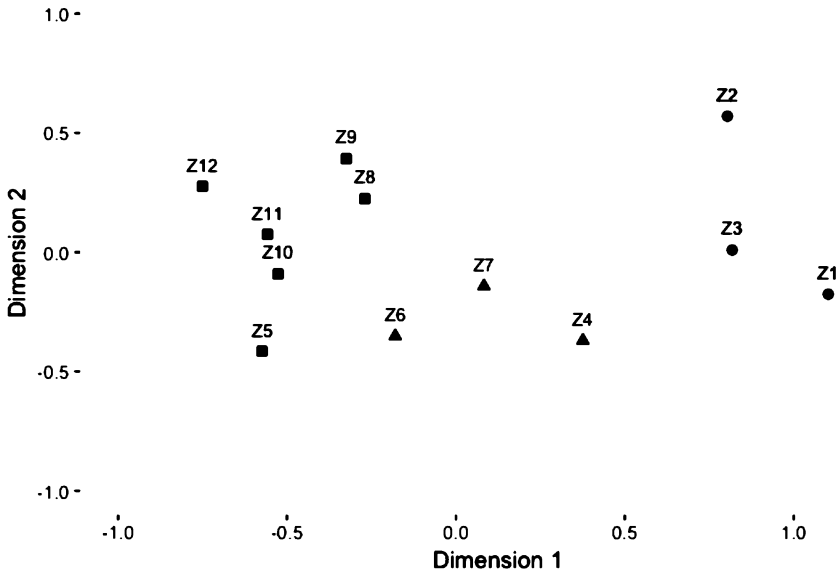


Figura 4.11.: Grupos formados a partir del dendograma



Estadísticos e-Books & Papers

le denomina, frecuentemente, **escalamiento de diferencias individuales** o INDSCAL (*Individual differences scaling*).

Esta técnica, como ilustraremos inmediatamente con un ejemplo, deriva un mapa de estímulos común a los individuos y equiparable al CMDS. Sin embargo, con esta técnica se obtiene un espacio adicional asociado a cada individuo donde aparece la ponderación que cada uno de ellos atribuye a las dimensiones obtenidas.

Caso 4.3 Apertura de una cadena de electrodomésticos

Supongamos que una gran cadena de hipermercados está planteándose la apertura de nuevos centros. Para decidir en qué comunidades autónomas puede resultar más oportuna la expansión plantea a dos de sus departamentos que agrupen las comunidades autónomas que consideren que tienen un atractivo similar. Estos departamentos son el de marketing y el encargado de la adquisición de los terrenos. Una vez obtenida la matriz de proximidades, al encargado de analizarla le asalta la duda de si estos departamentos pueden tener esquemas perceptuales distintos para juzgar la similitud de las comunidades autónomas. Quizás el departamento de marketing esté utilizando criterios de poder adquisitivo para establecer la similitud, mientras que el departamento que ha de comprar los terrenos considere aquello que más difícil hace su trabajo: la competencia de este tipo de superficie existente en cada una de ellas.

Atendiendo a estos razonamientos, decide aplicar el WMDS en lugar del CMDS, tratando las dos matrices de proximidades por separado. La figura 4.12 ofrece los resultados de este análisis, correspondiéndose las etiquetas de la misma con las equivalencias mostradas en el cuadro 4.11. El modelo se ha estimado con la función `smacofIndDiff{smacof}` como sigue:

```
datos_D1<-data.matrix(Datos_4_3a_Caso)
datos_D2<-data.matrix(Datos_4_3b_Caso)
rownames(datos_D1)<-colnames(datos_D1)
rownames(datos_D2)<-colnames(datos_D2)
datos<-list(datos_D1,datos_D2)
fit <- smacofIndDiff(delta=datos,type="ordinal",
constraint = "indscal")
```

El mapa de la figura 4.12 muestra la percepción que es común a los dos departamentos. De ella se deriva que las comunidades se diferencian poco unas de otras, salvo en 5 casos. Por un lado, la Comunidad de Madrid (C13) es percibida como muy distinta de todas las demás y algo parecido ocurre con Cataluña (C9) y Andalucía (C1). Por otra parte, Canarias (C5) y la Comunidad Valenciana (C10) tienen alguna particularidad que las diferencia, aunque de forma más suave, del grupo formado por las restantes comunidades.

Los indicadores de bondad de ajuste que se aplican en el WMDS son los mismos que los aplicados en el CMDS. En este caso, el *stress* obtenido para la solución dada es 0,0358.

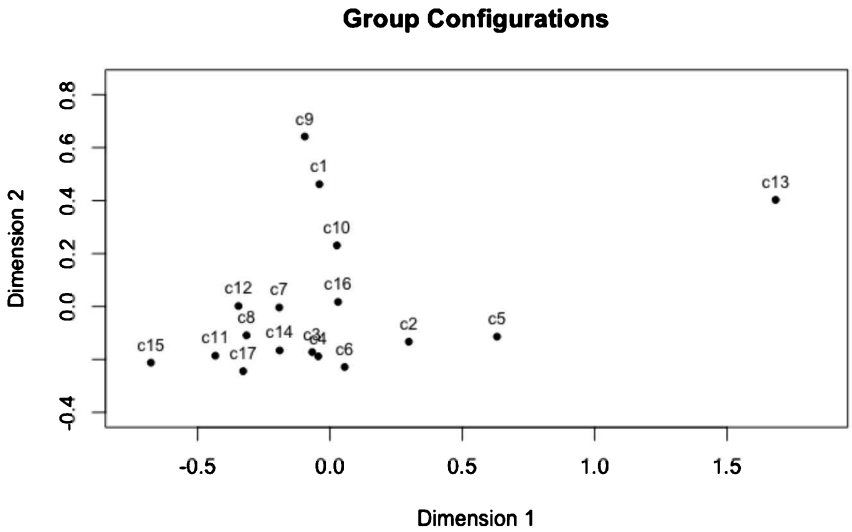
 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 4.11.: Etiquetas asociadas a las comunidades autónomas

Etiqueta	Comunidad autónoma
C1	Andalucía
C2	Aragón
C3	Asturias
C4	Baleares
C5	Canarias
C6	Cantabria
C7	Castilla y León
C8	Castilla-La Mancha
C9	Cataluña
C10	Comunidad Valenciana
C11	Extremadura
C12	Galicia
C13	Madrid
C14	Murcia
C15	Navarra
C16	País Vasco
C17	Rioja

Figura 4.12.: Mapa conjunto de estímulos derivados del WMDS



Estadísticos e-Books & Papers

Cuadro 4.12.: Pesos de la configuración para cada departamento

```
> fit$cweights
[[1]]
          D1       D2
D1 0.1092108 0.000000
D2 0.0000000 2.608886

[[2]]
         D1        D2
D1 1.346635 0.0000000
D2 0.000000 0.1602203
```

A diferencia del ejemplo utilizado para ilustrar el CMDS, este posicionamiento relativo es fruto de la valoración subjetiva de dos departamentos, es decir, no está basado en variables objetivas. Para realizar el análisis del MDS se requiere información adicional que permita explicar ese mapa. Lo primero que se puede hacer, dada la técnica empleada, es analizar la importancia relativa que el departamento de marketing (D1) y el de compra de terrenos (D2) atribuye a cada una de las dos dimensiones del mapa. Para ello debe analizarse el espacio ponderado (figura 4.13) que se obtiene de los pesos de la configuración ofrecidos por `smacofIndDiff{smacof}` (cuadro 4.12).

Este mapa debe interpretarse del siguiente modo: las diferencias individuales (departamentos) en la percepción vienen representadas por vectores. Cada vector tiene una longitud y forma un ángulo respecto a cada una de las dimensiones. El factor más importante es este último. Así, cuando el vector que representa a un individuo se separe del ángulo de 45° para acercarse a una dimensión dada, querrá decir que esa dimensión ha sido más importante para configurar el mapa perceptual de ese individuo. Si un vector tiene una mayor longitud que otro, este hecho solo indica que la matriz que generó ese individuo está mejor representada en la solución final que la de otro individuo con un vector más corto.

Entonces habría que recurrir a información secundaria que pudiera confirmarnos esta interpretación. Supongamos ahora que el investigador recoge la información que aparece en el cuadro 4.13. La *cuota de mercado* (Y_1) es un índice que expresa la capacidad de consumo comparativa de las comunidades autónomas y se obtiene como un promedio de índices de variables como el número de teléfonos, población, automóviles, oficinas bancarias y actividades comerciales minoristas. Por otra parte, la variable objetiva que el investigador considera que puede medir la *presión competitiva* son los metros cuadrados de grandes superficies comerciales por habitante (Y_2). El resto de columnas son las coordenadas X_1 y X_2 de cada comunidad autónoma en la configuración de estímulos derivada de la figura 4.12, cuya utilidad para la interpretación de este caso se expondrá a continuación.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

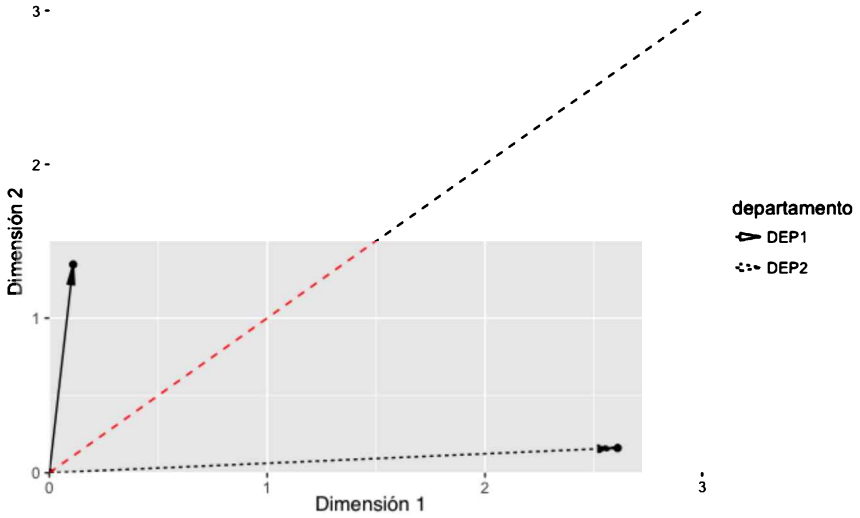
Cuadro 4.13.: *Indicadores objetivos del mapa comercial español*

Etiqueta	Comunidad autónoma	Cuota de mercado	m^2 de grandes superficies por habitante	Coord. X1	Coord. X2
C1	Andalucía	16.795	0,13	-0,040	0,462
C2	Aragón	3.132	0,18	0,298	-0,133
C3	Asturias	2.545	0,12	-0,067	-0,172
C4	Baleares	2.301	0,12	-0,044	-0,188
C5	Canarias	4.212	0,17	0,631	-0,114
C6	Cantabria	1.289	0,13	0,056	-0,228
C7	Castilla y León	6.286	0,10	-0,192	-0,004
C8	Castilla-La Mancha	4.376	0,08	-0,315	-0,109
C9	Cataluña	16.575	0,12	-0,095	0,642
C10	Comunidad Valenciana	10.436	0,14	0,026	0,231
C11	Extremadura	2.607	0,07	-0,433	-0,186
C12	Galicia	6.503	0,08	-0,345	0,002
C13	Madrid	12.707	0,26	1,683	0,403
C14	Murcia	2.725	0,10	-0,190	-0,166
C15	Navarra	1.434	0,06	-0,676	-0,212
C16	País Vasco	5.050	0,14	0,031	0,017
C17	Rioja	723	0,08	-0,328	-0,244

© Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 4.13.: Mapa de pesos de los sujetos del WMDS



Si la interpretación que hemos avanzado de las dimensiones es la correcta, cuando regresemos cada una de las variables objetivas obtenidas de fuentes secundarias que las representen sobre las coordenadas X_1 y X_2 de los estímulos, los cosenos directores deberían estar próximos a 1 en la coordenada de la dimensión que está mejor representada por la variable objetiva que hemos utilizado como dependiente. En general si regresamos una variable objetiva Y sobre las coordenadas X_1 a X_r de un espacio r dimensional:

$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_rX_r + u$$

entonces, el coseno director estimado respecto a cada una de las dimensiones vendrá dado genéricamente por:

$$\cos\theta_i = \frac{\hat{b_i}}{\sqrt{\hat{b}_1^2 + \hat{b}_2^2 + \cdots + \hat{b}_r^2}}$$

En nuestro caso, los planos que se estiman mediante la regresión lineal son los siguientes:

$$Y_1 = 5864,5 - 307,7X_1 + 19111,5X_2 \longrightarrow R^2 = 0,9725$$
$$Y_2 = 0,122353 + 0,088063X_1 + 0,010968X_2 \longrightarrow R^2 = 0,9262$$

Siendo Y_1 la cuota de mercado e Y_2 los metros cuadrados de grandes superficies por habitante. Los cosenos directores de Y_1 sobre las dimensiones 1 y 2 son los siguientes:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

$$\cos\theta_1 = \frac{-307,7}{\sqrt{307,7^2 + 19111,5^2}} = -0,01609$$

$$\cos\theta_2 = \frac{19111,5}{\sqrt{307,7^2 + 19111,5^2}} = 0,99987$$

De acuerdo con estos resultados, la cuota de mercado está muy bien representada por la dimensión 2, que, recordemos, era a la que más importancia daba el departamento de marketing (D1), lo que confirma como razonable la interpretación anteriormente efectuada. Si se repitiera el ejercicio para Y_2, se comprobaría como los metros cuadrados de superficie comercial están bien representados por la dimensión 1.

4.4.3. Escalamiento multidimensional clásico desdoblado

La última variante que trataremos de MDS es aquella cuyo input no es una matriz o matrices cuadradas, con los estímulos formando filas y columnas, sino rectangular, donde los estímulos forman las columnas y las variables que caracterizan a dichos estímulos forman las filas (o viceversa). A este tipo de MDS se le conoce en la literatura por las siglas CMDSU (*Classical Multidimensional Scaling Unfolding*). Su interés se comprenderá mejor con un ejemplo.

Imaginemos que un investigador en el campo del marketing desea efectuar un análisis de posicionamiento de franquicias de comida entre el público joven. Para ello elabora un cuestionario como el mostrado en el cuadro 4.14 en el que solicita que se señale con una cruz cuando se considera que la franquicia que aparece en la columna posee el atributo que se señala en la fila. La muestra es de 54 consumidores jóvenes y la matriz de frecuencias indicaría con un número más alto cuando más jóvenes asocian una cadena con un atributo. Los datos para el CMDSU, al menos como lo implementa el paquete que estamos utilizando `smacofRect{smacof}`, son datos de preferencias, es decir, cuanto más bajo es el número, mayor es la asociación del atributo con la marca, por lo que hemos de recodificar la matriz de frecuencias a preferencias simplemente asignando un rango a la frecuencia. Al restaurante que tenga la frecuencia más alta para un atributo se le asigna el rango 1, al segundo más alto 2 y así sucesivamente. El mismo cuadro 4.14 nos ofrece la base de datos de partida.

Sometemos a la matriz rectangular del cuadro 4.14 a un CMDSU con la siguiente sintaxis de `smacofRect{smacof}`:

```
library(smacof) datos<-Datos_4_4_Caso
fit <- smacofRect(datos, itmax = 1000)
plot(fit, joint = TRUE, xlim = c(-5, 5), asp=.6)
fit$conf.row
fit$conf.col
```

Donde hemos solicitado al paquete que nos ofrezca en la salida las configuraciones de las filas (atributos) y de la comida (franquicias) porque el gráfico

Estadísticos e-Books & Papers

Cuadro 4.14.: Datos para el estudio de posicionamiento

Atributo	Descripción	Cadena						
		Tagliatella	McDonalds	Burger King	Domino's	Vips	100 Monta-ditos	Fosters
1	Precio bajo	5	2	3	4	5	1	5
2	Instalaciones limpias	1	5	5	4	2	5	3
3	Buen trato personal	1	5	7	6	3	4	2
4	Rapidez servicio	7	1	3	4	6	2	5
5	Servicio a domicilio	3	3	2	1	6	6	5
6	Variedad de comida	1	5	6	6	2	4	2
7	Comida saludable	1	5	5	5	2	4	3
8	Comida de calidad	1	6	6	4	3	5	2
9	Opción vegetariana	2	3	6	7	1	5	4
10	Muchos establecimientos	6	3	1	3	7	2	5
11	Foto carta=producto final	1	6	7	4	3	4	2
12	Zona juegos/Infantil	5	1	2	5	3	5	4
13	Promociones	6	3	1	3	7	2	5
14	Familiar	4	2	3	6	5	7	1
15	Juvenil	6	1	2	2	7	4	5
16	Romántico	1	4	4	6	3	6	2
17	Compromiso social	3	1	3	7	5	6	2

© Ediciones Paraninfo
Estadísticos e-Books & Papers

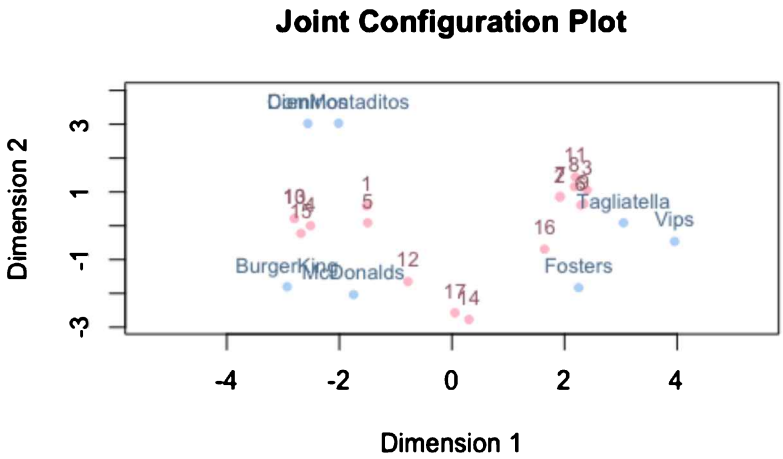
que saca por defecto con la opción *plot* el programa es de una calidad gráfica limitada, como se comprueba en el panel a) de la figura 4.14 y hemos querido mejorarlo en el panel b).

La interpretación del mapa es bastante inmediata, existen básicamente dos grupos de franquicias asociadas a distintos tipos de atributos. Por un lado, Burger King y McDonalds, aunque separados del mapa de Domino's y 100 montaditos, comparten con ellos los atributos que, fundamentalmente, son (1) precio bajo, (5) servicio a domicilio, (4) rapidez del servicio, (10) muchos establecimientos, (13) promociones y (15) juvenil. Probablemente la diferenciación entre ellos viene de una mayor asociación de Burger King y McDonalds a (12) zona de juegos, (14) familiar y (17) compromiso social.

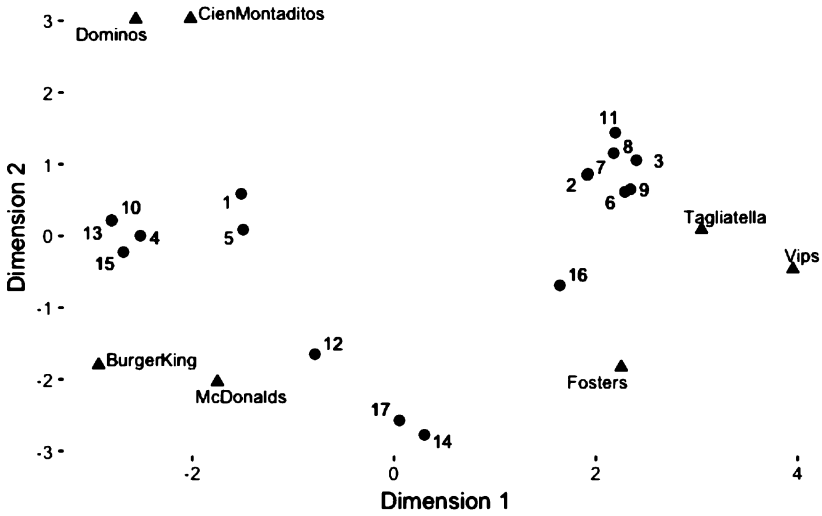
El segundo gran grupo estaría formado por Fosters, Tagliatella y Vips, que comparten atributos tales como (2) instalaciones limpias, (3) buen trato del personal, (6) variedad de comida, (7) comida saludable, (8) comida de calidad, (9) opción vegetariana, (11) la foto de la carta coincide con el resultado final y (16) carácter romántico.

Estadísticos e-Books & Papers

Figura 4.14.: Mapas perceptuales original (`smacof`) y elaborado



(a) Mapa perceptual original



(b) Mapa perceptual elaborado

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

5. Análisis de correspondencias

5.1. Introducción

El objetivo del análisis de correspondencias (CA) es muy parecido al del escalamiento multidimensional, lo que nos ha llevado a describirlos en capítulos consecutivos. Esta técnica estadística pretende representar en un espacio multidimensional reducido la relación existente entre las categorías de dos variables no métricas.

Es, por tanto, el tipo de datos que analiza, y no la finalidad del análisis lo que lo diferencia del MDS. En el escalamiento multidimensional el mapa perceptual que se obtenía mostraba las distancias percibidas entre diversos sujetos (objetos), atendiendo a un conjunto de variables que los caracterizaban. En el análisis de correspondencias, el mapa mostrará las distancias entre los distintos niveles de dos variables no métricas, por lo que se suele decir que el análisis de correspondencias sirve para visualizar tablas de contingencia. La figura 5.1 pretende ilustrar la finalidad de esta técnica. Como se verá en el desarrollo del tema, el análisis de correspondencias es generalizable a tablas resultantes del cruce de más de dos variables.

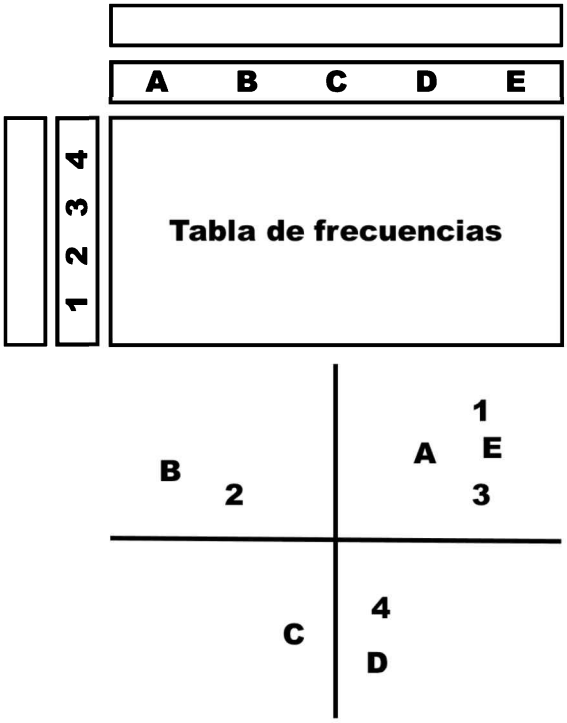
El análisis de correspondencias es una técnica que, pese a su innegable utilidad, no ha sido tan utilizada como otras en investigación en ciencias sociales. La razón cabe buscarla, sin ninguna duda, en el hecho de que su máximo desarrollo se produjo en un entorno no anglófono, gracias a los trabajos del francés Benzécri (1973). Este hecho ha provocado, también, una gran confusión en cuanto a la denominación misma de la técnica. Nosotros hemos optado por mantener la traducción más directa posible del *analyse des correspondances* francés o del *correspondence analysis* inglés, aunque en España algunos autores optan por el término análisis factorial de correspondencias que, en nuestra opinión, puede ocasionar confusión con otras técnicas descritas en este libro. En Holanda se refieren a ella como *Homogeneity Analysis*, *Dual Scaling* en Canadá, y en Estados Unidos se emplean diversos términos como: *Optimal scaling, reciprocal averaging, optimal scoring* y *appropriate scoring.*

Además del ya citado trabajo de Benzécri (1973), el lector que desee una descripción más profunda de la que aquí ofreceremos del análisis de correspondencias puede referirse a Lebart *et al.* (1984), Greenacre (1984) y Greenacre (1994).

A un nivel introductorio, son recomendables los trabajos de Weller y Romney (1990) y, sobre todo, el de Clausen (1998). Finalmente, otra buena introducción a la técnica, pero con aplicaciones al campo del marketing, puede encontrarse

Estadísticos e-Books & Papers

Figura 5.1.: Finalidad básica del análisis de correspondencias



Fuente: Adaptado de Clausen (1998)

© Ediciones Paraninfo

Estadísticos e-Books & Papers

en Hoffman y Franke (1986).

Son bastantes los paquetes de `R` que abordan el análisis de correspondencias. A lo largo de este capítulo nos centraremos fundamentalmente en el paquete `ca`[1] (Nenadic y Greenacre, 2007), `FactoMineR`[2] (Lê *et al.*, 2008) y, para mejorar la visualización de algunos gráficos, `factoextra`[3].

5.2. Funcionamiento del análisis de correspondencias

Como hemos venido haciendo en capítulos anteriores, utilizaremos un ejemplo sencillo para ilustrar el funcionamiento de la técnica. Una vez conocido el mecanismo de cálculo de la misma, será aplicada a un ejemplo más complejo.

Caso 5.1. Relación entre el hábito de fumar y la responsabilidad en el puesto de trabajo

Consideremos el caso, con datos ficticios, propuesto Greenacre (1984). Tras la publicación de los datos sobre los peligros del tabaco el CEO de una gran empresa realiza una encuesta sobre el hábito de fumar entre sus empleados. Realiza un muestreo aleatorio estratificado entre las principales categorías en que está clasificado el personal y se les pregunta por el hábito con posibles respuestas: (a) no fuma, (b) fuma entre 1-10 cigarrillos al día, (c) fuma entre 11-20 cigarrillos al día o (d) fuma más de 20 cigarrillos al día, categorías a las que denomina no fumadores, fumadores de intensidad baja, media y alta. La muestra final es de 193 empleados y los resultados aparecen ilustrados en el cuadro 5.1. El objetivo es saber si el hábito está más extendido entre unas categorías profesionales u otras.

Parece obvio que el simple cálculo de los porcentajes horizontales y verticales del cuadro 5.1 o una representación gráfica como la de la figura 5.2 nos permitiría responder a esta pregunta, pero ¿y si la tabla de contingencia fuera de 12×12? El análisis de correspondencias demuestra su utilidad en tablas cruzadas de mayor tamaño, pero para explicar su funcionamiento creemos más didáctico trabajar sobre una más reducida como la que se ofrece en el cuadro 5.1.

El análisis de correspondencias comienza transformando las frecuencias de la tabla de contingencia en porcentajes fila y columna, lo que se conoce, respectivamente, como perfiles fila y columna (*row profiles*, *column profiles*). El cuadro 5.2 muestra la salida de la función `CrossTable{gmodels}`, donde se ofrece la mencionada información calculada sobre los datos del cuadro 5.1. Así, el perfil fila del hábito de los directivos senior (SM), se obtiene simplemente dividiendo (4/11)=0,364; (2/11)=0,182, etc.

[1] https://cran.r-project.org/package=ca
[2] https://cran.r-project.org/package=FactoMineR
[3] https://cran.r-project.org/package=factoextra

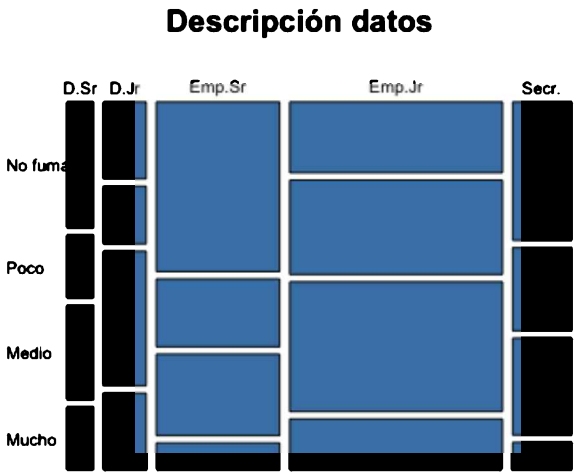
Estadísticos e-Books & Papers

Cuadro 5.1.: Relación entre puesto de trabajo y hábito de fumar

Puesto	Intensidad del hábito				Totales
	No fuma	Baja	Media	Alta	
Directivos senior	4	2	3	2	11
Directivos junior	4	3	7	4	18
Empleados senior	25	10	12	4	51
Empleados junior	18	24	33	13	88
Secretarios	10	6	7	2	25
Totales	61	45	62	25	193

Fuente: Greenacre (1984, p. 55)

Figura 5.2.: Gráfico descriptivo de los datos



Fuente: Adaptado de Clausen (1998)

© Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 5.2.: Perfiles fila y columna de la matriz de datos

	No fuma	Poco	Medio	Mucho	Total
D.Sr	4	2	3	2	11
	0.364	0.182	0.273	0.182	0.057
	0.066	0.044	0.048	0.080	
D.Jr	4	3	7	4	18
	0.222	0.167	0.389	0.222	0.093
	0.066	0.067	0.113	0.160	
Emp.Sr	25	10	12	4	51
	0.490	0.196	0.235	0.078	0.264
	0.410	0.222	0.194	0.160	
Emp.Jr	18	24	33	13	88
	0.205	0.273	0.375	0.148	0.456
	0.295	0.533	0.532	0.520	
Secr.	10	6	7	2	25
	0.400	0.240	0.280	0.080	0.130
	0.164	0.133	0.113	0.080	
Total	61	45	62	25	193
	0.316	0.233	0.321	0.130	

Pero cada perfil no puede tener el mismo peso en la configuración del mapa porque cada uno de ellos va asociado a un número diferente de casos. Así, como se observa en el cuadro 5.1, el perfil de los directivos senior (SM) está generado por 11 personas, mientras que el de los empleados junior viene de recoger la información en 88 casos. Esta ponderación de cada perfil la recoge el concepto de **masa** (*mass*). En la tabla cruzada 5.2 se observan claramente estos porcentajes (11/193=0,057) para los directivos senior, (18/193=0,093) para los directivos junior, etc. Pero esta información es clave para el desarrollo del análisis de correspondencias, luego deberá ser proporcionada por los paquetes específicos (`ca, FactoMineR`) y no obtenida al margen de los mismos en tablas cruzadas. Así si solicitamos la ejecución del CA mediante el primero de estos paquetes:

```
library(ca)
datos<-data("smoke")
ca(datos) summary(ca(datos),scree = TRUE,
rows=TRUE, columns=TRUE)
```

La salida (cuadro 5.3) nos ofrece las masas de cada fila y columna que, fácilmente, puede comprobarse que coinciden con los marginales del cuadro 5.2.

Los perfiles, por ejemplo, fila de cada tipo empleado en función de sus hábitos, podrían considerarse como vectores y ser así representados como puntos

Estadísticos e-Books & Papers

Cuadro 5.3.: Masas, inercia y coordenadas estandarizadas de los objetos representados

```
Rows:
             D.Sr      D.Jr    Emp.Sr   Emp.Jr     Secr.
Mass     0.056995  0.093264  0.264249 0.455959  0.129534
ChiDist  0.216559  0.356921  0.380779 0.240025  0.216169
Inertia  0.002673  0.011881  0.038314 0.026269  0.006053
Dim. 1  -0.240539  0.947105 -1.391973 0.851989 -0.735456
Dim. 2  -1.935708 -2.430958 -0.106508 0.576944  0.788435


 Columns:
          No fuma     Poco    Medio     Mucho
Mass     0.316062 0.233161 0.321244  0.129534
ChiDist  0.394490 0.173996 0.198127  0.355109
Inertia  0.049186 0.007059 0.012610  0.016335
Dim. 1  -1.438471 0.363746 0.718017  1.074445
Dim. 2  -0.304659 1.409433 0.073528 -1.975960
```

en un espacio tetradimensional del hábito. Cuanto más parecidos fueran los perfiles de cada tipo de empleado, más próximos aparecerían los puntos en ese hipotético mapa. El vector con las ponderaciones de cada columna (masas) [0,316 0,233 0,321 0,130] no es sino un promedio de los perfiles fila de los empleados, y puede considerarse como el centroide en esa hipotética representación tetradimensional. Las masas de las columnas son los perfiles promedios de las filas y viceversa.

Nuestra intención es que aquellos tipos empleados que tengan un perfil más parecido estén más cercanos en el espacio tetradimensional (posteriormente veremos cómo se produce su representación en un espacio dimensional más reducido de dos dimensiones que, por ello, es más fácilmente interpretable). Esto nos lleva necesariamente a calcular la distancia entre los distintos perfiles.

La matriz de distancias entre los perfiles se calcula recurriendo a las **distancias ji-cuadrado** entre los vectores de los perfiles fila (porque vamos a ilustrar con los tipos de empleados en función de sus hábitos, se procede análogamente para los perfiles fila). Una distancia ji-cuadrado es, simplemente, una distancia euclídea donde cada elemento del vector se pondera por la inversa de su masa correspondiente, esto es:

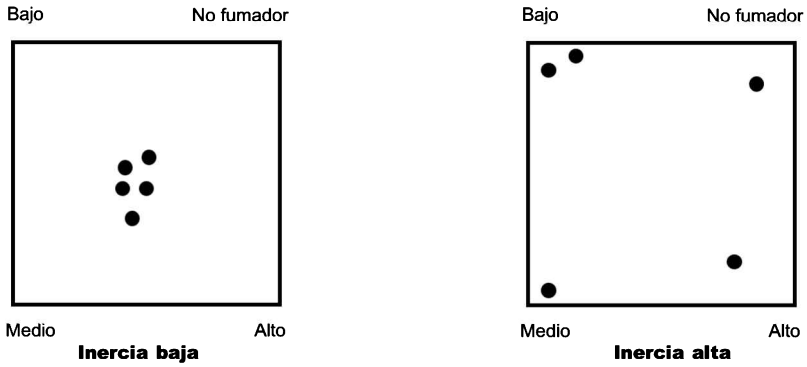
$$d(h,h') = \sqrt{\sum_j \frac{(a_{hj} - a_{h'j})^2}{a_{.j}}} \tag{5.1}$$

donde $d(h,h')$ es la distancia entre los puntos h y h' (digamos entre los directivos senior y los directivos junior), a_{jh} son los elementos del vector perfil fila del punto h, que es de dimensión J (cuatro en nuestro caso), y $a_{.j}$ son los elementos del centroide (masa). Con los datos de los cuadros 5.2 y 5.3, la distancia entre esos dos tipos de directivos se obtendría de la siguiente forma:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 5.3.: Ilustración geométrica del concepto de inercia



Fuente: Adaptado de Greenacre (1984)

$$d(h,h') = \sqrt{\frac{(0,364-0,222)^2}{0,316} + \cdots + \frac{(0,182-0,222)^2}{0,130}} = 1,496$$

Lo que el *software* proporciona no es la matriz de distancias entre los distintos perfiles, sino la distancia de cada perfil al centroide. Así, por ejemplo, la distancia al centroide del perfil de los directivos senior sería:

$$d(h,h') = \sqrt{\frac{(0,364-0,316)^2}{0,316} + \cdots + \frac{(0,182-0,130)^2}{0,130}} = 0,2165$$

esa información es la que aparece, para cada perfil, como `ChiDist` en el cuadro 5.3.

Greenacre (1994) considera que hay muchas razones para la utilización de este tipo de distancias. La primera de ellas es que la división de cada término por el valor medio tiene efectos de estandarización de la varianza, compensando la elevada varianza en frecuencias de ocurrencia altas y lo contrario en las frecuencias de ocurrencia bajas, lo que haría que las primeras tuvieran peso superior en el cálculo de la distancia.

El siguiente concepto importante para la interpretación del análisis de correspondencias es el concepto de **inercia**, en el que concurren los dos conceptos anteriores: la masa y la distancia entre cada perfil y el perfil promedio. La *inercia* es una medida de la dispersión de los perfiles en el espacio multidimensional. Cuanto mayor sea esta inercia, más alejados estarán unos de otros los puntos que representan a cada tipo de empleado, tal y como se ilustra hipotéticamente en la figura 5.3.

Estadísticos e-Books & Papers

Cuadro 5.4.: Inercia total y autovalores

```
Principal inertias (eigenvalues):

 dim    value      %   cum%   scree plot
 1      0.074759  87.8  87.8  **********************
 2      0.010017  11.8  99.5  ***
 3      0.000414   0.5 100.0
        -------- -----
 Total: 0.085190 100.0
```

Cuanto más cercano esté un punto representando un perfil fila (por ejemplo, los directivos senior) a uno de los vértices que representan la intensidad del hábito (los grandes fumadores), mayor correspondencia o asociación habrá entre las filas y columnas asociadas, que es, precisamente, el objetivo del análisis que estábamos efectuando. La *inercia* se calcula para cada perfil como el producto de la masa (w_h) por el cuadrado de la distancia ji-cuadrado de ese perfil al perfil promedio. Por ejemplo, para los directivos senior, su **contribución a la inercia total** sería:

$$I_{DSr} = 0,057 \times \left[\frac{(0,364 - 0,316)^2}{0,316} + \cdots + \frac{(0,182 - 0,130)^2}{0,130} \right] = 0,002673$$

que es la información que aparece en el cuadro 5.3 bajo la etiqueta de `Inertia`. Su total aparece en el cuadro 5.4 y podría obtenerse sumando las inercias de los perfiles fila y coincidiría con la suma de las inercias con los perfiles columna.

Como señalábamos al principio al explicitar el objetivo del análisis de correspondencias, se trata de hallar un espacio dimensional reducido (un plano si es posible) que mantenga lo más inalterablemente que sea posible las distancias ji-cuadrado relativas entre los distintos perfiles. Aunque en el epígrafe siguiente entraremos en detalles computacionales, ahora, siguiendo a Greenacre (1994), trataremos de ilustrar cómo el análisis de correspondencias encuentra ese plano que preserva en la medida de lo posible las distancias relativas.

Buscamos el plano que, conteniendo al centroide, esté lo más próximo posible a los perfiles. Habíamos definido la inercia (I_h) de un perfil fila h como el producto de su masa (w_h) por el cuadrado de su distancia ji-cuadrado al perfil promedio (d_{h^2}). Como se ilustra en la figura 5.4, esta distancia puede descomponerse como la distancia del perfil a su proyección en el plano y la distancia de esta proyección al centroide, cumpliéndose el teorema de Pitágoras:

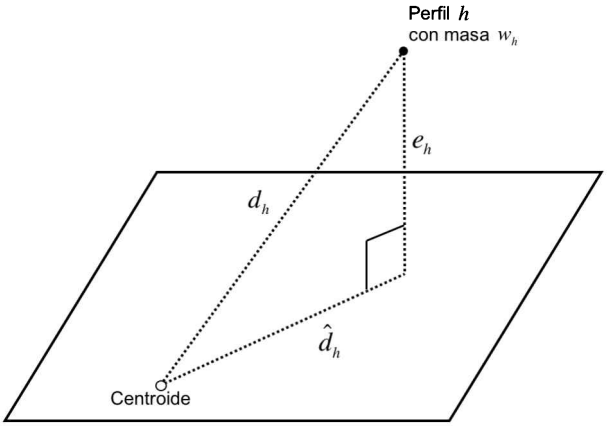
$$d_h^2 = \hat{d}_h^2 + e_h^2$$

por lo que la inercia total sería:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 5.4.: Descomposición de la inercia total



Fuente: Adaptado de Greenacre (1994)

$$I = \sum_{h=1}^{H} I_h = \sum_{h=1}^{H} w_h d_h^2 = \sum_{h=1}^{H} w_h \hat{d}_h^2 + \sum_{h=1}^{H} w_h e_h^2$$

esto es, se descompone entre la parte de la inercia contenida en el plano y la inercia residual. La proximidad entre los perfiles y el plano se mide pues en términos de mínimos cuadrados ponderados, mediante la inercia residual, que deberá ser minimizada en el proceso de cálculo del análisis de correspondencias (o análogamente maximizada la inercia en el plano).

El porcentaje que suponga la inercia del plano sobre la inercia total será un indicador de la bondad del ajuste de la solución obtenida. Como se observa en el cuadro 5.4, el plano de una solución bidimensional recogería el 99,5 % de la inercia total, de tal forma que la representación de los perfiles sobre el plano daría una visión muy adecuada de las distancias originales. Por ello, el añadir una tercera dimensión a la solución, además de ofrecer un mapa de mayor complejidad de interpretación, aporta nada más que el 0,5 % al poder explicativo del modelo.

También es de interés entender por qué la mayoría de programas ofrece, al realzar un CA, un test estadístico ji-cuadrado como indicador de la bondad del ajuste (así ocurre con `FactoMineR`, aunque no con `ca`). El estadístico ji-cuadrado sirve para contrastar la hipótesis nula de independencia entre las dos variables que conforman la tabla de contingencia. Si no pudiera rechazarse esta independencia, no habría asociación significativa entre los distintos niveles de las variables consideradas, puesto en la empresa y hábito de fumar, por lo que las frecuencias se repartirían con elevada homogeneidad por todas las celdas y la inercia total sería baja. En el ejemplo que hemos propuesto, no se puede rechazar la hipótesis nula de independencia y tendría un sentido dudoso pasar

Estadísticos e-Books & Papers

Cuadro 5.5.: Test χ^2

```
The chi square of independence between the two variables
is equal to 16.44164 (p-value =  0.1718348 ).
```

a analizar los resultados ofrecidos por la técnica utilizada, pero ha de tenerse en cuenta que existen en la tabla muchas celdas con frecuencia inferior a 5 y el test no es estrictamente fiable. La solicitud del análisis con `FactoMineR` sería la siguiente, mientras que el resultado del test se ofrece en el cuadro 5.5:

```
library(FactoMineR)
summary(CA(datos,ncp=2,graph=TRUE))
```

Con la información presentada hasta este momento, nos restaría únicamente mostrar el mapa bidimensional que los paquetes utilizados ofrecen para el ejemplo propuesto. Las coordenadas de cada perfil necesarias para dibujar el mapa ya se ofrecieron en el cuadro 5.3 y aparecen bajo las etiquetas `Dim. 1` y `Dim. 2`, tanto para dibujar los puntos fila como los puntos columna. El procedimiento que sigue el CA para su cálculo se mostrará de manera detallada en la sección siguiente y responde básicamente a una descomposición en autovalores, similar a la que veremos en el tema dedicado al análisis de componentes principales. Utilizando estas coordenadas, el mapa obtenido es el recogido en la figura 5.5. En la misma se observa una mayor asociación de la ausencia del hábito de fumar entre los empleados senior principalmente, pero también entre los secretarios, mientras que en los directivos junior el hábito es intenso.

Para concluir con el análisis del ejemplo que hemos sometido al análisis de correspondencias resta analizar la calidad de la representación obtenida, no en términos globales, dado que, en el cuadro 5.4, esta ya quedó constatada porque prácticamente el 100 % de la inercia aparecía recogida en las dos dimensiones representadas, sino de cada punto (perfil) en concreto. Esta tarea se realiza mediante el análisis de la contribución de los puntos a la inercia de la dimensión que aparece en el cuadro 5.6 para los perfiles filas y los perfiles de columna (etiquetados como `ctr`, `k=1` y `k=2` señalan la dimensión). El hecho de que la contribución al eje en la dimensión 1 se deba en más del 80 % a los empleados senior (51,2 %) y junior (33,1 %) (filas) y a los no fumadores (65,4 %) (columnas), hay que interpretar que son estos los puntos que han ejercido mayor influencia en que ese eje tenga la orientación que tiene. Tampoco es sorprendente el resultado en la medida en que son los puntos con mayor masa y, por ello, mayor capacidad de influencia. No siempre es así, a veces puntos con poca masa ejercen una gran influencia debido a su elevada distancia al centroide.

Bajo la etiqueta `cor` la salida ofrece la contribución de la dimensión a la inercia del punto que podemos interpretar de manera similar a la comunalidad en un análisis factorial, es decir, mide la calidad con la que el punto está re-

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

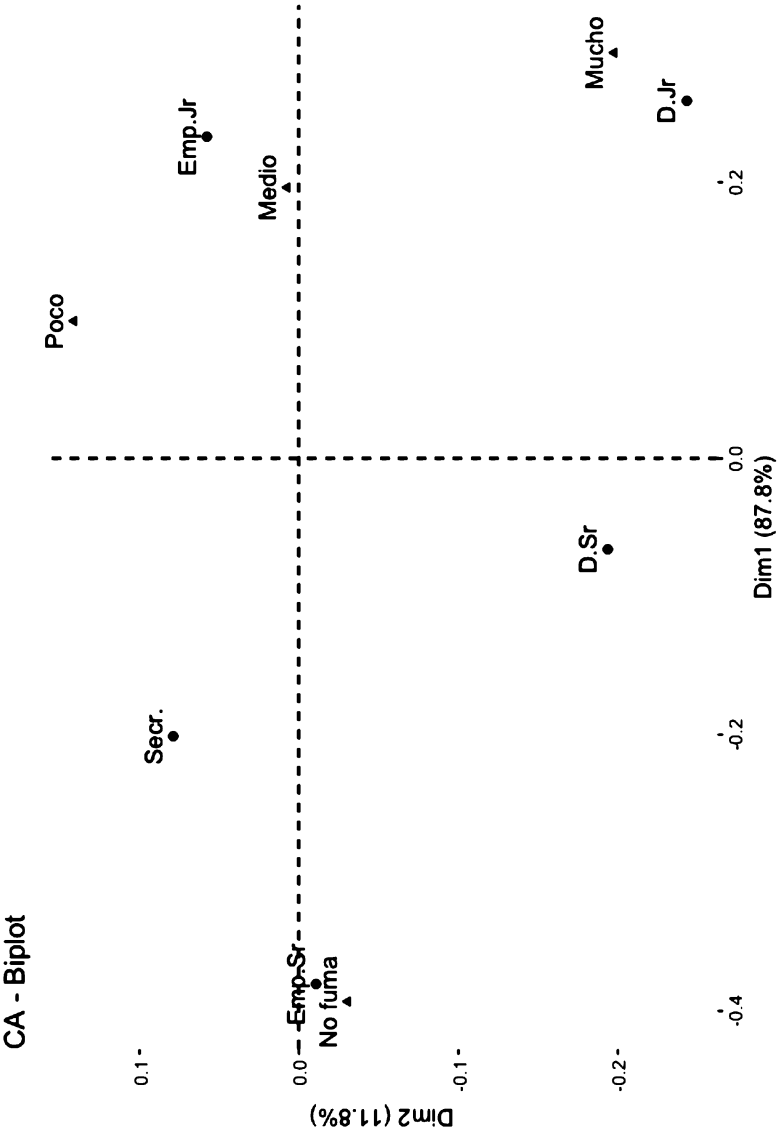


Figura 5.5.: Representación de la solución bidimensional

Estadísticos e-Books & Papers

Cuadro 5.6.: Contribución de los puntos a las dimensiones, de las dimensiones a los puntos y coordenadas para la representación

```
Rows:
    name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
1 |  DSr |   57  893   31 |  -66  92   3 | -194 800 214 |
2 |  DJr |   93  991  139 |  259 526  84 | -243 465 551 |
3 | EmpS |  264 1000  450 | -381 999 512 |  -11   1   3 |
4 | EmpJ |  456 1000  308 |  233 942 331 |   58  58 152 |
5 | Secr |  130  999   71 | -201 865  70 |   79 133  81 |

Columns:
    name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
1 | Nofm |  316 1000  577 | -393 994 654 |  -30   6  29 |
2 | Poco |  233  984   83 |   99 327  31 |  141 657 463 |
3 | Medi |  321  983  148 |  196 982 166 |    7   1   2 |
4 | Much |  130  995  192 |  294 684 150 | -198 310 506 |
```

presentado en la solución (contribución de la dimensión a la inercia del punto). Así si tomamos como ejemplo a los directivos senior, estos están pobremente representados por la dimensión 1 (0,092), pero sí lo están en la dimensión 2 (0,800), en su conjunto la solución bidimensional representa muy bien al punto, tal como muestra bajo la etiqueta `qlt` $(0,893 = 0,092 + 0,800)$. El valor de la contribución de cada dimensión es el coseno del ángulo que forma el vector que une el centroide con el punto respecto al plano bidimensional ajustado, tal como hemos representado en la figura 5.4. Así si la representación de los directivos senior tiene una correlación (`cor`) de 0,093, el ángulo será $\arccos(0,093) = 84,6°$, es decir, muy alejado del plano y por ello, como veíamos, mal representado.

5.3. Fundamentación matemática del análisis de correspondencias

El análisis de correspondencias se basa matemáticamente en la descomposición en valores singulares, que es una generalización de la descomposición en vectores propios que se ofrecerá al presentar el análisis de componentes principales. Una versión mucho más detallada de la que aquí se presenta, basada en Blasius y Greenacre (1994) y en Weller y Romney (1990), puede encontrarse en Greenacre (1984).

La descomposición en valores singulares (SVD: *Singular Value Descomposition*) consiste en la descomposición de una matriz rectangular $\mathbf{A}$ de orden $H \times J$ en el producto de tres matrices:

$$A = \mathbf{U}\mathbf{\Gamma}\mathbf{V}' \tag{5.2}$$

donde la matriz $\mathbf{\Gamma}$ es una matriz diagonal de números positivos en orden descendente $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_K > 0$, K es el rango de $\mathbf{A}$ y las columnas de las matrices $\mathbf{U}$ y $\mathbf{V}$ son ortonormales, esto es: $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}$. Los K números

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

$\gamma_1, \gamma_2 \cdots \gamma_K$ se denominan valores singulares (*singular values*), las K columnas de $\mathbf{U}$ se denominan vectores singulares izquierdos (*left singular vectors*) y las K columnas de $\mathbf{V}$, vectores singulares derechos.

Consideremos el caso general de un conjunto de H puntos en un espacio J dimensional, cuyas coordenadas son las filas de una matriz que denominaremos $\mathbf{Y}$. Cada uno de estos puntos (perfiles fila) tiene las siguientes masas, tal como se definieron en el apartado anterior: $w_1, w_2, ..., w_H$. A las masas asociadas a los perfiles columna las denotaremos como $q_1, q_2, ..., q_J$. Denotemos como $\mathbf{D_w}$ y $\mathbf{D}_q$ a las matrices diagonales que contienen a los valores anteriores y a $\mathbf{w}$ y $\mathbf{q}$ si en lugar de expresar las masas de los puntos en forma de matriz diagonal lo hacemos, respectivamente, en forma de vector.

Pues bien, el resultado general es que cualquier mapa de dimensión menor a J que contenga a los puntos señalados puede derivarse directamente mediante descomposición de valores singulares del siguiente modo:

$$A = \mathbf{D_w^{1/2}} \left(\mathbf{Y} - \mathbf{1}\bar{\mathbf{y}}'\right) \mathbf{D_q^{1/2}} \tag{5.3}$$

donde toda la notación es conocida salvo $\mathbf{1}$, que es un vector de 1, y que $\bar{\mathbf{y}}'$ es el centroide de las filas de la matriz $\mathbf{Y}$. Si expresamos la descomposición como en (5.2), entonces las coordenadas de los puntos en el espacio multidimensional reducido están contenidas en la matriz $\mathbf{F}$, que se obtiene:

$$\mathbf{F} = \mathbf{D_w^{1/2}} \mathbf{U} \mathbf{\Gamma} \tag{5.4}$$

por ejemplo, las coordenadas de los puntos en un espacio reducido bidimensional serían las dos primeras columnas de $\mathbf{F}$.

El resultado general anterior se traduce en dos casos particulares en el análisis de correspondencias, conocidos como problema filas y problema columnas, dado que los puntos que se someten a la descomposición pueden ser tanto H puntos de dimensión J o J puntos de dimensión H. Como veremos a continuación, ambos problemas se reducen a la descomposición en valores singulares de una misma matriz, llamada de residuos estandarizados. Ilustremos con el ejemplo que hemos venido utilizando lo expuesto hasta el momento.

Partimos de la tabla de contingencia del cuadro 5.1, que denotamos como matriz $\mathbf{X}$ de dimensiones $H \times J$, en este caso 5×4. Los vectores que contienen la suma de las filas [11 18 51 88 25] y las columnas [61 45 62 25] y el gran total (193) también aparecen en ese cuadro. Las masas de las filas $(w_1, w_2, ..., w_H)$ se obtienen simplemente dividiendo las sumas de las filas por el gran total. Así el primer elemento del panel a) del cuadro 5.7 se ha obtenido $11/193 = 0{,}057$, y el primer elemento del panel b) de ese cuadro, dividiendo $61/193 = 0{,}316$. Ambos cuadros ya estaban presentes en la salida del cuadro 5.2.

La matriz de correspondencias $\mathbf{P}$ se define como la matriz de frecuencias original $\mathbf{X}$ dividida por el gran total N, $\mathbf{P} = (1/N)\mathbf{X}$. Así, el primer elemento de la matriz $\mathbf{P}$ (cuadro 5.8) se obtiene como $4/193 = 0{,}021$. Nótese que la suma de las filas y columnas de la matriz $\boldsymbol{P}$ son, respectivamente, las masas fila y masas columna.

Estadísticos e-Books & Papers

Cuadro 5.7.: Masas fila y columna

Masas fila	
Directivos senior	0,057
Directivos junior	0,093
Empleados senior	0,264
Empleados junior	0,456
Secretarios	0,130

Masas columna	
No fuma	0,316
Poco	0,233
Medio	0,321
Mucho	0,130

Cuadro 5.8.: Matriz de correspondencias

	No	Poco	Medio	Alto
Directivos senior	0,021	0,010	0,016	0,010
Directivos junior	0,021	0,016	0,036	0,021
Empleados senior	0,130	0,052	0,062	0,021
Empleados junior	0,093	0,124	0,171	0,067
Secretarios	0,052	0,031	0,036	0,010

© Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 5.9.: Perfiles fila

	No	Poco	Medio	Alto	Total
Directivos senior	0,364	0,182	0,273	0,182	1,000
Directivos junior	0,222	0,167	0,389	0,222	1,000
Empleados senior	0,490	0,196	0,235	0,078	1,000
Empleados junior	0,205	0,273	0,375	0,148	1,000
Secretarios	0,400	0,240	0,280	0,080	1,000

Cuadro 5.10.: Perfiles columna

	No	Poco	Medio	Alto
Directivos senior	0,066	0,044	0,048	0,080
Directivos junior	0,066	0,067	0,113	0,160
Empleados senior	0,410	0,222	0,194	0,160
Empleados junior	0,295	0,533	0,532	0,520
Secretarios	0,164	0,133	0,113	0,080
Total	1,000	1,000	1,000	1,000

La matriz que contiene los perfiles fila (cuadro 5.9) puede obtenerse como la matriz de correspondencias $\mathbf{P}$ dividida por sus respectivas masas fila, lo que puede expresarse en notación matricial como $\mathbf{D_w^{-1}P}$, donde $\mathbf{D_w^{-1}}$ es la matriz diagonal de las masas fila. El procedimiento es análogo para los perfiles columna (cuadro 5.10). Si llamamos $\mathbf{D_q}$ a la matriz diagonal que contiene las masas columna en la diagonal, se obtendría como $\mathbf{PD_q^{-1}}$.

Así pues, el problema fila se reduce, como vimos en (5.2), a descomponer una matriz $\mathbf{A}$ que, de acuerdo con (5.3):

$$\mathbf{A} = \mathbf{D_w^{1/2}} \left(\mathbf{D_w^{-1}P} - \mathbf{1q'}\right) \mathbf{D_q^{-1/2}} \tag{5.5}$$

donde toda la notación es conocida salvo el traspuesto del vector de masas columna $\mathbf{q}$', que recordemos que coincide con el centroide de las filas. La expresión anterior, dado que dicho centroide se calcula del siguiente modo:

$$\mathbf{w'D_q^{-1}P} = \mathbf{1'P} = \mathbf{q'}$$

puede ponerse como:

$$\mathbf{A} = \mathbf{D_w^{-1/2}} \left(\mathbf{P} - \mathbf{wq'}\right) \mathbf{D_q^{-1/2}} \tag{5.6}$$

Por otro lado, el problema columna es totalmente análogo. El centroide de las columnas es ahora el traspuesto del vector de masas fila $\mathbf{w}$', que se obtiene del siguiente modo:

$$\mathbf{q'D_q^{-1}P} = \mathbf{1'P} = \mathbf{w'}$$

Estadísticos e-Books & Papers

Cuadro 5.11.: Matriz de residuos estandarizados

	No	Poco	Medio	Alto
Directivos senior	0,020	-0,025	-0,020	0,035
Directivos junior	-0,051	-0,042	0,036	0,079
Empleados senior	0,159	-0,039	-0,078	-0,073
Empleados junior	-0,134	0,055	0,064	0,034
Secretarios	0,054	0,005	-0,026	-0,050

y la matriz $\mathbf{A}$ que hay que descomponer es ahora:

$$\mathbf{A} = \mathbf{D_q^{1/2}} \left(\mathbf{D_q^{-1}P'} - \mathbf{1w'}\right) \mathbf{D_w^{-1/2}} \tag{5.7}$$

que, operando:

$$\mathbf{A} = \mathbf{D_q^{-1/2}} \left(\mathbf{P'} - \mathbf{qw'}\right) \mathbf{D_w^{-1/2}} \tag{5.8}$$

que es la traspuesta de la matriz derivada en (5.6), luego el problema de las filas es idéntico al de las columnas y se reduce a descomponer por valores singulares la misma matriz que se denomina, como ya indicamos, matriz de residuos estandarizados. Cada elemento de dicha matriz, si se prefiere expresado de manera algebraica, se obtendría:

$$a_{hj} = \frac{p_{hj} - w_h q_j}{\sqrt{w_h q_j}} \tag{5.9}$$

donde p_{hj} es el elemento h_j de la matriz de correspondencias $\mathbf{P}$, w_h es el elemento h del vector de masas fila, y q_j, el elemento j del vector de masas columna; de esta forma, en nuestro ejemplo, el primer elemento de dicha matriz sería:

$$a_{11} = \frac{0,021 - 0,057 \times 0,316}{\sqrt{0,057 \times 0,316}} = 0,020$$

repitiendo el cálculo para todos los elementos, la matriz quedaría de la siguiente forma:

Es importante señalar que, si sumáramos los cuadrados de estos residuos, obtendríamos la inercia total (0,085) y, si esta inercia se multiplicara por la frecuencia total (193), obtendríamos el estadístico ji-cuadrado (16,44) de la tabla de contingencia que vimos en el cuadro 5.5 y que ya vimos que era un indicador de la pertinencia del análisis en el caso de rechazarse la hipótesis nula de independencia:

$$\chi^2 = \left(\sum_h \sum_j a_{hj}^2\right) N = 0,085 \times 193 = 16,44$$

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 5.12.: Matriz de valores singulares ($\mathbf{\Gamma}$) y vectores singulares izquierdos ($\mathbf{U}$) y derechos ($\mathbf{V}$)

```
$d
[1] 2.734192e-01 1.000819e-01 2.033969e-02 5.024458e-06

$u
            [,1]        [,2]       [,3]        [,4]      [,5]
[1,] -0.05742425 -0.46210057  0.8333329  0.11164793 0.2761459
[2,]  0.28926053 -0.74239266 -0.5060816  0.04385465 0.3273220
[3,] -0.71554395 -0.05475551 -0.1304269 -0.56718037 0.3824839
[4,]  0.57529040  0.38961423  0.1097188 -0.38624835 0.5966705
[5,] -0.26470477  0.28375832 -0.1427879  0.71744697 0.5606194

$v
           [,1]        [,2]        [,3]      [,4]
[1,] -0.8087034 -0.17127154 -0.02469825 0.5621876
[2,]  0.1756443  0.68056247  0.52225563 0.4829419
[3,]  0.4069508  0.04169165 -0.71521349 0.5666767
[4,]  0.3867026 -0.71116987  0.46379813 0.3599858
```

Pues bien, es precisamente la matriz $\mathbf{A}$ la que hay que someter a la descomposición en valores singulares indicada en (5.2). Obtenemos mediante R las matrices con los valores singulares ($\mathbf{\Gamma}$) y los vectores singulares izquierdos ($\mathbf{U}$) y derechos ($\mathbf{V}$). Los resultados de la descomposición se ofrecen en el cuadro 5.12 y se han obtenido mediante la sintaxis:

```
a1<-c( 0.02020,-0.02538,-0.02044, 0.03468)
a2<-c(-0.05098,-0.04205, 0.03645, 0.07865)
a3<-c( 0.15922,-0.03948,-0.07795,-0.07299)
a4<-c(-0.13394, 0.05533, 0.06404, 0.03413)
a5<-c( 0.05374, 0.00510,-0.02619,-0.04953)
A<-t(cbind(a1,a2,a3,a4,a5))
svd(A,n=5,nv=4)
```

Los valores singulares que aparecen en la primera fila del cuadro 5.12 sirven para descomponer la inercia total entre las dimensiones posibles de la solución que, recordemos, pueden ser hasta $K = min(I-1,\ J-1)$, esto es $min(5-1,\ 4-1) = 3$ en nuestro caso. Así, la inercia correspondiente a la primera dimensión es el cuadrado del primer valor singular ($0,2734^2 = 0,0747$) que se ofrecía en el cuadro 5.4. Normalmente, como en el análisis de componentes principales, estos valores se expresan como porcentajes de la inercia total, como se veía en ese mismo cuadro 5.4.

Llegados a este punto, es el momento de obtener las **coordenadas de cada perfil** en los tres ejes resultantes de la descomposición, independientemente de que, para facilitar la interpretación, el investigador decida retener solo dos (lo que es razonable dada la escasa proporción de inercia que explica el tercero). Esta es la información que denominábamos "puntuación en la dimensión" y que se ofrecía en el cuadro 5.3 en su versión estándar y en el cuadro 5.6 en su versión principal, que es la utilizada en la representación. Para el caso de las filas, por ejemplo, estas coordenadas se obtendrían mediante la expresión

Estadísticos e-Books & Papers

(5.4). Algebraicamente, cualquier componente de la matriz que ofrece dichas coordenadas respondería a la expresión:

$$f_{hk} = \frac{u_{hk}\gamma_k}{\sqrt{w_h}}$$

donde toda la notación es conocida, pero la recordamos: u_{hk} es el elemento hk de la matriz que contiene los vectores singulares izquierdos (cuadro 5.12), γ_k es el valor singular correspondiente a la dimensión k (cuadro 5.12) y w_h es la masa del perfil fila h (cuadro 5.7). Así, el primer elemento de la matriz que recoge las coordenadas (cuadro 5.12) se obtiene de esta forma:

$$f_{11} = \frac{-0,05742425 \times 0,2734192}{\sqrt{0,056995}} = -0,066$$

que vemos coincide con la coordenada que ofrece el cuadro 5.6, aunque ese cuadro ofrece las coordenadas en valores multiplicados por 1000 y se visualiza como -66.

Para el caso de las columnas, el procedimiento es análogo, simplemente que operando del siguiente modo:

$$\mathbf{G} = \mathbf{D_q^{-1/2}V\Gamma}$$

Por lo tanto, un elemento genérico g_{hj} de $\mathbf{G}$ es igual a:

$$g_{jk} = \frac{v_{jk}\gamma_k}{\sqrt{q_j}}$$

donde la notación es equivalente a la ofrecida para las filas, v_{jk} es el elemento jk de la matriz que contiene los vectores singulares derechos y que tenemos recogidos en el cuadro 5.12, γ_k es el valor singular correspondiente a la dimensión k (cuadro 5.12) y q_j es la masa del perfil columna j (cuadro 5.7). Así, el primer elemento de la matriz que recoge las coordenadas (cuadro 5.12), se obtiene del siguiente modo:

$$g_{11} = \frac{-0,8087034 \times 0,2734192}{\sqrt{0,316062}} = -0,393$$

que también vemos que coincide con el ofrecido en el cuadro 5.6, aunque una vez más multiplicado por 1000 (-393).

Como hemos señalado, esas son las coordenadas principales que se utilizan en la representación, sin embargo, el paquete `ca` ofrece también las **coordenadas estandarizadas,** que, debido a que son las que se van a utilizar posteriormente para calcular la contribución de los puntos a las dimensiones y de las dimensiones a los puntos, es necesario derivar. La coordenada estandarizada se obtiene, simplemente, dividiendo por el autovalor de la dimensión la coordenada principal que acabamos de derivar. Por ejemplo, para las coordenadas en las dos primeras dimensiones de los directivos senior (fila) y de los no fumadores (columna) sería:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

$$f_{11}^s = \frac{f_{11}}{\gamma_1} = \frac{-0,066}{0,2734192} = -0,241$$

$$f_{12}^s = \frac{f_{12}}{\gamma_2} = \frac{-0,194}{0,1000819} = -1,936$$

$$g_{11}^s = \frac{g_{11}}{\gamma_1} = \frac{-0,393}{0,2734192} = -1,438$$

$$g_{12}^s = \frac{f_{12}}{\gamma_2} = \frac{-0,030}{0,1000819} = -0,304$$

Donde el lector puede comprobar que las coordenadas estandarizadas coinciden con las que muestra `ca` en el cuadro 5.3, los datos de los autovalores proceden del cuadro 5.12 y las coordenadas principales del cuadro 5.6, bajo las etiquetas `k=1` y `k=2`.

Para terminar de obtener toda la información necesaria para interpretar un análisis de correspondencias solo restaría por indicar cómo calcular las **contribuciones de la dimensión a la inercia del punto y del punto a la inercia de la dimensión** (cuadro 5.6), que, recordemos, nos indicaban, respectivamente, la calidad de la representación de cada punto en la solución propuesta, al mismo tiempo que sirven de guía para interpretar tentativamente las dimensiones resultantes.

Para ambos casos es necesario descomponer la inercia de cada dimensión dada en el cuadro 5.4 para cada punto fila y columna. La inercia total se obtendrá de la suma por filas y columnas de los siguientes elementos (ambas sumas coincidirán):

$$\lambda_{hk} = w_h (f_{hk}^s)^2$$

$$\lambda_{jk} = q_j (g_{jk}^s)^2$$

donde w_h es la masa de la fila h y f_{hk}^s es la coordenada estandarizada de la fila h en la dimensión k. Para la inercia de los directivos senior, los elementos de cada uno de los ejes son los siguientes:

$$\lambda_{11} = 0,057 \times (-0,241)^2 = 0,003$$
$$\lambda_{12} = 0,057 \times (-1,936)^2 = 0,214$$

que puede comprobarse que se corresponden con la contribución etiquetada como `ctr` en el cuadro 5.6. Análogamente, para las columnas (por ejemplo, no fumadores):

$$\lambda_{11} = 0,316 \times (-1,438)^2 = 0,654$$
$$\lambda_{12} = 0,316 \times (-0,304)^2 = 0,029$$

Calculadas las contribuciones de los puntos a las dimensiones, nos resta por

Estadísticos e-Books & Papers

calcular las **contribuciones de las dimensiones a los puntos**, es decir, la calidad de la representación de cada punto que aparece etiquetada como `cor` en el cuadro 5.6. Nótese que en este cuadro las correlaciones no suman 100 en la medida en que hay una dimensión que no aparece en el cuadro porque se ha decidido trabajar con una solución bidimensional que recogía casi la totalidad de la inercia. El proceso es elemental, una vez hemos calculado las contribuciones. Basta calcular el porcentaje horizontal que representan dichas contribuciones si estuvieran las 3 dimensiones. Por ejemplo, centrando la atención en el cuadro 5.6 y los cálculos que acabamos de hacer, vemos que la contribución `ctr` de la fila de los directivos senior es de 3 (0,003) en la primera dimensión y de 214 (0,214) para la segunda. Faltaría el cálculo para la tercera dimensión. Calculando esos porcentajes, la primera dimensión aporta un 9,2 % (92 en la tabla) a la explicación del punto, y la segunda, un 80 % (800 en la tabla). El 1,8 % restante correspondería a la tercera dimensión no representada.

5.4. Incorporación de puntos suplementarios al análisis de correspondencias simple

En algunas ocasiones el investigador puede querer incorporar al mapa la visualización de variables que no haya incorporado para la obtención del mapa inicial. A diferencia del análisis de correspondencias múltiple (MCA), que veremos a continuación, los puntos fila o columna suplementarios se dibujan en el mapa conjunto pero no se utilizan para derivar los ejes. El objetivo puede ser intentar clarificar la interpretación de las asociaciones que se producen con variables que estén relacionadas, por ejemplo incorporando medias nacionales de alguna de las variables utilizadas para ver diferencias entre los patrones de la población y de la muestra.

Siguiendo el ejemplo de Greenacre (1984), este autor incorpora a la matriz que relaciona los puestos de trabajo en la empresa analizada y el hábito de fumar una fila adicional (el porcentaje de población en el país que no fuma, o que lo hace de manera ligera, media o alta) y añade dos columnas más. En una de ellas añade los resultados de la encuesta sobre el número de empleados que toma alcohol y, en la otra, el número que no lo hace. El cuadro 5.13 ofrece la nueva matriz de datos. Nótense dos cosas: que esa matriz tendrá valores perdidos (el cruce entre los datos nacionales del hábito y la muestra de la empresa de empleados que toman o no alcohol no tiene sentido) y, en segundo lugar, que los datos nacionales son un porcentaje, pero la forma en que se introduzca (porcentajes, proporciones) no tiene importancia debido al proceso de estandarización de filas y columnas que vimos con anterioridad.

El primer paso es incorporar a la matriz de datos esas filas y columnas suplementarias y etiquetarlas:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 5.13.: Relación entre puesto de trabajo y hábito de fumar con filas y columnas adicionales

Puesto	Intensidad del hábito				Totales	Alcohol	
	No fuma	Baja	Media	Alta		Sí	No
Directivos senior	4	2	3	2	11	0	11
Directivos junior	4	3	7	4	18	1	17
Empleados senior	25	10	12	4	51	5	46
Empleados junior	18	24	33	13	88	10	78
Secretarios	10	6	7	2	25	7	18
Totales	61	45	62	25	193	23	170
Media nacional	42	29	20	9	100	–	–

Fuente: Greenacre (1984, cuadros 3.4 y 3.5).

```
col1.sup<-c(0,1,5,10,7)
col2.sup<-c(11,17,46,78,18)
fila.sup<-c(0.42,0.29,0.20,0.09,NA,NA)

datos.sup<-cbind(datos,col1.sup,col2.sup)
colnames(datos.sup)<-c("No fuma","Poco","Medio","Mucho",
"Alch.Si","Alch.No")
datos.sup<-rbind(datos.sup,fila.sup)
rownames(datos.sup)<-c("D.Sr","D.Jr","Emp.Sr","Emp.Jr",
"Secr.","Media.Pob")
```

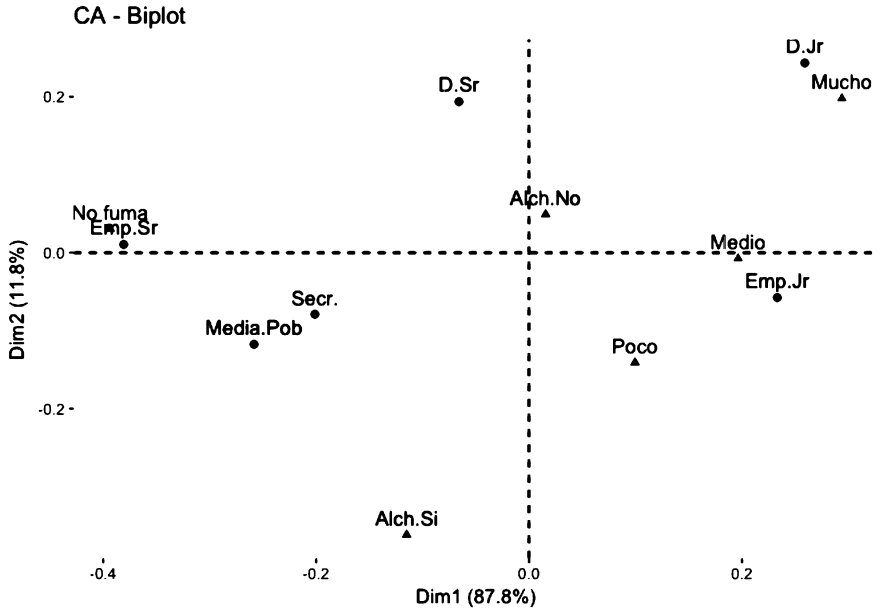
En segundo lugar, estimamos el modelo. Utilizamos en esta segunda etapa la función `CA{FactoMineR}` con el fin de introducir distintos paquetes y que el lector se pueda familiarizar con ellos. El único elemento importante de la sintaxis es que se ha de señalar cuáles son las columnas suplementarias, en nuestro caso la 5 y la 6 (`col.sup=5:6`), y la fila complementaria, en nuestro caso la 6 (`row.sup=6:6`). Indicamos que no obtenga el gráfico (`graph=FALSE`) simplemente porque queremos obtenerlo con la función `fviz_ca_biplot`, mucho más flexible y, en cierta forma, estética.

```
fit.sup <- CA (datos.sup, row.sup = 6:6,
col.sup = 5:6, graph = FALSE)
summary(fit.sup)
fviz_ca_biplot(fit.sup) + theme_grey()

# coordenadas de las filas y columnas suplementarias
fit.sup$row.sup fit.sup$col.sup
```

La figura 5.6 nos ofrece el mapa de correspondencias conjunto donde aparecen representados los puntos originales más los adicionales. Varias son las

Estadísticos e-Books & Papers

Figura 5.6.: Representación de la solución bidimensional con puntos fila y columna adicionales



conclusiones que podemos derivar. El punto con el promedio de la población está a mitad de camino entre los niveles de "no fuma" y "fuma poco", lo que indica que nuestra muestra está formada por una proporción más elevada de fumadores que el promedio nacional. Si nos fijamos en la ubicación de los puntos que señalan a los empleados que no toman alcohol «Alch.No» o que sí que lo hacen «Alch.Si», vemos que estos están orientados de manera bastante paralela a la dimensión 2 que, recordemos, ordenaba a los fumadores por intensidad (la dimensión 1 separaba a los no fumadores de los que sí que lo eran). Esta orientación paralela sugiere que sí que hay una asociación entre los dos hábitos, al menos entre los que fuman más intensamente.

Aunque ya hemos señalado que los puntos adicionales no contribuyen a la formación del eje, sí que podemos evaluar lo bien que los ejes los explican. El cuadro 5.14 nos proporciona esta información. La función `CA{FactoMineR}` nos ofrece la calidad de la representación (`cor` en las salidas del paquete `ca`) bajo la etiqueta de `cos2`. La suma de las contribuciones en cada dimensión equivaldrá a la calidad de la representación que aparecía, por ejemplo en el cuadro 5.6, bajo la etiqueta `qlt`.

Parece que la solución bidimensional representa muy bien la fila con el promedio de consumo nacional (la suma de las correlaciones en las dos dimensiones da un 76 %), mientras que en los nuevos puntos columna esta calidad, sin ser excesiva, alcanza alrededor del 40 %.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 5.14.: Calidad de la representación de los puntos fila y columna suplementarios

```
$cos2
              Dim 1     Dim 2     Dim 3
Media.Pob 0.6305782 0.1307462 0.2386756

$cos2
             Dim 1     Dim 2      Dim 3
Alch.Si 0.0401038 0.3984713 0.09576565
Alch.No 0.0401038 0.3984713 0.09576565
```

5.5. Análisis de correspondencias múltiple

A diferencia de añadir variables filas o columna a la solución de un análisis de correspondencias simple, en el análisis de correspondencias múltiple (MCA) todas las variables contribuyen a conformar los ejes. El esquema de interpretación será muy similar solo que se añade la complejidad de variables adicionales.

Los datos pueden aparecer presentados de distinta forma. Podemos tenerlos con las distintas variables categóricas en columnas y los individuos en filas (*raw data*) o como un conjunto anidado de tablas cruzadas donde todas las variables se cruzan con todas, lo que se conoce como *matriz de Burt* ($\mathbf{C}$). También pueden tenerse como *matriz de indicadores* ($\mathbf{Z}$), que tendría tantas filas como casos y las columnas no serían las variables, sino las categorías de las mismas que tomarían el valor 1 en el caso en que ese individuo tenga el atributo que representa esa categoría de la variable en cuestión, y cero en caso contrario. Ambas variables están relacionadas: $\mathbf{C} = \mathbf{Z}^{\intercal}\mathbf{Z}$ (Greenacre y Blasius, 2006). Cuando el algoritmo que hemos presentado en las secciones anteriores se aplica a una matriz de Burt o una matriz de indicadores, al método se le denomina *Multiple Correspondence Analysis* (MCA). Veremos su estimación e interpretación con el caso 5.2.

Caso 5.2. Satisfacción con los servicios prestados por un taller
Este caso se utiliza como ilustración del software XLSTAT[4]. Los datos se corresponden con una encuesta realizada por un concesionario de automóviles a 28 clientes de su taller una semana después de que recogieran su automóvil tras una reparación. Las preguntas (y las variables que se encuentran en la base de datos) fueron las siguientes:

1. [`sat`] ¿Está usted satisfecho de manera global con el servicio? (Sí/No).
2. [`soluc`] ¿Considera que el problema que tenía el automóvil ha sido resuelto? (Sí/No/No lo sé).
3. [`bienv`] ¿En qué medida fue correcta la recepción del automóvil por el taller? (Escala de 1 a 5 con 1=Muy incorrecta, 5=Muy correcta).
4. [`q_p`] ¿Fue la relación calidad/precio satisfactoria? (Sí/No).

[4]https://www.xlstat.com/en/company/about-us

Estadísticos e-Books & Papers

Cuadro 5.15.: Autovalores e inercia explicada

```
Eigenvalues
                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7   Dim.8   Dim.9
Variance                0.539   0.324   0.242   0.223   0.211   0.179   0.137   0.074   0.049
% of var.              26.968  16.222  12.076  11.126  10.526   8.966   6.858   3.711   2.443
Cumulative % of var.   26.968  43.190  55.266  66.393  76.918  85.884  92.742  96.453  98.896
                       Dim.10
Variance                0.022
% of var.               1.104
Cumulative % of var. 100.000
```

5. [rep] ¿Volverá a utilizar los servicios de nuestro taller? (Sí/No/No lo sé).

El objetivo es analizar mediante un MCA si existe algún tipo de relación entre las variables que arroje luz sobre los determinantes de la satisfacción. Utilizaremos, para este caso, la función `MCA{FactoMineR}`. La estimación del modelo es muy sencilla, tras cargar los datos, podemos plantear un análisis descriptivo de las frecuencias de las categorías de cada variable para tener una idea general de las respuestas de los entrevistados.

La figura 5.7 nos muestra que el taller tiene un problema en la medida en que, aunque mayoritariamente se considera que el problema del coche queda resuelto, la intención de repetir es muy baja y no se aprecia una buena relación calidad/precio, siendo los que se declaran satisfechos solo ligeramente más que los que no lo están. Esta misma dispersión se observa entre los que valoran positiva y negativamente la recepción del automóvil por el taller. Pedimos la estimación del MCA. Solo estamos limitando el número de dimensiones al número de variables (`ncp=5`) y que en la presentación de los resultados nos limite la información a dos dimensiones (`ncp=2`) con tres decimales (`nb.dec=3`).

```
fit2<-MCA(datos, ncp=5, graph=FALSE)
summary(fit2, nb.dec=3, ncp=2)
```

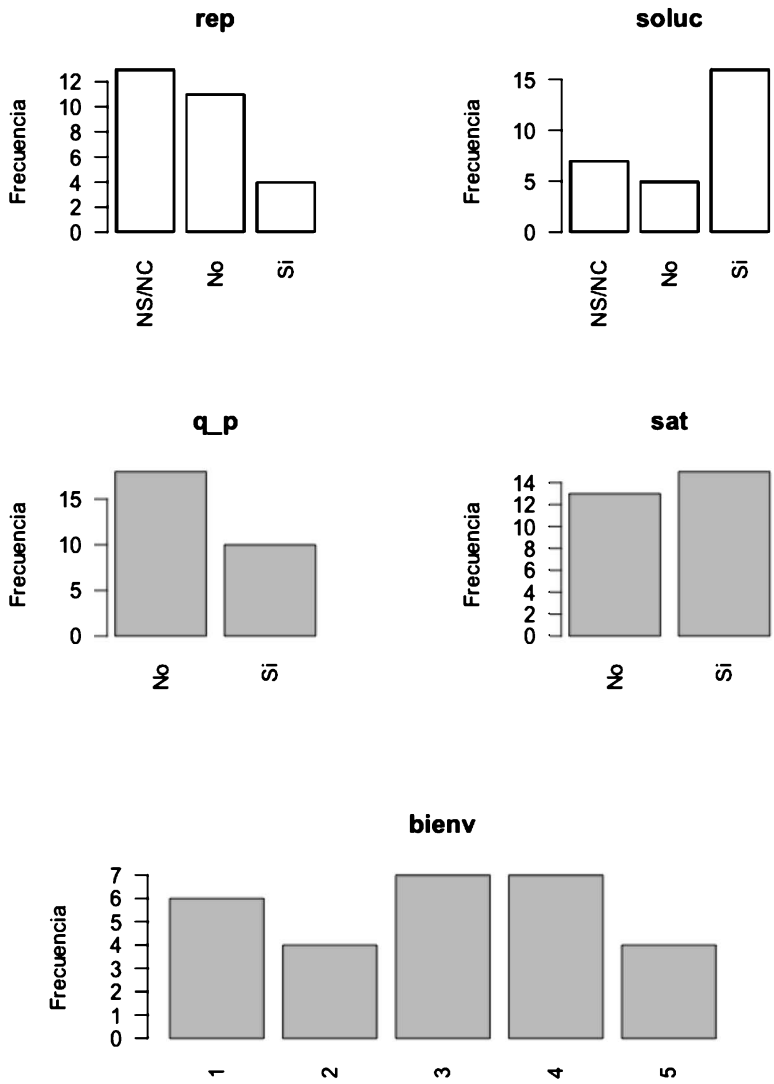
Aunque el cuadro 5.15 muestra que dos dimensiones explican algo menos de la mitad de la varianza, también vemos en el gráfico de sedimentación (*scree plot*) que, a partir de la segunda dimensión, la aportación de cada una de las siguientes es marginal respecto a las dos primeras.

El paso siguiente es obtener el mapa perceptual que relaciona los distintos niveles de las variables para intentar intuir la relación entre ellas y, posteriormente, intentar dar una interpretación a los ejes dependiendo de la contribución de cada nivel de las variables a dichos ejes. La figura 5.9 ofrece el mapa. En lugar de optar por la opción por defecto de `MCA{FactoMineR}` recurrimos a la función `fviz_mca_var{factoextra}`, que nos da más flexibilidad y gráficos de mayor calidad visual.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 5.7.: Descripción de la base de datos



Estadísticos e-Books & Papers

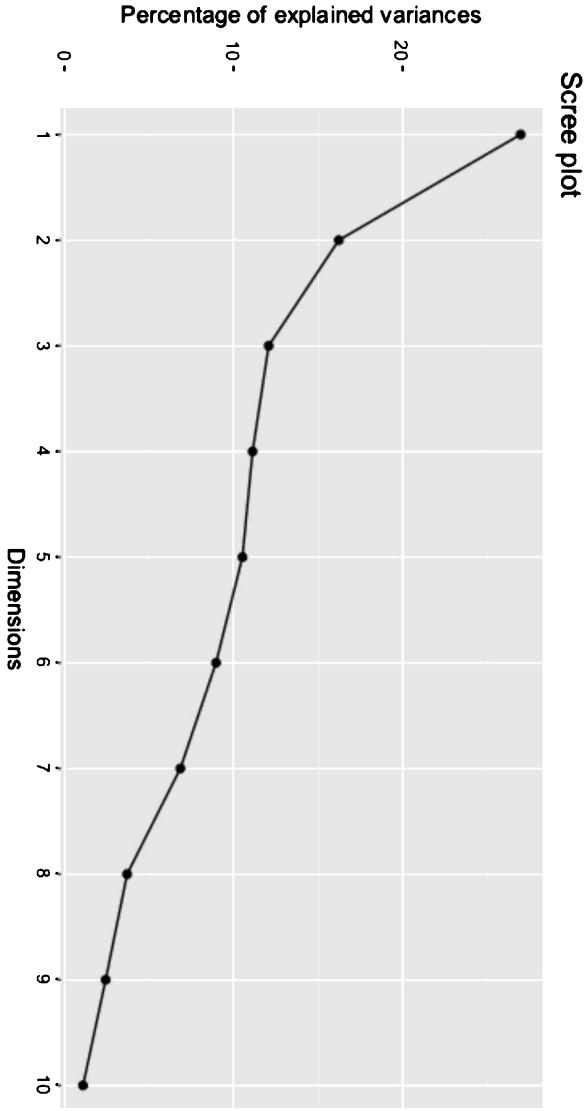
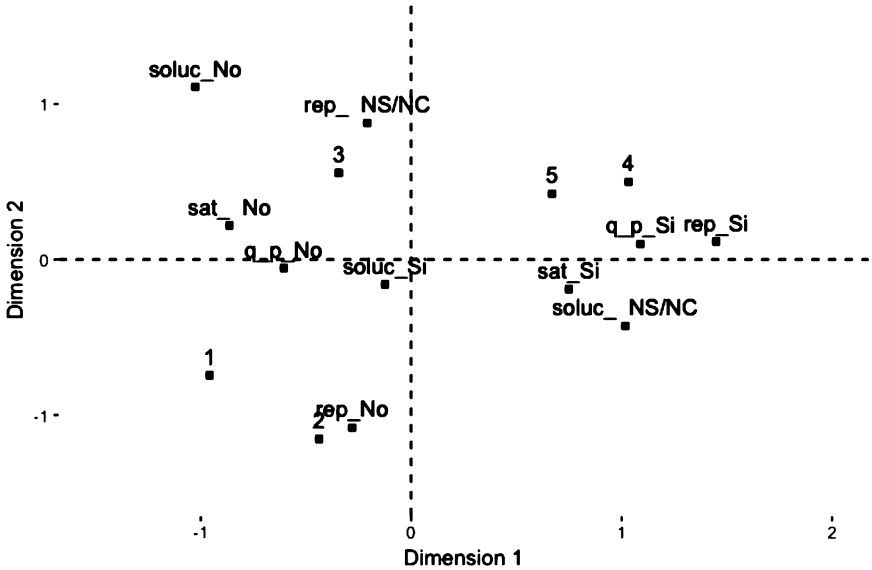


Figura 5.8.: Gráfico de sedimentación

© Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 5.9.: Mapa perceptual




```
fviz_mca_var(fit2, col.var="black", shape.var = 15+ggtitle("")
+expand_limits(x=c(-1.5,2), y=c(-1.5, 1.5))+
labs (x="Dimension 1", y="Dimension 2")+
guides(colour=FALSE, shape=FALSE)+
theme_gray()
```

La inspección del gráfico de la figura 5.9 ya nos permite intuir las primeras conclusiones. Si nos fijamos, el punto que representa a los individuos que tienen intención de repetir y ser leales (rep_Si) se encuentra rodeado de aquellos que declaran estar satisfechos (sat_Si) y perciben una buena relación calidad precio (q_p_Si) junto a los que mejor valoran (4 y 5) la recepción del automóvil. El que el problema del automóvil se haya resuelto ocupa un lugar a mitad de camino de los que piensan repetir y los que no. Esto parece indicar que la resolución del problema se da por descontado en un servicio técnico y lo que realmente cuestiona la intención de repetir es la mala atención en la recepción —vemos a rep_No junto a los niveles 1 y 2 de esa variable— siendo esa la primera consecuencia de gestión que se debe considerar.

Si intentamos confirmar el análisis mediante una interpretación de la contribución de las variables a los ejes, la salida a analizar es muy similar al CA. Así el cuadro 5.16, además de las coordenadas para la representación, refleja la contribución de cada categoría a cada una de las dimensiones. Observamos que la dimensión 1 está explicada fundamentalmente por la satisfacción en sus

Estadísticos e-Books & Papers

Cuadro 5.16.: Calidad de la representación de los niveles de las variables

```
Categories
                  Dim.1    ctr   cos2 v.test    Dim.2    ctr   cos2 v.test
sat_  No        | -0.864 12.839  0.646 -4.177 |  0.219  1.378  0.042  1.061 |
sat_Si          |  0.748 11.127  0.646  4.177 | -0.190  1.194  0.042 -1.061 |
soluc_  NS/NC   |  1.017  9.583  0.345  3.050 | -0.428  2.818  0.061 -1.283 |
soluc_No        | -1.025  6.963  0.229 -2.484 |  1.111 13.586  0.268  2.692 |
soluc_Si        | -0.124  0.328  0.021 -0.746 | -0.160  0.903  0.034 -0.961 |
1               | -0.958  7.286  0.250 -2.598 | -0.745  7.335  0.151 -2.022 |
2               | -0.438  1.017  0.032 -0.930 | -1.153 11.715  0.222 -2.447 |
3               | -0.344  1.099  0.040 -1.033 |  0.558  4.792  0.104  1.673 |
4               |  1.033  9.897  0.356  3.100 |  0.499  3.836  0.083  1.497 |
5               |  0.669  2.370  0.075  1.419 |  0.422  1.570  0.030  0.896 |
q_p_No          | -0.606  8.741  0.660 -4.222 | -0.056  0.124  0.006 -0.390 |
q_p_Si          |  1.090 15.734  0.660  4.222 |  0.101  0.223  0.006  0.390 |
rep_  NS/NC     | -0.209  0.750  0.038 -1.010 |  0.879 22.115  0.670  4.252 |
rep_No          | -0.280  1.143  0.051 -1.171 | -1.081 28.293  0.756 -4.518 |
rep_Si          |  1.449 11.121  0.350  3.074 |  0.116  0.118  0.002  0.245 |
```

extremos, junto a la intención de repetir y la valoración positiva de la calidad precio. La segunda dimensión tiene una mayor contribución de la consideración de que no se ha resuelto el problema, se ha recibido una mala recepción del vehículo y no se tiene intención de repetir. Podemos interpretar el primero como el eje de la satisfacción/lealtad y sus causas y el segundo como el origen de los problemas de la insatisfacción.

Si se quiere una representación gráfica de la contribución —aunque con la claridad del cuadro no la vemos necesaria— basta indicarle al programa que nos la ofrezca en un gráfico de barras para cada dimensión (figura 5.10). La línea roja señala el valor de 1/número de categorías, que sería la contribución promedio si todas contribuyeran igual. Solo las que superan ese valor se considera que tienen una contribución razonable.

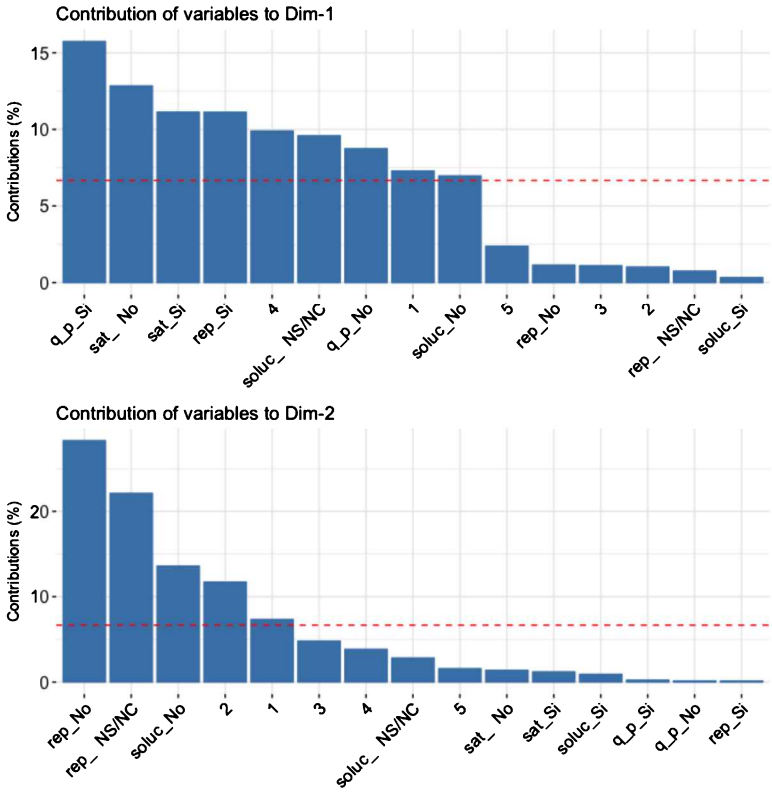
```
fviz_contrib(fit2, choice = "var", axes = 1)
fviz_contrib(fit2, choice = "var", axes = 2)
fviz_contrib(fit2, choice = "var", axes = 1:2)
```

Finalmente el MCA nos da la opción de representar sobre el mapa a los individuos. En nuestro caso son entrevistados anónimos pero no siempre será así y tendría interés analizar si los individuos son interpretables en sí mismos como países o empresas, si algunos de ellos se concentran alrededor de determinados valores de las variables. A veces, interpretadas las dimensiones, ponerlos solos en el mapa puede ser interesante. Por ejemplo, la figura 5.11 muestra los individuos sobre las mismas coordenadas donde representamos en el gráfico 5.10 las variables, pero con distintas intensidades de gris en función de su intención de repetir. Se observa claramente como los que pretenden hacerlo están muy alineados en la primera dimensión específicamente en su zona positiva. Lo importante es que los que no quieren volver están en la zona negativa de la segunda dimensión. Lógicamente el mayor interés está cuando superponemos las variables a los individuos (panel b de la figura 5.11) que aclara la asociación

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

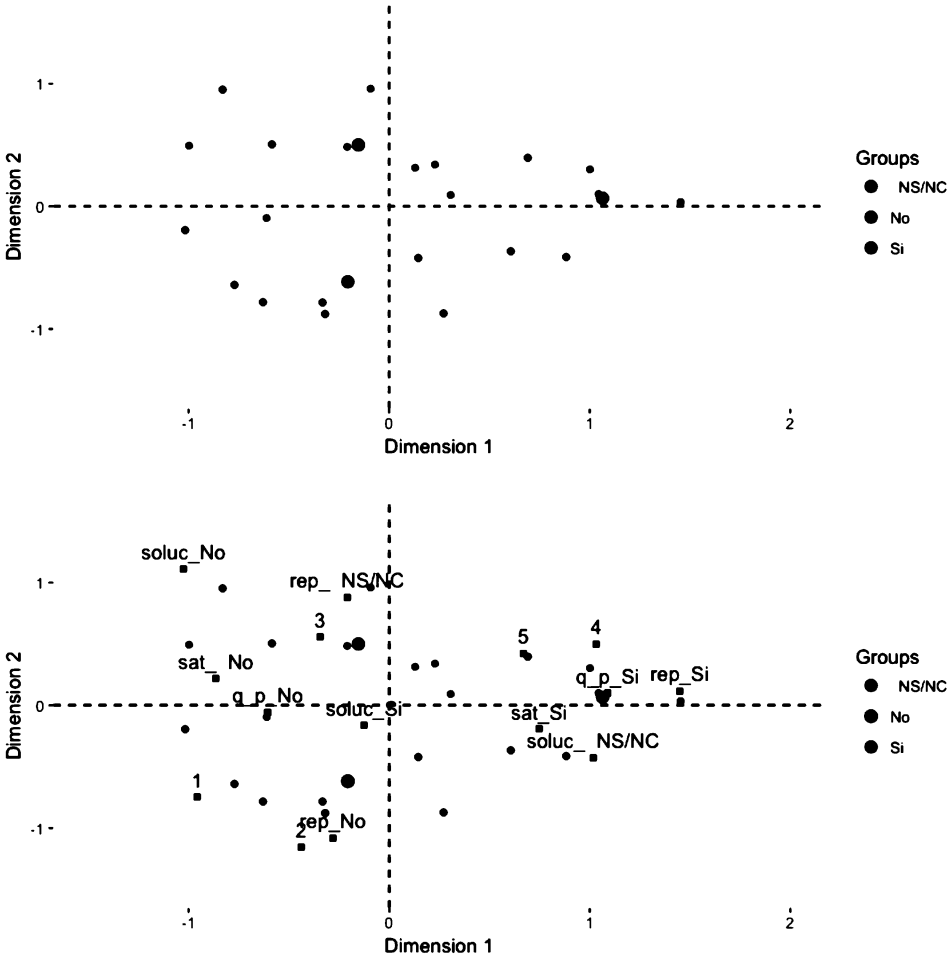
Figura 5.10.: Gráfico de contribuciones a las dimensiones



Estadísticos e-Books & Papers

entre ambas situaciones. Los puntos de no repetición (bajo el eje de abscisas) están asociados, además de lógicamente a la variable que lo señala (rep_No) a los niveles muy bajos de la atención en la recepción del vehículo (1 y 2).

Figura 5.11.: Mapa con los individuos



 © Ediciones Paraninfo

Estadísticos e-Books & Papers

6. Análisis de la varianza

6.1. Introducción

El análisis de la varianza es un método estadístico que nos permite analizar si la media de una variable —por tanto, métrica— toma valores estadísticamente distintos en los grupos que crea otra variable —por tanto, no métrica—. Originalmente el análisis de la varianza se utilizó para determinar si las cosechas que se obtenían con distintos tratamientos o niveles de fertilizantes diferían o no.

Se denomina *factor* a la variable que supuestamente ejerce una influencia sobre la variable estudiada, a la que denominaremos variable *dependiente*. En el caso anterior, el factor es el fertilizante, y la variable dependiente, la cosecha.

En el análisis de la varianza, el *factor* cuya influencia se quiere corroborar se introduce de forma discreta, independientemente de que sea de naturaleza continua o no. Así, la cantidad de fertilizante aplicada tiene una naturaleza intrínsecamente continua pero en un estudio de análisis de la varianza solamente se consideran un número determinado de niveles. Cuando el factor sea de naturaleza discreta, que será la situación más frecuente, utilizaremos de forma equivalente los términos de *grupo* o *nivel* para referirnos a una característica concreta.

El análisis de la varianza, especialmente por la difusión de programas de ordenador, es conocido por las siglas inglesas ANOVA (*ANalysis Of VArriance*). Normalmente hay dos formas de recoger los datos. Podemos exponer a distintas personas a cada uno de los grupos de tratamiento del factor —*diseño independiente* o *entre grupos*— o podemos coger a un único grupo de personas y someterlas a distintas manipulaciones en momentos distintos del tiempo —diseño de *medidas repetidas*—. Es un planteamiento similar al de las pruebas t para diferencia de medias independientes o relacionadas. La diferencia es que, en una prueba t, la variable independiente solo puede generar dos grupos en los que comparar la dependiente, mientras que, en un ANOVA, el número de grupos que genera la independiente puede ser superior a dos.

En relación con este punto, una cuestión habitual es ¿por qué es necesario realizar un análisis de la varianza cuando se cuenta con más de dos grupos y no se puede sustituir por el encadenamiento de varias pruebas t? Es decir, si el factor genera tres grupos, digamos A, B y C, podríamos comparar la media de la dependiente entre los grupos A-B, B-C, A-C con tres pruebas distintas. Field (2005) ilustra la razón, que no es otra que la inflación del error tipo I. Si en cada una de las tres pruebas t que realizamos exigiéramos un nivel de significatividad

Estadísticos e-Books & Papers

del 5 % (probabilidad de rechazar erróneamente la hipótesis nula de igualdad de medias, esto es, error tipo I), la probabilidad de no cometer ese error sería del 95 %. Si cada prueba es independiente, la probabilidad global en ellas de no cometer un error tipo I sería de $(0{,}95)^3 = 0{,}857$, por lo que la probabilidad de cometer, al menos, un error tipo I sería $1 - 0,857 = 0{,}143$, es decir, un 14,3 %. Y 3 es un número pequeño de grupos, en general la probabilidad de cometer un error tipo I en el encadenamiento, para una significatividad del 5 %, es de $1 - (0,95)^n$, siendo n el número de comparaciones que hay que realizar[1]. Para 5 grupos, por ejemplo, como se realizan 10 comparaciones, la probabilidad de cometer un error tipo I, es decir, detectar un efecto significativo del factor donde no lo hay, sería del 40 %.

6.2. Análisis de la varianza de un factor

Para desarrollar la lógica del análisis de la varianza, comenzaremos por el de un factor, es decir, cuando solo queremos analizar la influencia de una única variable independiente generadora de los grupos de tratamiento. Partiremos de un caso muy sencillo para ilustrar más fácilmente los desarrollos.

Caso 6.1. Métodos de adiestramiento de personal en Tricotosas Cerlerinas. En la empresa Tricotosas Cerlerinas se están considerando tres métodos alternativos para el adiestramiento de 14 operarios en una determinada técnica, tratándose de determinar si existen diferencias significativas entre los tres métodos. Para ello se forman tres grupos formados por 4, 5 y 5 operarios. La asignación de cada operario a un grupo se hace de forma aleatoria a fin de evitar que existan disparidades en la composición de los grupos. A cada uno de los grupos se le instruye en uno de los tres métodos, a los que se designa con las letras A, B y C. Al final del periodo de entrenamiento se somete a los 14 operarios adiestrados a una prueba común. Los resultados se han recogido en el cuadro 6.1.

En este caso existen tres poblaciones distintas, a cada una de las cuales se le ha aplicado un método diferente. La cuestión que se puede plantear la dirección de la empresa es la siguiente: ¿Son diferentes o, mejor dicho, significativamente diferentes los resultados obtenidos con los tres métodos?

El modelo de partida de un ANOVA es muy sencillo y viene dado por la siguiente expresión:

$$Y_g = \mu_g + \varepsilon_g = 1, 2, \cdots, G \tag{6.1}$$

que nos señala que la variable dependiente Y_g es igual a una media teórica más una variable aleatoria ε_g. Existen G subpoblaciones, es decir, G niveles del factor analizado. ¿Cuál es la **hipótesis nula** que ha de analizarse en un ANOVA? El punto de partida es que la media de la variable dependiente Y_g

[1] Para un factor con cinco grupos el número de comparaciones sería $C = \frac{k!}{2(k-2)!} = \frac{5!}{2\times 3!} = 10$.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 6.1.: Datos de las puntuaciones obtenidas por los empleados adiestrados con distintos métodos

Método	A	B	C
Resultados	6	9	9
	7	8	8
	8	9	7
	7	8	8
		9	9

es la misma en todos los subgrupos, es decir, que el factor no ejerce ningún efecto sobre ella. Hay que tener mucho cuidado con la **hipótesis alternativa**, porque nos generará una tarea adicional que abordaremos posteriormente, las pruebas *post hoc*. Si se rechaza la hipótesis nula, la alternativa no es que todas las medias son distintas, sino que al menos una es distinta a las demás. Las pruebas *post hoc* a las que aludimos son las que nos tendrán que decir qué medias son distintas de cuáles. Por lo tanto:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \cdots = \mu_G \\ H_1 &: \text{No todas } \mu_g \text{ son iguales} \end{aligned} \tag{6.2}$$

Para nuestro caso de ejemplo, la hipótesis nula es que los desempeños en el puesto de trabajo de los trabajadores que han sido sometidos a formación no varían en función de la metodología aplicada en la misma. Veamos cuáles son los pasos que vamos a dar para el contraste de esas hipótesis. El proceso es el estándar de cualquier prueba inferencial: construir un estadístico que recogerá la evidencia muestral a favor o en contra de la hipótesis nula y nos indicará la probabilidad de errar al no aceptarla.

6.2.1. Construcción del estadístico F

La figura 6.1 nos va a servir de base para ilustrar la lógica de construcción del estadístico F, que es el que se va a utilizar para el contraste de las hipótesis de la expresión (6.1). Es obviamente una ilustración idealizada pero creemos que puede servir para entender la descomposición de la varianza y el cálculo del estadístico F.

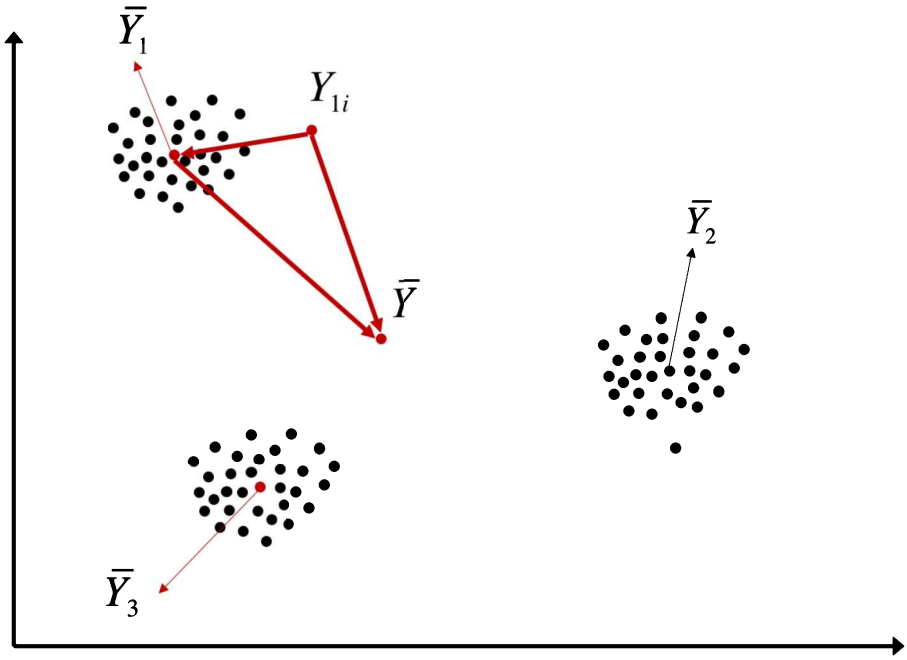
En esa figura vemos como la distancia de cualquier caso i que pertenece a uno de los g grupos que ha formado la variable de tratamiento o factor —denominémosle por ello Y_{gi}— al valor medio de la variable dependiente en el conjunto de la muestra $\bar{Y}$ o centroide de la muestra puede descomponerse en dos partes: lo que ese punto dista del centroide del grupo al que pertenece $\bar{Y}_g$ y lo que el grupo dista del promedio de la muestra $\bar{Y}$.

Dicho de otro modo:

$$\left(Y_{gi} - \bar{Y}\right) = \left(Y_{gi} - \bar{Y}_g\right) + \left(\bar{Y}_g - \bar{Y}\right) \tag{6.3}$$

Estadísticos e-Books & Papers

Figura 6.1.: Ilustración de un ANOVA



© Ediciones Paraninfo

Estadísticos e-Books & Papers

Por lo tanto, vemos que no todo lo que un caso difiere de la media de toda la muestra se debe a que pertenezca a un grupo dado, solo una de las partes tiene que ver con la pertenencia al grupo —lo que su grupo difiere de la media $\left(\bar{Y}_g - \bar{Y}\right)$— hay otra parte de esa diferencia que no se explica por pertenecer al grupo sino a su diferencia con el resto de casos de su propio grupo representados en su centroide: $\left(Y_{gi} - \bar{Y}_g\right)$. Por lo tanto, el factor, la variable de tratamiento, el pertenecer a un grupo explica una parte de la diferencia de los individuos de la muestra, pero no toda. $\left(\bar{Y}_g - \bar{Y}\right)$ es la *desviación explicada por el factor,* mientras que la no explicada es $\left(Y_{gi} - \bar{Y}_g\right)$, y la denominamos *desviación residual.*

Elevando al cuadrado ambos miembros de (6.3) se tiene que:

$$\left(Y_{gi} - \bar{Y}\right)^2 = \left(Y_{gi} - \bar{Y}_g\right)^2 + \left(\bar{Y}_g - \bar{Y}\right)^2 + 2\left(Y_{gi} - \bar{Y}_g\right)\left(\bar{Y}_g - \bar{Y}\right) \tag{6.4}$$

Si en (6.4) sumamos para todos los G grupos y para todos los individuos de cada grupo, se obtiene que:

$$\sum_{g=1}^{G}\sum_{n=1}^{n_g}\left(Y_{gi} - \bar{Y}\right)^2 = \sum_{g=1}^{G}\sum_{n=1}^{n_g}\left(Y_{gi} - \bar{Y}_g\right)^2 + \sum_{g=1}^{G}\sum_{n=1}^{n_g}\left(\bar{Y}_g - \bar{Y}\right)^2 =$$

$$= \sum_{g=1}^{G} n_g\left(\bar{Y}_g - \bar{Y}\right)^2 + \sum_{g=1}^{G}\sum_{n=1}^{n_g}\left(Y_{gi} - \bar{Y}_g\right)^2 \tag{6.5}$$

donde se ha tenido en cuenta para derivar la expresión que:

$$\sum_{g=1}^{G}\sum_{n=1}^{n_g}\left(Y_{gi} - \bar{Y}_g\right)\left(\bar{Y}_g - \bar{Y}\right) = \sum_{g=1}^{G}\left(\bar{Y}_g - \bar{Y}\right)\sum_{n=1}^{n_g}\left(Y_{gi} - \bar{Y}_g\right) = 0$$

ya que la suma de las desviaciones respecto a la media de cada grupo es por construcción igual a 0. En el primer miembro de (6.5) aparece la suma de cuadrados de las desviaciones de cada observación respecto a la media global. A esta suma la denominaremos **suma de cuadrados total (SCT)** y refleja la variabilidad total. Si se divide por el tamaño total de muestra, se obtiene la varianza total. La SCT, de acuerdo con (6.5), se descompone en dos partes:

1. La suma de cuadrados de las desviaciones entre la media de cada grupo y la media general. Esta es la suma de cuadrados explicada por el factor considerado, a la que denominaremos **suma de cuadrados del factor (SCF)** o variabilidad explicada.

2. La suma de cuadrados de las desviaciones entre cada dato y la media de su grupo. Esta es la suma de cuadrados no explicada, a la que denominaremos **suma de cuadrados residual (SCR)** o variabilidad residual. Así pues la descomposición de la ecuación (6.5) se puede presentar de la siguiente forma:

Estadísticos e-Books & Papers

$$SCT = SCF + SCR \tag{6.6}$$

A su vez, la SCR se puede expresar como agregación de la suma de cuadrados residuales correspondientes a cada uno de los grupos, es decir:

$$SCR = SCR_1 + SCR_2 + \cdots + SCR_G \tag{6.7}$$

Veamos la descomposición de la suma de cuadrados aplicados a los datos del caso 6.1. El cuadro 6.2 nos ofrece los cálculos necesarios.

$$SCT = \sum_{g=1}^{G} \sum_{n=1}^{n_g} \left(Y_{gi} - \bar{Y}\right)^2 = 12$$

$$SCF = \sum_{g=1}^{G} \sum_{n=1}^{n_g} \left(\bar{Y}_g - \bar{Y}\right)^2 = 6$$

$$SCR = \sum_{g=1}^{G} \sum_{n=1}^{n_g} \left(Y_{gi} - \bar{Y}_g\right)^2 = 6$$

El concepto de **grados de libertad**, como señala Field (2005), no es sencillo de explicar, pero es necesario para continuar con la explicación de la lógica del estadístico F que utilizaremos para el contraste de la hipótesis nula. Este autor realiza la siguiente analogía. En estadística el número de grados de libertad es el número de observaciones que pueden variar. Imaginemos que tenemos una muestra de cuatro observaciones extraídas de una población. Estas observaciones pueden tomar cualquier valor. Sin embargo si usamos esta muestra para calcular la media como una estimación de la media poblacional y que nos da el valor 10, entonces asumimos que la media poblacional es 10 y este valor queda constante. Con este parámetro fijado, ¿pueden seguir variando los cuatro valores? No, como hay que mantener la media constante, solo pueden hacerlo 3 de ellos. Por ejemplo, si la muestra era 8, 9, 11, 12 (media = 10) y cambiamos tres de estos valores a 7, 15 y 9, entonces el tercer valor ha de ser 9 para mantener la media constante. Si se mantiene un parámetro constante, el número de grados de libertad ha de ser el tamaño muestral menos uno. Esto explica, por ejemplo, por qué para calcular la desviación típica de una población, dividimos la suma de los cuadrados por $n-1$ y no por n. Para obtener las sumas de los cuadrados del ANOVA, hemos usado la muestra completa, por lo tanto, el número de grados de libertad será $n-1$.

De una manera más precisa, y para los datos del caso 6.1.

1. Los grados de libertad de SCT son iguales al número total de datos menos 1, es decir, $n-1$. La restricción viene determinada porque hemos ya estimado un parámetro, la media global $\bar{Y}$.

2. Los grados de libertad de SCF son iguales al número total de grupos menos 1, es decir, $G-1$. La restricción viene también determinada por

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 6.2.: Cálculos para la descomposición de la suma de cuadrados en el caso 6.1

Método	Y_{gi}	$\bar{Y}_g$	$Y_{gi}-\bar{Y}$	$(Y_{gi}-\bar{Y})^2$	$Y_{gi}-\bar{Y}_g$	$(Y_{gi}-\bar{Y}_g)^2$	$\bar{Y}_g-\bar{Y}$	$(\bar{Y}_g-\bar{Y})^2$
A	6	7,0	-2	4	-1,0	1,00	-1,0	1,00
A	7		-1	1	0,0	0,00	-1,0	1,00
A	8		0	0	1,0	1,00	-1,0	1,00
A	7		-1	1	0,0	0,00	-1,0	1,00
B	9	8,6	1	1	0,4	0,16	0,6	0,36
B	8		0	0	-0,6	0,36	0,6	0,36
B	9		1	1	0,4	0,16	0,6	0,36
B	8		0	0	-0,6	0,36	0,6	0,36
B	9		1	1	0,4	0,16	0,6	0,36
C	9	8,2	1	1	0,8	0,64	0,2	0,04
C	8		0	0	-0,2	0,04	0,2	0,04
C	7		-1	1	-1,2	1,44	0,2	0,04
C	8		0	0	-0,2	0,04	0,2	0,04
C	9		1	1	0,8	0,64	0,2	0,04
Total		**8,0**	**0**	**12**	**0,0**	**6,00**	**0,0**	**6,00**

© Ediciones Paraninfo

Estadísticos e-Books & Papers

el cálculo de $\bar{Y}$. Las g medias de los grupos están condicionadas a la restricción de la media general:

$$n_1\bar{Y}_1 + n_2\bar{Y}_2 + \cdots + n_G\bar{Y}_G = n\bar{Y}$$

3. Los grados de libertad de la SCR son iguales al número total de datos menos G, es decir, $n - G$, ya que la media muestral de cada grupo actúa como una restricción al cálculo de las desviaciones de dicho grupo. Puede comprobarse como el número de grados de libertad gl de la SCT es igual a la suma de los grados de libertad de cada uno de sus componentes, es decir:

$$gl\,(SCT) = gl\,(SCF) + gl\,(SCR) = (G-1) + (n-G) = n-1$$

La SCF nos dice cuánta variación explica el factor y la SCR cuánta se debe a factores externos al manipulado. Sin embargo, como ambos elementos son suma de casos, estarán afectados por el número de observaciones que incluyen, por ejemplo para la SCF hemos sumado solo cuatro medias (véase cuadro 6.2), mientras que para la SCR y la SCT hemos sumado 14 valores. Para eliminar este sesgo calculamos el promedio de la suma de los cuadrados dividiendo por el número de grados de libertad. A este concepto lo denominamos **media cuadrática**. Dividimos por los grados de libertad en lugar de por el número de casos utilizados debido a que pretendemos extrapolar un resultado a la población y alguno de esos datos se ha de mantener constante, tal y como hemos planteado al explicar el concepto de grados de libertad.

Aplicado a nuestro caso:

$$MCF = \frac{SCF}{G-1} = \frac{6}{3-1} = 3 \tag{6.8}$$

$$MCR = \frac{SCR}{n-G} = \frac{6}{14-3} = 0{,}545 \tag{6.9}$$

Nos resta, finalmente, por proponer un estadístico que recoja la siguiente lógica. Cuanto más explique la pertenencia al grupo (factor) la variación de la variable dependiente, más grande debería hacerse ese estadístico. Como, además, si crece la varianza explicada por el factor, decaerá la residual, si la MCR está en el denominador de ese estadístico, todavía este se hará mayor. En resumen, el **estadístico** F es la ratio entre MCF y MCR y cumple la lógica señalada:

$$F = \frac{MCF}{MCR} = \frac{3}{0{,}545} = 5{,}5 \tag{6.10}$$

estadístico que se distribuye como una F de Snedecor con $G-1$ grados de libertad en el numerador y $n-G$ grados en el denominador.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 6.3.: Resultados de la estimación del ANOVA del caso 6.1

```
            Df Sum Sq Mean Sq F value Pr(>F)
metodo       2      6  3.0000     5.5 0.0221 *
Residuals   11      6  0.5455
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La sintaxis para la obtención del ANOVA es sencilla. Usaremos la función `aov{stats}`, basta simplemente definir cuál es la variable dependiente (`resultado`) y cuál el factor (`metodo`). Los resultados se muestran en el cuadro 6.3 y en él se puede observar la coincidencia con los cálculos manuales de todos los parámetros del modelo, SCM, SCR, MCM, MCR, sus respectivos grados de libertad y el estadístico F.

```
datos<-Datos_6_1_Caso
fit<-aov(data=datos, resultado~metodo)
summary(fit)
```

6.2.2. Supuestos del ANOVA

Los supuestos del ANOVA son los mismos que los de todas las pruebas paramétricas basadas en la distribución normal: los datos deben proceder de una distribución normal (**normalidad**), las varianzas de la variable dependiente en cada nivel del factor deben ser aproximadamente iguales (**homocedasticidad**), las observaciones deben ser independientes en los grupos que crea el factor y la variable dependiente debe ser **métrica**.

Field (2005) señala que no todas las violaciones de estas hipótesis tienen la misma influencia sobre los resultados, siendo el ANOVA bastante robusto frente a las mismas. Por ejemplo Lunney (1970) analiza el uso del ANOVA con variables dependientes dicotómicas (0,1), y por tanto no métricas, obteniendo que, cuando los tamaños de los grupos eran iguales, la prueba era precisa si había al menos 20 grados de libertad y la categoría de la dependiente más pequeña tenía al menos el 20 % de las respuestas. Si esto último no era así hacían falta más de 40 grados de libertad.

En cuanto a la violación de la **homocedasticidad**, el ANOVA es bastante robusto cuando los tamaños muestrales en cada grupo son iguales. Glass *et al.* (1972a) obtuvieron que, cuando los grupos con mayores tamaños tenían mayores varianzas, el test F tendía a mantener la hipótesis nula cuando sí había diferencias significativas en las medias. Lo contrario ocurría cuando los grupos con mayor tamaño tenían varianzas menores. El test de Levene, que ya presentamos en el capítulo 2, permite comprobar este supuesto. En el caso en que la hipótesis de homoscedasticidad se viole, el investigador dispone dos

Estadísticos e-Books & Papers

versiones alternativas del estadístico F, el test de Brown y Forsythe (1974) y el test de Welch (1951). La lógica de la corrección del test de Brown y Forsythe (1974) es bastante directa. La razón de que, como acabamos de señalar, cuando los tamaños de los grupos están desequilibrados y el grupo mayor tiene más varianza, el test F tenga tendencia a ser conservador radica en que, si observamos su expresión, en su denominador está el término SCR, que, recordemos, se calculaba:

$$SCR = \sum_{g=1}^{G} \sum_{n=1}^{n_g} \left(Y_{gi} - \bar{Y}_g\right)^2$$

y dado que la varianza de la variable dependiente en un grupo determinado se puede poner como:

$$s_g^2 = \frac{1}{n_g - 1} \sum_{n=1}^{n_g} \left(Y_{gi} - \bar{Y}_g\right)^2$$

entonces:

$$SCR = \sum_{g=1}^{G} (n_g - 1)\, s_g^2$$

Por lo tanto, si el grupo con mayor tamaño n_g es, además, el que tiene mayor varianza s_g^2 y ambos términos están multiplicados, el denominador del test tiene tendencia a crecer y el test F a ser menor y más conservador. Pues bien, Brown y Forsythe (1974) proponen la siguiente corrección del denominador para corregir ese efecto:

$$SCR = \sum_{g=1}^{G} \left(1 - \frac{n_g}{n}\right) s_g^2 \tag{6.11}$$

de tal forma que la varianza de cada grupo está multiplicada por un término $(1 - n_g/n)$ que es más pequeño cuanto más grande es el tamaño del grupo. El estadístico F resultante se evalúa también con una corrección del número de grados de libertad en el denominador.

La obtención a través de la función `bf.test{onewaytests}` es muy sencilla y su resultado se ofrece en el cuadro 6.4:

```
bf.test(resultado2,metodo2)
```

El test de Welch (Welch, 1951) es menos intuitivo que el de Brown-Forsythe y no plantearemos su desarrollo. Su cálculo a partir de la función `bf.test {onewaytests}` es también muy sencilla.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 6.4.: Test de Brown-Forsythe

```
      Brown-Forsythe Test

data:  y vs group
F = 5.3617, num df = 2.0000, denom df = 9.2412, p-value = 0.02846
```

Cuadro 6.5.: Test de Welch

```
      Welch's Heteroscedastic F Test

data:  y vs group
F = 5.1283, num df = 2.0000, denom df = 6.5715, p-value = 0.04552
```

```
welch.test(resultado,metodo)
```

A la hora de decantarse por una corrección u otra en el supuesto de violación de la homoscedasticidad, Field (2005) sigue la recomendación de Tomarken y Serlin (1986) que concluyen que ambas correcciones corrigen bien el error tipo I, pero que el test de Welch (1951) tiene más potencia (capacidad de detección de un efecto significativo cuando este existe).

El paso previo antes de aplicar cualquiera de las dos correcciones del test F, por tanto, sería el cálculo del test de Levene (cuadro 6.6), cuyo desarrollo vimos en el capítulo 2, para evaluar si tenemos problemas de heteroscedasticidad.

```
library(car)
leveneTest(datos$resultado,datos$metodo,center=mean)
```

El resultado (cuadro 6.6) muestra que no puede rechazarse la hipótesis nula de igualdad de varianzas, por lo que la aplicación del test de Brown-Forsythe y el test de Welch serían innecesarios.

La hipótesis cuya violación causa más distorsiones es la de la **independencia de las observaciones**. Scariano y Davenport (1987) demostraron que, cuando

Cuadro 6.6.: Test de Levene

```
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group  2   0.227 0.8006
      11
```


Estadísticos e-Books & Papers

las observaciones están correlacionadas, el error tipo I se incrementa significativamente. Como ya apuntábamos en el capítulo 2, sin embargo, la única manera de corregir esta violación es ser cuidadosos en el diseño de los experimentos. Sin embargo, si se quiere una aproximación al diagnóstico, cuando los datos se obtienen de forma secuencial puede ocurrir que este supuesto no se cumpla. Una forma de averiguar si se da esta circunstancia es a través de la representación gráfica de los residuos. Si los residuos siguen una cierta estela, en lugar de estar distribuidos de forma aleatoria en torno al eje de abscisas, será un indicio de falta de independencia.

La alternativa ante **problemas de normalidad** cuya detección ya abordamos en el capítulo 2 está en planteamientos no paramétricos aunque se considera que el ANOVA es bastante robusto ante estas violaciones. El **test de Kruskal-Wallis** (Kruskal y Wallis, 1952) es una alternativa ante estas situaciones. Su planteamiento es bastante similar a una prueba de Mann-Whitney para comparar medias en dos grupos y se basa en el cálculo de rangos. En primer lugar, como se muestra en el cuadro 6.7 para los datos de nuestro caso, se ordenan los valores de la variable dependiente de menor a mayor sin tener en cuenta el grupo en el que están y luego se separan por grupos conservando el rango calculado. Nótese en el cuadro 6.7 que, al haber rangos empatados, se asignan a los que ocupan la misma posición el promedio de los rangos que les corresponderían. Veremos como el hecho de que existan rangos empatados exigirá una corrección del estadístico de Kruskal-Wallis.

El estadístico de Kruskal-Wallis se define como:

$$K = \frac{12}{n\,(n+1)} \sum_{i=1}^{G} \frac{R_i^2}{n_i} - 3\,(n+1) \tag{6.12}$$

donde toda la notación es conocida salvo R_i, que es la suma de los rangos para cada grupo y que está disponible en el cuadro 6.7. Para el caso que estamos utilizando como ilustración:

$$K = \frac{12}{14\,(14+1)} \left(\frac{14^2}{4} + \frac{50^2}{5} + \frac{41^2}{5} \right) - 3\,(14+1) = 5{,}5828$$

Pero, como hemos señalado, el test de Kruskal-Wallis necesita de una corrección cuando existen rangos empatados como es el caso. Esta corrección es la siguiente:

$$k = 1 - \frac{\sum_{i=1}^{G} \left(t_i^3 - t_1\right)}{n^3 - n} \tag{6.13}$$

donde, una vez más toda la notación es conocida salvo t_i, que es el número de rangos repetidos dentro de cada grupo (por ejemplo, en el grupo A se repiten dos rangos, en el grupo B son 5 y también 5 en el C tal y como se comprueba en la penúltima columna del cuadro 6.7. Pues bien, el estadístico de Kruskal-Wallis corregido es:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 6.7.: Cálculo del test de Kruskal-Wallis

Datos iniciales		Datos iniciales ordenados		Cálculo del rango medio para casos empatados		Suma de rangos por grupo		
Método	Y_{gi}	Método	Y_{gi}	Orden	R_i	Método	Y_{gi}	$\sum R_i$
A	6	A	6	1	1	A	1	14
A	7	A	7	2	3	A	3	
A	8	A	7	3	3	A	3	
A	7	C	7	4	3	A	7	
B	9	A	8	5	7	B	7	50
B	8	B	8	6	7	B	7	
B	9	B	8	7	7	B	12	
B	8	C	8	8	7	B	12	
B	9	C	8	9	7	B	12	
C	9	B	9	10	12	C	3	41
C	8	B	9	11	12	C	7	
C	7	B	9	12	12	C	7	
C	8	C	9	13	12	C	12	
C	9	C	9	14	12	C	12	

© Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 6.8.: Resultados del test de Kruskal-Wallis

```
        Kruskal-Wallis Test

data:  y vs group
X-squared = 6.1805, df = 2, p-value = 0.04549
```

$$K' = \frac{K}{k} = \frac{5{,}5828}{\frac{(2^3-2)+(5^3-5)+(5^3-5)}{14^3-14}} = \frac{5{,}5828}{0{,}9032} = 6{,}1805$$

El estadístico de Kruskal-Wallis se distribuye como una χ^2 con $G-1$ grados de libertad, lo que en nuestro caso es 2, permitiendo de esta forma evaluar su significatividad.

Si efectuamos el cálculo mediante la función `kw.test{onewaytests}`, vemos que este coincide con los resultados obtenidos más arriba (cuadro 6.8).

```
kw.test(resultado,metodo)
```

6.2.3. Medida de bondad del ajuste y tamaño del efecto

Para determinar si es importante la parte de la variabilidad total explicada por el factor se utiliza el coeficiente de determinación. El **coeficiente de determinación** viene dado por la siguiente expresión:

$$R^2 = \frac{SCF}{SCT} \tag{6.14}$$

A este coeficiente se le denomina también *eta cuadrado* (η^2). Un valor próximo a 1 indica que la mayor parte de la variabilidad total puede atribuirse al factor, mientras que un valor próximo a 0 significa que el factor explica muy poco de esa variabilidad total. A partir de los datos del cuadro 6.3 es inmediato deducir que en nuestro caso toma el valor $6/12 = 0{,}50$. El coeficiente de determinación es una medida del denominado **tamaño del efecto**.

La creciente inclusión en la presentación de los resultados de un test del tamaño del efecto —que ahora definiremos— tiene que ver con la sacralización del nivel de significación $p < 0{,}05$ como criterio de rechazo de hipótesis nulas —es muy recomendable el artículo de Cohen (1994) al respecto— que no deja de ser un nivel arbitrario fruto de la costumbre de ofrecer en los textos históricos de estadística los valores críticos a un nivel de 0,05 y 0,01. La idea básica del contraste de hipótesis es, como se ha visto, plantear una hipótesis nula, ajustar un modelo estadístico a los datos y el cálculo de un estadístico que nos permite establecer, cuando la probabilidad de obtener ese valor del estadístico de manera casual sea de menos del 5 %, que hay un *efecto significativo* del factor. El término *significativo* no debe confundirnos, significativo no significa

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

que el efecto del factor sea importante. El tamaño del efecto es la medida de la fuerza de la relación entre las variables más allá de su significatividad. Este tamaño del efecto es una medida estandarizada de la magnitud de la relación entre dos variables, y estandarizada quiere decir que podemos comparar los efectos a lo largo de distintos estudios aunque hayan utilizado distintas escalas de medida. Como señala Field (2005), este hecho ha llevado a la American Psychological Association a recomendar que todos los informes de resultados incorporen los tamaños de los efectos.

Veremos a lo largo del capítulo y del libro que existen distintas medidas del tamaño del efecto: coeficiente de correlación de Pearson, d de Cohen, *odds ratio*... En cada momento iremos viendo los más adecuados. Cohen (1988; 1992) ha hecho algunas sugerencias muy extendidas acerca de lo que se considera un tamaño del efecto pequeño, medio o grande cuando se utiliza el coeficiente de correlación de Pearson (r):

- $r = 0{,}10$ (efecto pequeño), se explica apenas un 1 % de la varianza.
- $r = 0{,}30$ (efecto medio), se explica un 9 % de la varianza.
- $r = 0{,}50$ (efecto grande), se explica un 25 % de la varianza.

Aunque estos niveles de referencia son útiles, Bahuley (2004) y Lenth (2001), como apunta Field (2005), señalan que no pueden sustituir a la necesaria contextualización dentro del dominio de investigación en que tiene lugar la prueba, sus resultados previos y sus costumbres.

En nuestro ejemplo, un coeficiente de determinación de 0,50 equivaldría a un r de Pearson de $\sqrt{0{,}50} = 0{,}70$ y, por lo tanto, a un efecto grande. Field (2005) apunta que esta medida del tamaño del efecto está ligeramente sesgada al estar basada en suma de cuadrados obtenidos de la muestra y no realizar ajuste alguno para el hecho de que estemos estimando el tamaño del efecto en la población por lo que propone para el ANOVA una medida más compleja, omega al cuadrado (ω^2), incorporando los grados de libertad y las medias de los cuadrados:

$$\omega^2 = \frac{SCF - (gl_F)\,MCR}{SCT + MCR} = \frac{6 - 2 \times 0{,}5455}{12 + 0{,}5455} = 0{,}3913 \longrightarrow \omega = 0{,}6255$$

que lleva a una estimación algo menor que la obtenida mediante el r (es básicamente una estimación insesgada de r). Normalmente se reporta ω^2 y Kirk (1996) propone 0,01, 0,06 y 0,14 como tamaños de efecto pequeño, medio y alto.

6.2.4. Pruebas *post hoc*: comparaciones múltiples

En el caso de que en un análisis de la varianza se acepte la hipótesis nula de que los distintos grupos tienen la misma media puede darse por concluido el

Estadísticos e-Books & Papers

estudio del caso.

Por el contrario, si se rechaza la hipótesis nula de igualdad es necesario investigar qué grupos han sido determinantes en ese rechazo. A este tipo de investigación complementaria y posterior a los contrastes básicos del análisis de la varianza lo denominaremos pruebas *post hoc*.

Las pruebas *post hoc* implican la comparación de las medias entre cada par de grupos que genera el factor. En nuestro caso A-B, A-C, B-C. Sabemos que al menos una de esas medias es diferente, pero no sabemos si las tres comparaciones generan medias distintas, una de ellas, o dos. Pero ya hemos planteado de manera detallada los problemas de inflación del error tipo I que encadenar pruebas pareadas conlleva cuando explicábamos porque encadenar pruebas t no es una opción frente a la realización de un análisis de la varianza.

Esto será así salvo que, precisamente, las pruebas *post hoc* estén diseñadas de tal forma que controlen la inflación del error tipo I. Y así ocurre. Básicamente lo hacen corrigiendo el nivel de significatividad α para que el error tipo I se mantenga constante al encadenar las pruebas.

La forma más sencilla de plantearlo es conocida como corrección de Bonferroni y consiste en dividir α por el número de comparaciones k que impliquen las pruebas que se quieren realizar. Así la diferencia de medias entre los grupos A-B no se evaluaría a un nivel del 5 %, sino que solo sería una diferencia significativa si la significatividad asociada al estadístico t fuera inferior al valor crítico p_{crit}:

$$p_{crit} = \frac{\alpha}{k} = \frac{0{,}05}{3} = 0{,}016$$

Obviamente, como señala Field (2005), hay un efecto de sustitución entre los errores tipo I y tipo II. Al controlar para no rechazar una hipótesis nula que es cierta, incrementamos la probabilidad de no dar como significativo un efecto que sí que lo es, es decir, perdemos potencia de prueba $1 - \beta$ o probabilidad de rechazar la hipótesis nula cuando hay que hacerlo y constatar la significatividad del efecto del factor. Por lo tanto, para evaluar qué test elegir, hay que considerar su capacidad para controlar los errores tipos I y II y su solidez ante violaciones de las hipótesis que subyacen en el análisis de la varianza. Si seguimos las recomendaciones de Toothaker (1993), pese a la gran diversidad, pueden extraerse algunas directrices útiles:

- Cuando hay homocedasticidad:
  - Cuando los tamaños de los grupos son iguales: Tukey o Bonferroni.
  - Si los tamaños de los grupos son ligeramente diferentes: Gabriel.
  - Si los tamaños de los grupos son muy distintos: Hochberg.
- Cuando no hay homoscedasticidad: Games-Howell.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Desgraciadamente no todos los test mencionados están disponibles en R en el momento de redacción de este libro. Teniendo en cuenta las recomendaciones de Toothaker (1993) y los paquetes disponibles, Field *et al.* (2012) plantean la siguiente recomendación:

- Cuando hay homoscedasticidad: Tukey o Bonferroni.
- Cuando se tienen dudas sobre el cumplimiento de alguna de las hipótesis del análisis de la varianza, métodos robustos basados en el *bootstrap*, medias recortadas o estimadores-M como los propuestos por Wilcox (2003; 2005).

Veamos la aplicación al caso 6.1. En primer lugar debemos hacer notar que, como se observa en el cuadro 6.3, el rechazo de la hipótesis nula se da a un nivel del 5 % y no del 1 %, lo que ya nos hace prever que el número de comparaciones donde las medias sean distintas pueden ser pocas. Hay varias formas de implementar los test *post hoc*. La función `pairwise.t.test {stats}` incorpora la corrección de Bonferroni (`p.adj = bonferroni`), la de Holm (1979), que compara el valor p con un nivel de error tipo I que se va reduciendo en comparaciones sucesivas (`p.adj = holm`), Hommel (1988) (`p.adj = hommel`), Hochberg (1988) (`p.adj = hochberg`), Benjamini y Hochberg (1995) (`p.adj=BH`) y Benjamini y Yekutieli (2001) (`p.adj = BY`). A modo de ejemplo, las correcciones de Bonferroni y de Holm, llevarían al resultado del cuadro 6.9. En ambos casos parece confirmarse que solo las medias de los métodos A y B son distintas ($p < 0{,}05$).

```
pairwise.t.test(resultado2,metodo2,p.adjust="bonferroni")
pairwise.t.test(resultado2,metodo2,p.adjust="holm")
```

El test de Tukey y el de Dunnet se pueden obtener mediante la función `TukeyHSD{stats}`. Hay que fijarse en los intervalos de confianza de las diferencias de las medias del método que no deberían contener al cero. Vemos que el método 1(A)-2(B) no lo contiene, pero sí los demás, de manera coherente con los resultados anteriores.

```
posthoc.tukey<-TukeyHSD(fit)
print(posthoc.tukey)
plot(posthoc.tukey)
```

Finalmente las pruebas robustas basadas en medias recortadas `lincon{WRS2}` y *boostrapping*, `mcppb20{WRS2}` señalan que no hay ningún grupo con medias distintas en el primer caso y el 1 frente al 2 (A frente al B) en el segundo. Nótese que, para la primera prueba (medias recortadas), nos fijamos en el cuadro 6.11 que el intervalo de confianza de las diferencias no contenga al 0, y no en la significatividad dado que esta no está corregida para controlar el error tipo I.

Estadísticos e-Books & Papers

Cuadro 6.9.: Pruebas *post hoc* de Bonferroni y Holm

```
      Pairwise comparisons using t tests with pooled SD

data:  resultado2 and metodo2

  1     2
2 0.024 -
3 0.102 1.000

P value adjustment method: bonferroni

      Pairwise comparisons using t tests with pooled SD

data:  resultado2 and metodo2

  1     2
2 0.024 -
3 0.068 0.410

P value adjustment method: holm
```

Cuadro 6.10.: Test *post hoc* de Tukey

```
$metodo
                  diff        lwr       upr     p adj
       2-       1  1.6  0.2619046 2.9380954 0.0201810
       3-       1  1.2 -0.1380954 2.5380954 0.0799612
       3-       2 -0.4 -1.6615685 0.8615685 0.6773938
```

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 6.11.: Test *post hoc* robustos

```
lincon(formula = datos2)

          psihat ci.lower ci.upper p.value
1 vs. 2 -1.66667 -3.79415  0.46081 0.04403
1 vs. 3 -1.33333 -3.46081  0.79415 0.08368
2 vs. 3  0.33333 -2.03205  2.69872 0.62604

mcppb20(formula = datos2)

          psihat ci.lower ci.upper p-value
1 vs. 2 -1.66667 -2.75000 -0.50000 0.00000
1 vs. 3 -1.33333 -2.25000  0.00000 0.02003
2 vs. 3  0.33333 -0.66667  1.66667 0.44407
```

En la segunda prueba, la obtenida por *boostrapping*, podemos fijarnos en la significatividad porque sí está corregida. En este caso, además de diferencias entre los métodos A y B, también la habría al 5 % para A y C.

```
library(DTK)
resultado2<-c(6,7,8,7,9,8,9,8,9,9,8,7,8,9)
metodo2<-gl.unequal(n=3,k=c(4,5,5))
datos2<-data.frame(resultado2,metodo2)
lincon(datos2)
mcppb20(datos2)
```

Caso 3.2 Actitudes hacia el tabaco de fumadores y no fumadores

Con el fin de consolidar la presentación realizada del análisis de la varianza de un factor, vamos a desarrollar un caso con datos reales en lugar de la versión simulada de los datos del caso 3.1, que era voluntariamente sencillo para ejemplificar el cálculo de los distintos estadísticos.

Se ha realizado una encuesta para saber la actitud hacia el tabaco de los jóvenes. A una muestra de 241 de ellos se les han realizado distintas preguntas sobre dicha **actitud** pidiéndoseles que valoren en una escala de 1 "Totalmente en desacuerdo" hasta 5 "Totalmente de acuerdo" su opinión respecto a afirmaciones como "Fumar perjudica la salud" o "Deben aumentarse los impuestos sobre el tabaco". Para evaluar si hay diferencias entre las actitudes medidas por las mencionadas preguntas se les pidió que indicaran cuál era su **hábito** respecto al tabaco, con tres posibilidades:

1. Nunca he fumado.
2. He fumado, pero lo he dejado.
3. Soy fumador.

Estadísticos e-Books & Papers

Cuadro 6.12.: Descriptivos de las variables dependientes en los grupos

```
> # Para "fumar perjudica la salud"

            habito   N     mean        sd         se
1             Fuma 107 4.598131 0.8780318 0.08488254
2 Ha dejado fumar  14 4.714286 0.6112498 0.16336339
3          No fuma 120 4.641667 0.9855752 0.08997030


> # Para "Deben subirse los impuestos sobre el tabaco"

            habito   N     mean       sd        se
1             Fuma 107 1.728972 1.153911 0.1115528
2 Ha dejado fumar  14 2.857143 1.561909 0.4174378
3          No fuma 120 3.466667 1.201773 0.1097063
```

Queremos saber si, fundamentalmente, las actitudes hacia el tabaco difieren en función de que se sea o no fumador. También interesa saber si, en general, los que han dejado de fumar tienen una actitud más cercana a la de los fumadores, por haberlo sido, o a los no fumadores.

Aunque el análisis de la varianza de un factor analiza una variable dependiente cada vez, duplicaremos los análisis tanto para la pregunta sobre el acuerdo con que fumar perjudica la salud, como sobre que se suban los impuestos sobre el tabaco. La razón es didáctica: mostrar relaciones donde las diferencias son significativas y también donde no lo son.

El primer paso es, siempre, realizar un análisis descriptivo de las variables dependientes (las dos preguntas sobre la actitud) en los grupos que genera el factor "hábito". Usamos para ello la función `ddply{plyr}` y solicitamos que se calcule la frecuencia, media, desviación típica y error estándar de la media.

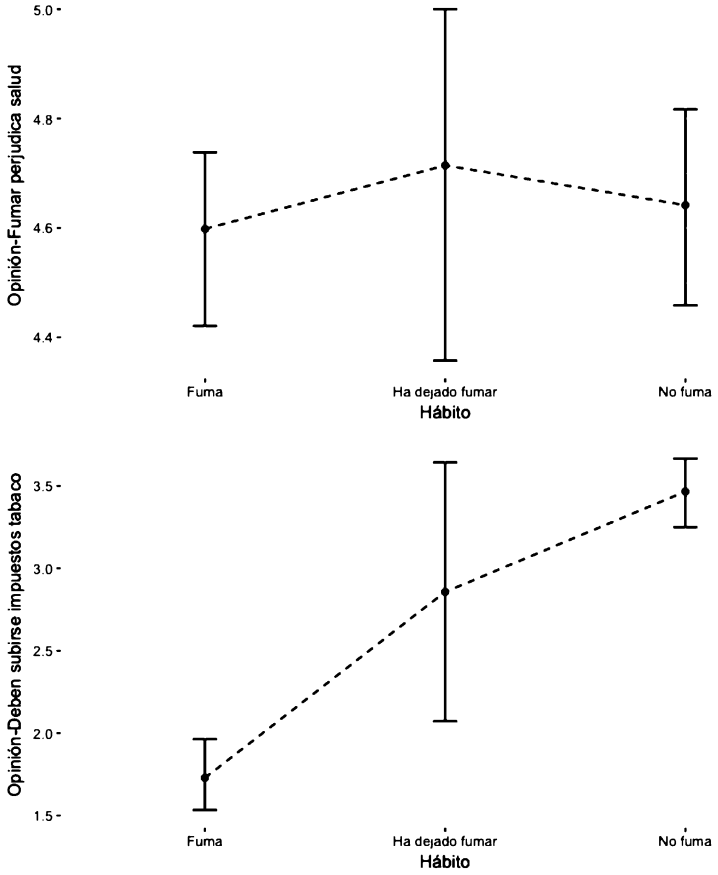
```
# Para "fumar perjudica la salud"
ddply(datos, c("habito"), summarise,
      N    = length(opinion_perjudica),
      mean = mean(opinion_perjudica),
      sd   = sd(opinion_perjudica),
      se   = sd / sqrt(N) )
# Para "Deben subirse los impuestos sobre el tabaco"
ddply(datos, c("habito"),summarise,
      N    = length(opinion_impuestos),
      mean = mean(opinion_impuestos),
      sd   = sd(opinion_impuestos),
      se   = sd / sqrt(N) )
```

Los resultados (cuadro 6.12)ya parecen apuntar que las medias respecto a la opinión que fumar perjudica la salud son altas (toda superiores a 4,5) y muy

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 6.2.: Valores medios e intervalos de confianza al 95 % para las medias



parecidas entre los tres grupos. Sin embargo, en la pregunta respecto a si deben subirse los impuestos sobre el tabaco parece intuirse que el nivel de acuerdo es mucho más alto entre los no fumadores (3,46) que entre los fumadores (1,72). Lógicamente el análisis de la varianza lo estamos realizando para establecer si estas diferencias son o no significativas. También observamos que el error estándar de la media siempre es más alto en el colectivo de entrevistados que dejaron de fumar. La figura 6.2 nos muestra, para las dos variables dependientes analizadas, los valores medios de las mismas en los niveles del factor y el intervalo de confianza al 95 % de esas medias. En la medida en que esos intervalos se solapen, las diferencias entre las medias no serían significativas.

El siguiente paso es comprobar el supuesto de homoscedasticidad para establecer si podemos utilizar el estadístico F o hemos de recurrir a las correcciones que vimos de Brown-Forsythe o de Welch como consecuencia de la ausencia de

Estadísticos e-Books & Papers

Cuadro 6.13.: Test de Levene

```
> # Para "fumar perjudica la salud"
Levene's Test for Homogeneity of Variance (center = mean)
       Df F value Pr(>F)
group   2  0.3335 0.7167
      238


> # Para "Deben subirse los impuestos sobre el tabaco"
Levene's Test for Homogeneity of Variance (center = mean)
       Df F value  Pr(>F)
group   2  2.4387 0.08945 .
      238
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

la misma.

```
# Para "fumar perjudica la salud"
leveneTest(datos$opinion_perjudica,datos$habito,center=mean
# Para "Deben subirse los impuestos sobre el tabaco"
leveneTest(datos$opinion_impuestos,datos$habito,center=mean)
```

Dado que en ambos casos (cuadro 6.13) no puede rechazarse la hipótesis nula de homoscedasticidad, estimaremos el análisis de la varianza sin la corrección del estadístico F.

```
# Para "fumar perjudica la salud"
fit1<-aov(data=datos, opinion_perjudica~habito)
summary(fit1) plot(fit1)
# Para "Deben subirse los impuestos sobre el tabaco"
fit2<-aov(data=datos, opinion_impuestos~habito)
summary(fit2) plot(fit2)
```

Puede comprobarse en el cuadro 6.14 como, para el caso de la opinión acerca de si el tabaco perjudica la salud, no puede rechazarse la hipótesis nula de igualdad de medias ($F(2,238) = 0{,}13; p > 0{,}05$), mientras sí que podemos afirmar que hay diferencias de opinión en cuanto a que se suban los impuestos sobre el tabaco ($F(2,238) = 59{,}16; p < 0{,}01$). Este resultado nos lleva a centrar las pruebas *post hoc*, solo en la segunda variable, en la medida en que al no haber medias distintas en la primera no tiene sentido preguntarse qué grupos en función del hábito difieren. Solicitamos para ello alguno de los test *post hoc* que vimos con anterioridad.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 6.14.: Contraste de hipótesis

```
> # Para "fumar perjudica la salud"

             Df Sum Sq Mean Sq F value Pr(>F)
habito        2   0.22  0.1108    0.13  0.878
Residuals   238 202.17  0.8494

> # Para "Deben subirse los impuestos sobre el tabaco"

             Df Sum Sq Mean Sq F value Pr(>F)
habito        2  171.4   85.69   59.16 <2e-16 ***
Residuals   238  344.7    1.45
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Basados en bonferroni
pairwise.t.test(datos$opinion_impuestos,datos$habito,
p.adjust="bonferroni")
pairwise.t.test(datos$opinion_impuestos,datos$habito,
p.adjust="holm")

# Tukey
posthoc.tukey<-TukeyHSD(fit2)
print(posthoc.tukey) plot(posthoc.tukey)

# bootstrapping y medias recortadas
metodo2<-gl.unequal(n=3,k=c(107,14,120))
datos2<-data.frame(datos$opinion_impuestos,datos$habito)
lincon(datos2) mcppb20(datos2)
```

Observamos que los resultados (cuadro 6.15) son muy coherentes y confirman lo que ya anticipábamos al analizar la figura 6.2, que respecto a la subida de impuestos sobre el tabaco los fumadores discrepan de los que no fuman o lo han dejado, al ser mucho más contrarios a ello, mientras que los fumadores y los que han dejado de fumar no difieren entre sí.

Nótese que al principio del caso no hemos efectuado el análisis de normalidad de las dos variables dependientes. Aunque ya hemos señalado que el análisis de varianza es bastante robusto a problemas de normalidad, no está de más que, si este problema existe, se realice un análisis de la varianza no paramétrico de Kruskal-Wallis para confirmar el resultado obtenido por el planteamiento paramétrico. Recordemos del capítulo 2 que la solicitud de los estadísticos de Kolmogorov-Smirnov y Shapiro para realizar esta comprobación de normalidad era sencilla:

Estadísticos e-Books & Papers

Cuadro 6.15.: Pruebas *post hoc*

```
      Pairwise comparisons using t tests with pooled SD

data:  datos$opinion_impuestos and datos$habito

                Fuma   Ha dejado fumar
Ha dejado fumar 0.0034 -
No fuma         <2e-16 0.2226

P value adjustment method: bonferroni


      Pairwise comparisons using t tests with pooled SD

data:  datos$opinion_impuestos and datos$habito

                Fuma   Ha dejado fumar
Ha dejado fumar 0.0022 -
No fuma         <2e-16 0.0742

P value adjustment method: holm

Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = opinion_impuestos ~ habito, data = datos)

$habito
                             diff        lwr      upr     p adj
Ha dejado fumar-Fuma    1.1281709  0.3214694 1.934872 0.0032067
No fuma-Fuma            1.7376947  1.3602907 2.115099 0.0000000
No fuma-Ha dejado fumar 0.6095238 -0.1921060 1.411154 0.1740187

lincon(formula = datos2)

                              psihat ci.lower ci.upper p.value
No fuma vs. Fuma            -1.49231 -3.21053  0.22591 0.03346
No fuma vs. Ha dejado fumar -2.27564 -2.62490 -1.92638 0.00000
Fuma vs. Ha dejado fumar    -0.78333 -2.50599  0.93932 0.22591

mcppb20(formula = datos2)

                              psihat ci.lower ci.upper p-value
No fuma vs. Fuma            -1.49231 -2.70000 -0.48462 0.00000
No fuma vs. Ha dejado fumar -2.27564 -2.61944 -1.87073 0.00000
Fuma vs. Ha dejado fumar    -0.78333 -1.90000  0.52500 0.17696
```

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 6.16.: Test de normalidad

```
        Shapiro-Wilk normality test

data:  datos$opinion_perjudica
W = 0.46191, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  datos$opinion_impuestos
W = 0.85752, p-value = 3.789e-14


        Lilliefors (Kolmogorov-Smirnov) normality test

data:  datos$opinion_perjudica
D = 0.4587, p-value < 2.2e-16


        Lilliefors (Kolmogorov-Smirnov) normality test

data:  datos$opinion_impuestos
D = 0.2031, p-value < 2.2e-16
```

```
# Para "fumar perjudica la salud"
shapiro.test(datos$opinion_perjudica)
lillie.test(datos$opinion_perjudica)

# Para "Deben subirse los impuestos sobre el tabaco"
shapiro.test(datos$opinion_impuestos)
lillie.test(datos$opinion_impuestos)
```

Dado que, como se aprecia en el cuadro 6.16 con ambos test y para las dos variables, siempre se puede rechazar la hipótesis nula de normalidad, deberemos efectuar el análisis de la varianza no paramétrico de Kruskal-Wallis:

```
# Para "fumar perjudica la salud"
kw.test(datos$opinion_perjudica,datos$habito)
# Para "Deben subirse los impuestos sobre el tabaco"
kw.test(datos$opinion_impuestos,datos$habito)
```

Observamos como los resultados del cuadro 6.17 confirman los resultaos del análisis de la varianza paramétrico al no detectar diferencias significativas para la variable "Fumar perjudica la salud" y sí hacerlo para la variable "Deben subirse los impuestos sobre el tabaco".

Estadísticos e-Books & Papers

Cuadro 6.17.: ANOVA no paramétrico de Kruskal-Wallis

```
# Para "fumar perjudica la salud"

      Kruskal-Wallis Test

data:  y vs group
X-squared = 1.9572, df = 2, p-value = 0.3758

# Para "Deben subirse los impuestos sobre el tabaco"

      Kruskal-Wallis Test

data:  y vs group
X-squared = 82.311, df = 2, p-value < 2.2e-16
```

6.3. Análisis de la varianza de dos factores

En el análisis de la varianza con 2 factores aparece una novedad en relación con el análisis de la varianza con un solo factor: la interacción entre factores. Conceptualmente, el proceso es análogo a cuando se utiliza un único factor como variable independiente, simplemente hay que centrar la atención, al plantear la descomposición de las varianzas, sobre la aparición de ese efecto de interacción y sobre todo sobre su interpretación. No reiteraremos los elementos comunes, como las hipótesis en las que se basa el análisis de la varianza y cómo han de comprobarse.

Caso 6.3. Número de horas de exposición a la televisión
Un investigador se plantea en qué medida el nivel educativo y el sexo afectan al número de horas que las personas ven la televisión y, si, de existir, ese efecto, la forma en que el nivel educativo afecta al visionado varía entre hombres y mujeres (interacción). Para el análisis cuenta con los datos de la General Social Survey estadounidense en la que, a los entrevistados, se les ha preguntado lo siguiente:

- [niveduc] Nivel educativo del entrevistado, siendo:
  - 0. Sin estudios.
  - 1. Primarios.
  - 2. Secundarios.
  - 3. Grado.
  - 4. Máster.
- [sexo] del entrevistado, siendo:
  - 1. Hombre.
  - 2. Mujer.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

- [horastv] Número de horas de exposición diaria a la televisión en la semana de la entrevista.

6.3.1. Construcción del estadístico F

El proceso de **descomposición de la varianza** es paralelo al del análisis de la varianza de un factor que ilustrábamos con la ecuación (6.3). La distancia de cualquier individuo al centroide de la muestra puede ponerse del siguiente modo, teniendo en cuenta que ahora hay dos factores generando los grupos, G para el factor 1 (5 niveles educativos en nuestro caso) y J para el factor 2 (dos niveles de la variable sexo):

$$\sum_{g=1}^{G}\sum_{j=1}^{J}\sum_{i=1}^{n_{gj}}\left(Y_{gj}-\bar{Y}\right)^2=\sum_{g=1}^{G}\sum_{j=1}^{J}n_{gj}\left(\bar{Y}_{gj}-\bar{Y}\right)^2+\sum_{g=1}^{G}\sum_{j=1}^{J}\sum_{i=1}^{n_{gj}}\left(Y_{gj}-\bar{Y_{gj}}\right)^2 \tag{6.15}$$

El primer miembro de (6.15) refleja la suma de cuadrados total con respecto a la media muestral global. Se utiliza un triple sumatorio: el primero se refiere a los distintos niveles del factor A, el segundo, a los distintos niveles del factor B, y el tercero, a los datos de cada celda. El segundo término del segundo miembro es la suma de cuadrados residual (SCR), mientras que el primero registra las diferencias al cuadrado entre la media de cada celda y la media global. Ahora bien, estas últimas diferencias, ¿a qué se deben? Pueden deberse bien a la influencia del factor A, bien a la influencia del factor B, o bien a la interacción entre ambos factores. Para aislar estas influencias, el primer término del segundo miembro de (6.15), al que denominaremos SCF (suma de cuadrados de los factores), se puede descomponer, realizando sencillas manipulaciones algebraicas, de la siguiente forma:

$$\begin{aligned}&\sum_{g=1}^{G}\sum_{j=1}^{J}n_{gj}\left(\bar{Y}_{gj}-\bar{Y}\right)^2=\\&\sum_{g=1}^{G}\sum_{j=1}^{J}n_{gj}\left(\bar{Y}_{g}-\bar{Y}\right)^2+\sum_{g=1}^{G}\sum_{j=1}^{J}n_{gj}\left(\bar{Y}_{j}-\bar{Y}\right)^2+\\&+\sum_{g=1}^{G}\sum_{j=1}^{J}n_{gj}\left(\bar{Y}_{gj}-\bar{Y}_{g}-\bar{Y}_{j}+\bar{Y}\right)^2\end{aligned} \tag{6.16}$$

Se ve claramente en la expresión anterior que los términos primero y segundo del segundo miembro son las sumas de cuadrados de los factores A y B respectivamente. A estas sumas las denominaremos SCF_A y SCF_B. El último término, que tiene una configuración menos clara, debe reflejar la interacción de los factores A y B. En efecto, si los dos primeros términos del segundo miembro reflejan el efecto individual de cada uno de los factores, entonces el término

Estadísticos e-Books & Papers

restante tiene que registrar el efecto conjunto, o de interacción, entre los factores A y B, es decir, el efecto de estos dos factores no recogido individualmente. El efecto de interacción será denominado $SCF_{A\times B}$. Por lo tanto una manera alternativa de escribir (6.16) es la siguiente:

$$SCT = SCF_A + SCF_B + SCF_{A\times B} + SCR \tag{6.17}$$

Los **grados de libertad**, asociados a cada una de las sumas de cuadrados que se han calculado, son los siguientes:

1. Los grados de libertad de SCT son iguales al número total de datos menos 1, es decir, $n-1$.

2. Los grados de libertad de SCF_A son iguales al número total de grupos menos 1, es decir $G-1$. Análogamente, para la SCF_B, los grados de libertad son $J-1$.

3. El número total de combinaciones de niveles de los factores A y B es igual a $G \times J$. Ahora bien, una vez calculadas las medias marginales por filas y por columnas a que nos hemos referido en el punto anterior, los grados de libertad de la $SCF_{A\times B}$ quedan reducidos a $(G-1)(J-1)$.

4. Los grados de libertad de la SCR son iguales al número total de datos menos el número total de celdas, es decir, $n - G \times J$.

Se debe verificar que, como en el caso de un factor, los grados de libertad de la SCT son iguales a la suma de los grados de libertad de cada uno de los componentes, es decir,

$$gl(SCT) = gl(SCF_A) + gl(SCF_B) + gl(SCF_{A\times B}) + gl(SCR)$$

$$n-1 = (G-1) + (J-1) + (G-1)(J-1) + (n - G \times J)$$

A partir de aquí, las medias cuadráticas se derivan de manera paralela al caso de un factor dividiendo las sumas de los cuadrados por los respectivos grados de libertad y ahora deberemos contar con tres estadísticos F, uno para analizar el efecto principal de un factor (si existen diferencias de medias significativas en el número de horas de televisión visionadas por nivel educativo), para el segundo (por sexo) y el efecto interacción de ambas variables, es decir:

$$F = \frac{MCF_A}{MCR}$$

$$F = \frac{MCF_B}{MCR}$$

$$F = \frac{MCF_{A\times B}}{MCR}$$

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 6.18.: Estadísticos descriptivos

```
        niveduc   sexo   N     mean       sd        se
1  Sin Estudios Hombre  43 3.930233 3.692949 0.5631695
2  Sin Estudios  Mujer  32 2.375000 1.288911 0.2278493
3     Primarios Hombre 276 2.786232 2.294965 0.1381407
4     Primarios  Mujer 251 2.494024 2.310620 0.1458450
5   Secundarios Hombre  46 2.326087 1.212057 0.1787081
6   Secundarios  Mujer  41 2.073171 1.034172 0.1615106
7         Grado Hombre  90 2.377778 2.074878 0.2187113
8         Grado  Mujer  94 1.957447 1.585654 0.1635477
9        Master Hombre  58 1.517241 1.217437 0.1598574
10       Master  Mujer  36 1.638889 1.312637 0.2187728
```

Cuadro 6.19.: Resultados de la estimación del ANOVA de dos factores

```
              Df Sum Sq Mean Sq F value   Pr(>F)
niveduc        4    165   41.13   9.201 2.69e-07 ***
sexo           1     33   33.09   7.402  0.00663 **
niveduc:sexo   4     32    8.09   1.809  0.12485
Residuals    957   4278    4.47
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Previamente, igual que hicimos para el caso de un factor, es muy importante realizar un análisis descriptivo de las variables (cuadro 6.18) y visualizar ese análisis en un gráfico de medias (figura 6.3). Del estudio de ambos se observan claramente dos tendencias, parece que en general los niveles de visualización de televisión de los hombres es superior al de las mujeres, lo que podría traducirse en un efecto principal significativo para el sexo, mientras que también se observa un decaimiento progresivo para ambos sexos del número de horas de televisión cuando aumenta el nivel educativo. De confirmarse estaríamos ante un efecto principal significativo para el nivel educativo.

Si efectuamos la estimación del modelo, la sintaxis es inmediata y totalmente paralela al caso de un factor, solo ha de notarse como se introduce en la misma el efecto interacción como `niveduc*sexo`. El cuadro 6.19 nos permite confirmar estadísticamente la impresión que nos mostraban los estadísticos descriptivos. Vemos que tanto el efecto del nivel educativo ($F(4,957) = 9{,}201; p < 0{,}01$) como el del sexo ($F(1,957) = 7{,}402; p < 0{,}01$) son significativos, con la interpretación que hemos dado de estos hechos al comentar los descriptivos. Sin embargo, el efecto interacción no es significativo ($F(4,957) = 8{,}09; p > 0{,}05$). Qué implica este resultado lo hemos de ver en la subsección siguiente.

```
fit<-aov(data=datos, tvhoras~niveduc+sexo+niveduc*sexo)
summary(fit)
```


Estadísticos e-Books & Papers

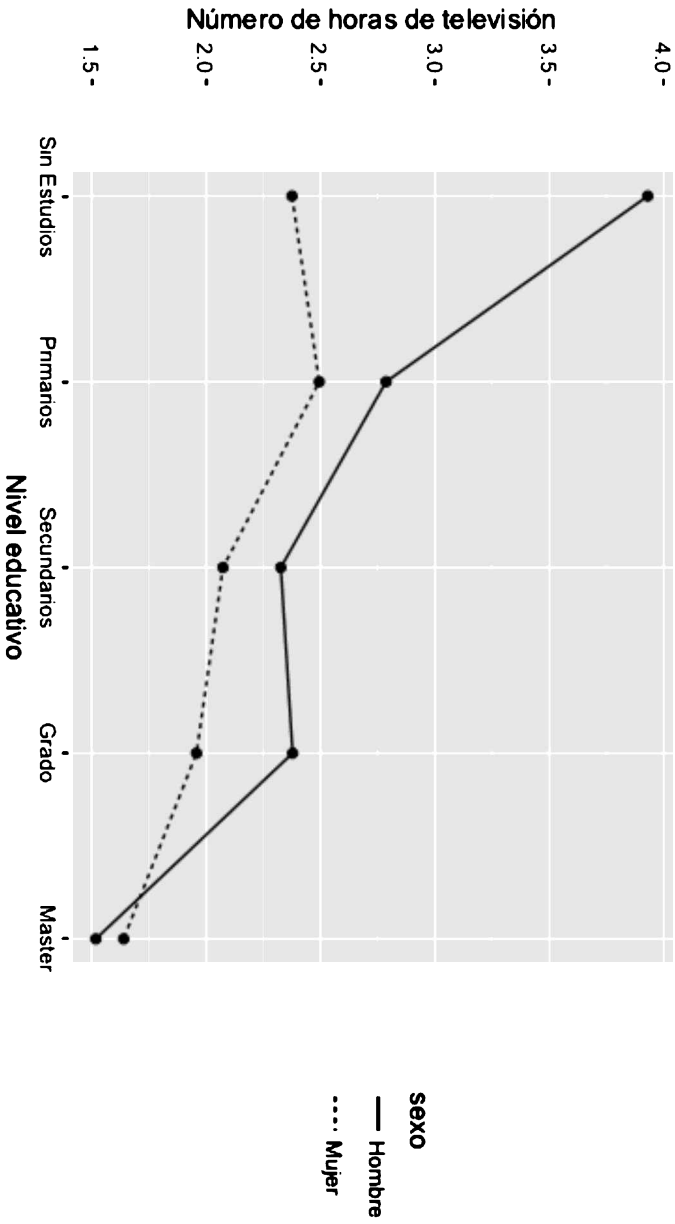
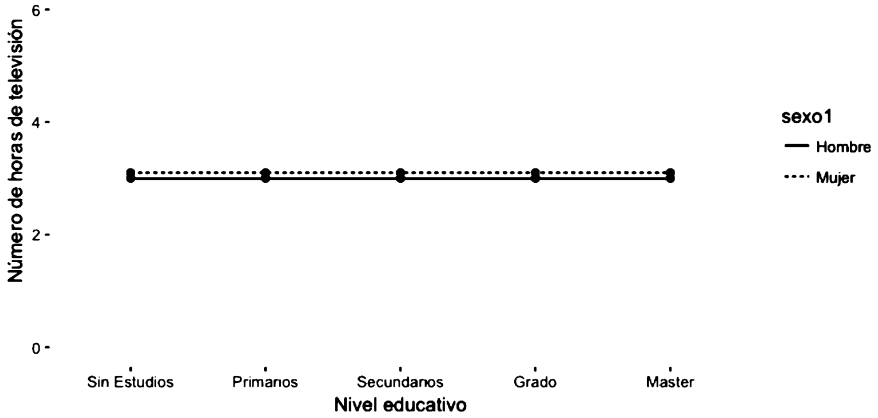


Figura 6.3.: Valores medios del consumo de televisión por sexo y por nivel educativo

© Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 6.4.: Ilustración del efecto interacción. Ni efectos principales ni efecto interacción



6.3.2. Ilustración del efecto interacción

De todo lo expuesto, creemos que lo único que requiere que el lector desarrolle una intuición clara es sobre el concepto de **interacción entre los factores**. La interacción del factor A y el B hace referencia a en qué medida uno de ellos —digamos el B— acentúa o inhibe el efecto del factor A sobre la variable dependiente. Hemos intentado, mediante las siguientes figuras, ilustrar distintas situaciones que pueden darse respecto a los **efectos principales** de los factores y el efecto interacción.

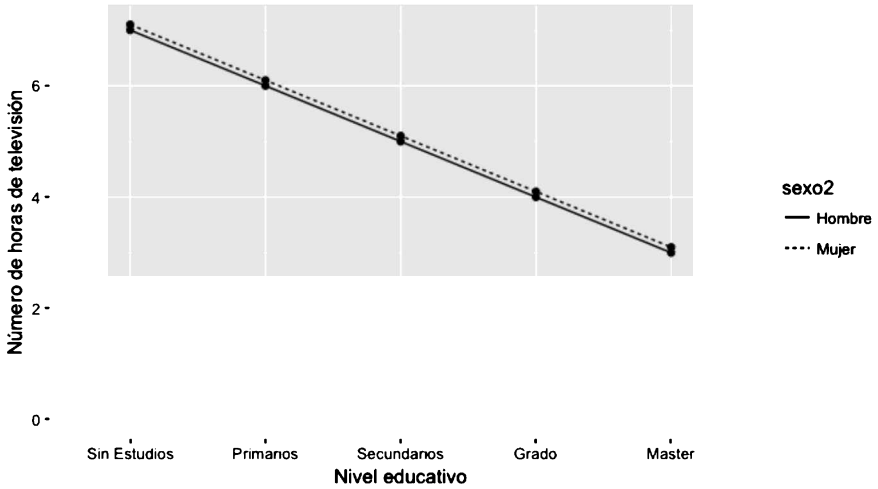
En la figura 6.4 se muestra una situación en que ninguno de los efectos principales, ni el nivel educativo ni el sexo, influyen sobre el número de horas de visionado, que es constante ante cambios en ambas variables. En el panel a) de la figura 6.5, sin embargo, el efecto del sexo no es significativo, pero el del nivel educativo sí hace bajar el número de horas de visionado. Que ambas líneas estén juntas implica, como hemos dicho, que, para cada nivel educativo, el número de horas que hombres y mujeres ven la televisión es el mismo.

Si pasamos al panel b) de la figura 6.5, vemos como ahora sí que hay diferencias por sexo, los hombres ven más la televisión que las mujeres, pero no influye el nivel educativo, pues el nivel de visionado se mantiene constante en cada sexo al aumentar el nivel educativo. La figura 6.6, al contrario, muestra que ahora los dos efectos son significativos, pues siempre los hombres ven más horas de televisión que la mujeres, pero además el nivel educativo hace caer en ambos casos a la misma tasa el número de horas de televisión.

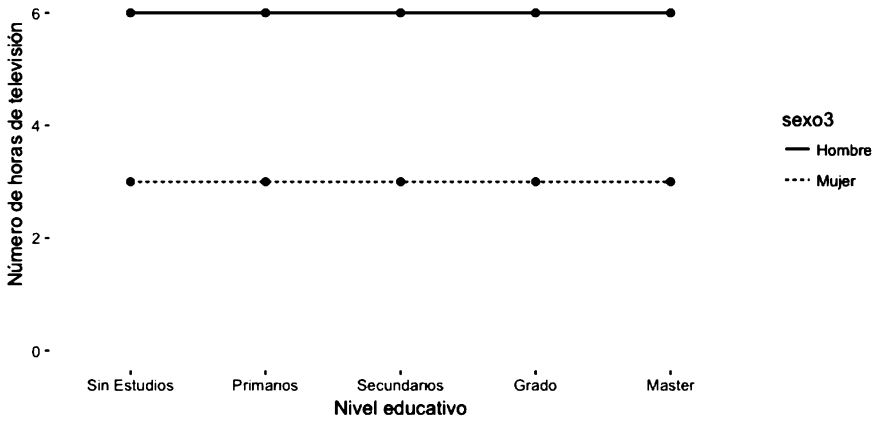
La figura relevante es la 6.7, que debe compararse con el que acabamos de comentar. Mientras que en el panel b) de la figura 6.6, el incremento en el nivel

Estadísticos e-Books & Papers

Figura 6.5.: Ilustración del efecto interacción. Solo uno de los efectos principales significativos. Sin interacción.



(a) Efecto nivel educativo significativo



(b) Efecto sexo significativo

© Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 6.6.: Ilustración del efecto interacción. Ambos efectos principales significativos. Sin interacción.

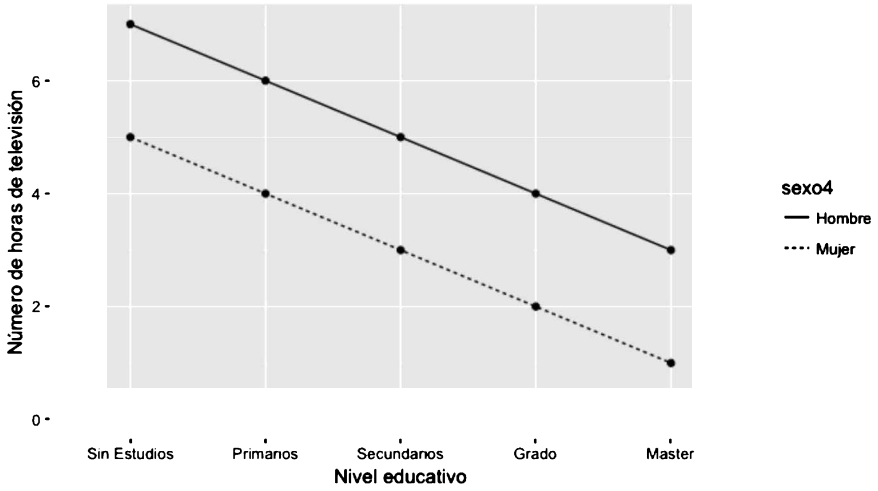
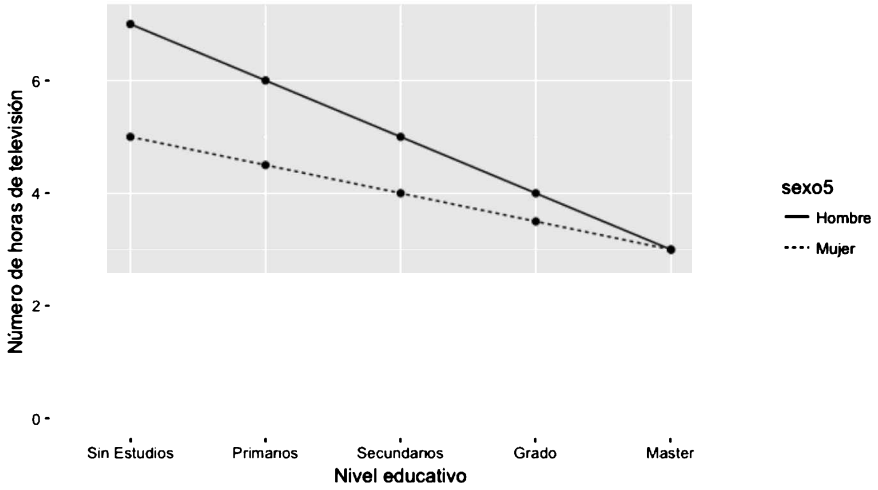


Figura 6.7.: Ilustración del efecto interacción. Efecto interacción significativo.



Estadísticos e-Books & Papers

educativo hace caer a la misma tasa —afecta igual— el número de horas de visionado en hombres y en mujeres, en la figura 6.7, el nivel educativo inhibe con mayor intensidad —hace caer a una mayor velocidad— el número de horas de televisión vista en el colectivo masculino que en el femenino. Esa diferente tasa es lo que denominamos un efecto de interacción significativo.

En el caso 6.3 obteníamos un resultado no significativo para el efecto interacción ($F(4,957) = 8{,}09; p > 0{,}05$). ¿Cómo debemos interpretarlo? Este resultado implica que, al crecer el nivel educativo, la velocidad a la que se reduce el número de horas de visionado no es estadísticamente distinta entre los hombres que entre las mujeres.

6.3.3. Medida de bondad del ajuste y tamaño del efecto

Hemos visto como en el análisis de la varianza con un factor se utilizaba el coeficiente de determinación (o estadístico eta cuadrado) definido en (6.14) como una medida de la variabilidad total explicada por el factor. En el análisis de la varianza con dos factores sigue siendo válido dicho coeficiente como una medida general del efecto de los factores sobre la variable dependiente. Con objeto de medir el efecto de cada factor (y de la interacción) separadamente se utilizan los estadísticos eta cuadrados parciales, que se definen de la siguiente forma:

$$\begin{aligned} \eta_A^2 &= \frac{SCF_A}{SCF_A + SCR} \\ \eta_B^2 &= \frac{SCF_B}{SCF_B + SCR} \\ \eta_{A\times B}^2 &= \times \frac{SCF_{A\times B}}{SCF_{AB} + SCR} \end{aligned} \tag{6.18}$$

Basta sustituir la información del cuadro 6.19 en las expresiones anteriores:

$$\eta_A^2 = \frac{165}{165 + 4278} = 0{,}037$$

$$\eta_B^2 = \frac{33}{33 + 4278} = 0{,}008$$

$$\eta_{A\times B}^2 = \frac{32}{32 + 4278} = 0{,}008$$

Pero, como vimos para el caso de un factor, era recomendable el cálculo del coeficiente ω^2, solo que este resulta algo más complicado en el caso de los diseños factoriales. Siguiendo a Field *et al.* (2012) y a Howell (2006), primero se ha de calcular la varianza para cada uno de los efectos (principales e interacción):

$$\hat{\sigma}_\alpha^2 = \frac{(a-1)\,(MCF_A - MCR)}{nab}$$

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

$$\hat{\sigma}^2_\beta = \frac{(b-1)\,(MCF_B - MCR)}{nab}$$

$$\hat{\sigma}^2_{\alpha\beta} = \frac{(a-1)\,(b-1)\,(MCF_{A\times B} - MCR)}{nab}$$

donde toda la notación es conocida salvo n, que es el tamaño muestral de cada condición, a es el número de niveles del primer factor y b el del segundo. En nuestro caso, el nivel educativo tiene $a = 5$ niveles, el sexo $b = 2$ y, dado que tenemos un tamaño muestral de 494 casos, asumiremos que en cada nivel hay, redondeando, 50 casos[2]. Con la información del cuadro 6.19 tendríamos:

$$\hat{\sigma}^2_\alpha = \frac{(5-1)\,(41{,}13 - 4{,}47)}{5\times 2\times 50} = 0{,}2933$$

$$\hat{\sigma}^2_\beta = \frac{(2-1)\,(33{,}09 - 4{,}47)}{5\times 2\times 50} = 0{,}0572$$

$$\hat{\sigma}^2_{\alpha\beta} = \frac{(5-1)\,(2-1)\,(8{,}09 - 4{,}47)}{5\times 2\times 50} = 0{,}0290$$

La variabilidad total, que es simplemente la suma de las anteriores, sería:

$$\hat{\sigma}^2_{total} = \hat{\sigma}^2_\alpha + \hat{\sigma}^2_\beta + \hat{\sigma}^2_{\alpha\beta} + MCR = 0{,}2933 + 0{,}0572 + 0{,}0290 + 4,47 = 4{,}8495$$

El tamaño del efecto es, entonces, para cada factor, la varianza calculada respecto a la total:

$$\omega^2_A = \frac{\hat{\sigma}^2_\alpha}{\hat{\sigma}^2_{total}} = \frac{0{,}2933}{4{,}8495} = 0{,}0653$$

$$\omega^2_B = \frac{\hat{\sigma}^2_\beta}{\hat{\sigma}^2_{total}} = \frac{0{,}0572}{4{,}8495} = 0{,}0127$$

$$\omega^2_{A\times B} = \frac{\hat{\sigma}^2_\alpha}{\hat{\sigma}^2_{total}} = \frac{0{,}0290}{4{,}8495} = 0{,}0065$$

como vemos algo superiores a los calculados mediante la eta cuadrado pero, en cualquier caso, pequeños. Si queremos compararlos con la r para poder aplicar los criterios de evaluación de Cohen (1988) habría que tomar raíces cuadradas, lo que nos da un 0,25 para el nivel de estudios, 0,11 para el sexo y 0,08 para la interacción, como decíamos, efectos pequeños.

[2]Esto no es así, no tenemos un diseño equilibrado y, por lo tanto, no tenemos $494/(5\times 2)$ sujetos por condición, pero creemos importante ilustrar el cálculo.

Estadísticos e-Books & Papers

6.3.4. Pruebas *post hoc*: comparaciones múltiples

Solo tiene sentido efectuar las comparaciones múltiples sobre los efectos principales y si estos son significativos. En nuestro caso, los dos efectos principales lo son. En la medida en que el sexo solo tiene dos grupos, carece de relevancia una prueba a posteriori, si el efecto es significativo, lo son las diferencias entre los dos únicos grupos.

Más interés tiene determinar entre qué grupos de nivel educativo hay diferencias, en cuanto que tenemos 5 de ellos. No repetiremos la sintaxis puesto que esta es análoga a la empleada en el caso de un factor. El cuadro 6.20 nos muestra los resultados de los principales test, de los que podemos concluir que las diferencias afloran principalmente entre los consumidores sin estudios y con estudios primarios frente a los que tienen estudios secundarios o superiores.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 6.20.: Pruebas *post hoc* para el nivel de estudios

```
      Pairwise comparisons using t tests with pooled SD

data:  datos$tvhoras and datos$niveduc

            Sin Estudios Primarios Secundarios Grado
Primarios   0.1834       -         -           -
Secundarios 0.0160       0.7376    -           -
Grado       0.0016       0.0794    1.0000      -
Master      2.8e-06      6.0e-05   0.4220      0.2636

P value adjustment method: bonferroni
> pairwise.t.test(datos$tvhoras,datos$niveduc,p.adjust="holm")

      Pairwise comparisons using t tests with pooled SD

data:  datos$tvhoras and datos$niveduc

            Sin Estudios Primarios Secundarios Grado
Primarios   0.0917       -         -           -
Secundarios 0.0112       0.1475    -           -
Grado       0.0013       0.0476    0.8740      -
Master      2.8e-06      5.4e-05   0.1266      0.1054

P value adjustment method: holm

  Tukey multiple comparisons of means
    95% family-wise confidence level

$niveduc
                              diff        lwr         upr     p adj
Primarios-Sin Estudios   -0.61960784 -1.3327369  0.09352119 0.1231745
Secundarios-Sin Estudios -1.05977011 -1.9702557 -0.14928455 0.0131227
Grado-Sin Estudios       -1.10362319 -1.8952421 -0.31200429 0.0013885
Master-Sin Estudios      -1.70283688 -2.5974898 -0.80818391 0.0000024
Secundarios-Primarios    -0.44016227 -1.1088526  0.22852809 0.3747058
Grado-Primarios          -0.48401535 -0.9788116  0.01078090 0.0587005
Master-Primarios         -1.08322904 -1.7301964 -0.43626166 0.0000526
Grado-Secundarios        -0.04385307 -0.7956870  0.70798086 0.9998551
Master-Secundarios       -0.64306676 -1.5027165  0.21658300 0.2456453
Master-Grado             -0.59921369 -1.3317942  0.13336684 0.1675528

lincon(formula = datos2)

                            psihat ci.lower ci.upper p.value
Grado vs. Primarios         0.35710 -0.35042  1.06462 0.14776
Grado vs. Secundarios       0.50566 -0.25727  1.26859 0.06077
Grado vs. Master            0.79643  0.06737  1.52549 0.00242
Grado vs. Sin Estudios      1.18621  0.47520  1.89722 0.00001
Primarios vs. Secundarios   0.14856 -0.29252  0.58964 0.34078
Primarios vs. Master        0.43933  0.07419  0.80447 0.00091
Primarios vs. Sin Estudios  0.82911  0.50523  1.15299 0.00000
Secundarios vs. Master      0.29077 -0.18537  0.76690 0.08717
Secundarios vs. Sin Estudios 0.68055  0.23275  1.12834 0.00004
Master vs. Sin Estudios     0.38978  0.01671  0.76285 0.00393
```


Estadísticos e-Books & Papers

Estadísticos e-Books & Papers

7. Análisis multivariante de la varianza

7.1. Introducción

El análisis multivariante de la varianza es una generalización del análisis de la varianza. A este último se le debería denominar, por contraposición, univariante, pero casi siempre se omite este calificativo. En el análisis multivariante de la varianza se consideran simultáneamente varias variables dependientes que supuestamente están relacionadas entre sí, en lugar de una sola variable dependiente que se examina en el análisis univariante. Si las variables analizadas no tuvieran relación entre ellas, no tendría interés el aplicar un análisis multivariante. En ese caso lo más indicado sería aplicar el análisis univariante de la varianza a cada una de las variables investigadas. El análisis multivariante de la varianza es conocido por las siglas inglesas MANOVA (*Multivariate ANalysis Of VAriance*). Esta es también la denominación usual en los programas de ordenador para análisis de datos.

Como ocurría con el análisis de la varianza, podemos utilizar una variable independiente que genere los grupos en los que se comparen las medias de una variable —ANOVA— o el vector de medias de distintas variables —MANOVA—, hablaríamos entonces de MANOVA de un factor, o podemos utilizar más variables independientes (MANOVA de dos factores, etc.)

Utilizando una ilustración de pocos casos para facilitar los cálculos, desarrollaremos el procedimiento de inferencia del MANOVA: hipótesis nula, lógica de los estadísticos para el contraste, análisis *post-hoc*, etc. Luego generalizaremos el análisis a la situación de dos factores y realizaremos los pertinentes estudios de casos más realistas con bases de datos más amplias.

7.2. Análisis multivariante de la varianza con un factor

Como ya se ha indicado, el análisis multivariante de la varianza es una generalización del análisis univariante de la varianza que se examinó en el apartado 6.2. En el análisis univariante nos referimos a la variable dependiente mediante un escalar; en el análisis multivariante designaremos al conjunto de variables dependientes mediante un vector, en el que cada uno de los elementos es una variable dependiente. Es decir, dada una variable dependiente X cuya media

Estadísticos e-Books & Papers

queremos evaluar en G grupos, la hipótesis nula de un ANOVA era:

$$H_0 : \overline{X}_1 = \overline{X}_2 = \cdots = \overline{X}_G$$

y la hipótesis se contrastaba mediante el estadístico F, cuya lógica de construcción desarrollamos. Sin embargo, en un MANOVA lo que queremos evaluar es si un conjunto de variables toma simultáneamente valores estadísticamente distintos en los G grupos que genera una variable dependiente, es decir, ahora la hipótesis nula es que :

$$H_0 : \boldsymbol{\mu_1} = \boldsymbol{\mu_2} = \cdots = \boldsymbol{\mu_G}$$

donde μ es el vector de medias para, digamos, K variables dependientes. Quizás se entienda mejor desdoblando el vector:

$$H_0 : \begin{bmatrix} \overline{X}_{11} \\ \overline{X}_{21} \\ \vdots \\ \overline{X}_{K1} \end{bmatrix} = \begin{bmatrix} \overline{X}_{12} \\ \overline{X}_{22} \\ \vdots \\ \overline{X}_{K2} \end{bmatrix} = \cdots = \begin{bmatrix} \overline{X}_{1k} \\ \overline{X}_{2k} \\ \vdots \\ \overline{X}_{KG} \end{bmatrix}$$

El lector se preguntará inmediatamente por qué no podemos encadenar K ANOVA en lugar de realizar un MANOVA, y la respuesta vuelve a ser la misma que dimos para explicar por qué no podíamos enlazar distintas pruebas t en lugar de realizar un ANOVA, porque cuantas más pruebas realizamos sobre los mismos datos, más se incrementa la probabilidad de cometer un error tipo I. Pero Field (2005) alega razones adicionales para realizar un MANOVA: no perdemos la información de la correlación entre las variables dependientes, por ello no solo podremos saber si las variables difieren en los grupos, sino también si combinaciones de las variables dependientes difieren entre los grupos. En esto abunda Stevens (2009) al señalar que medias que pueden no ser estadísticamente distintas individualmente, pueden serlo consideradas en su conjunto.

Ilustración 7.1 Datos para ilustración del MANOVA
Utilizaremos para ilustrar la lógica que sigue el MANOVA para el contraste de las hipótesis el siguiente ejemplo tomado de Stevens (2009). Partimos de que tenemos dos variables dependientes y_1 e y_2 y una variable independiente que nos genera tres grupos $G = 3$. El cuadro 7.1 ofrece los datos y una primera información descriptiva que se corresponde con la media de cada una de esas variables en cada uno de los grupos.

7.2.1. Descomposición de la matriz de covarianzas

Como para calcular las varianzas nos harán falta las distintas medias, obtengámoslas. El vector de las medias muestrales globales (es decir, de las dos variables en el conjunto de la muestra) se obtiene sumando para cada variable todos los valores de la muestra y dividiendo por el total de la muestra, es decir:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 7.1.: Datos para la ilustración del MANOVA

Grupo	Datos		$\overline{\boldsymbol{y}}_{\boldsymbol{g}}$	
	y_1	y_2	$\overline{y}_{1g}$	$\overline{y}_{2g}$
1	2	3	3	4
	3	4		
	5	4		
	2	5		
2	4	8	5	7
	5	6		
	6	7		
3	7	6	8.2	6.4
	8	7		
	10	8		
	9	5		
	7	6		
$\overline{\boldsymbol{y}}$	5.67	5.75		

Fuente: Stevens (2009, p. 178).

$$\overline{\boldsymbol{y}} = \begin{bmatrix} \overline{Y}_1 \\ \overline{Y}_2 \\ \vdots \\ \overline{Y}_K \end{bmatrix} = \begin{bmatrix} \dfrac{\sum_{g=1}^{G}\sum_{i=1}^{n_g} Y_{1gi}}{n} \\ \dfrac{\sum_{g=1}^{G}\sum_{i=1}^{n_g} Y_{2gi}}{n} \\ \vdots \\ \dfrac{\sum_{g=1}^{G}\sum_{i=1}^{n_g} Y_{Kgi}}{n} \end{bmatrix} \tag{7.1}$$

con los datos de la ilustración (cuadro 7.1):

$$\overline{\boldsymbol{y}} = \begin{bmatrix} \overline{Y}_1 \\ \overline{Y}_2 \end{bmatrix} = \begin{bmatrix} 5{,}67 \\ 5{,}75 \end{bmatrix}$$

Dentro de cada grupo se puede obtener el correspondiente vector de medias muestrales. Así, el vector de medias del grupo g vendrá dada por:

Estadísticos e-Books & Papers

$$\overline{\boldsymbol{y}}_g = \begin{bmatrix} \overline{Y}_{1g} \\ \overline{Y}_{2g} \\ \vdots \\ \overline{Y}_{Kg} \end{bmatrix} = \begin{bmatrix} \dfrac{\sum_{i=1}^{n_g} Y_{1gi}}{n_g} \\ \dfrac{\sum_{i=1}^{n_g} Y_{2gi}}{n_g} \\ \vdots \\ \dfrac{\sum_{i=1}^{n_g} Y_{Kgi}}{n_g} \end{bmatrix} \tag{7.2}$$

y que, de nuevo con los datos del cuadro 7.1, sería:

$$\overline{\boldsymbol{y}}_1 = \begin{bmatrix} \overline{Y}_{11} \\ \overline{Y}_{12} \\ \overline{Y}_{13} \end{bmatrix} = \begin{bmatrix} 3 \\ 5 \\ 8{,}2 \end{bmatrix}$$

$$\overline{\boldsymbol{y}}_2 = \begin{bmatrix} \overline{Y}_{21} \\ \overline{Y}_{22} \\ \overline{Y}_{23} \end{bmatrix} = \begin{bmatrix} 4 \\ 7 \\ 6{,}4 \end{bmatrix}$$

En el análisis de la varianza (univariante) se descompone la suma de cuadrados de las desviaciones con respecto a la media global en dos partes: la suma de cuadrados de las desviaciones entre la media de cada grupo y la media global —la varianza explicada por el factor—, y la suma de cuadrados de las desviaciones entre cada dato y la media de cada grupo —la varianza residual o no explicada por el factor—. Cuando se generaliza a K variables, en lugar de una varianza escalar, se dispone de una **matriz de covarianzas**. Es decir, en los momentos de segundo orden se pasa de escalares a matrices. La matriz a descomponer, a la que denominaremos **T**, es la matriz de suma de cuadrados y de productos cruzados en desviaciones con respecto a la media global, o, en forma abreviada, la **matriz de la suma de cuadrados y productos cruzados total (SCPCT)**.

¿Qué componentes tiene esta matriz **T**? Recordemos que en el ANOVA los cuadrados totales eran la suma de los cuadrados de las diferencias entre el valor de cada elemento y el centroide total de la muestra. En nuestro caso, lógicamente habrá un valor para cada variable así calculado en la diagonal de esa matriz. ¿Qué habrá fuera de la diagonal? La relación entre las variables, es decir, el producto cruzado de la diferencia entre cada valor y la media en la muestra para una variable multiplicada por lo mismo para la otra variable. En el fondo, la diagonal sería el cálculo que obtendríamos si hiciéramos dos ANOVA de un factor, la diferencia del MANOVA la aportan los elementos fuera de la diagonal. Construyamos la matriz:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

$$\mathbf{T} = \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix}$$

$$t_{11} = (2-5{,}67)^2 + (3-5{,}67)^2 + (5-5{,}67)^2 + (2-5{,}67)^2 +$$

$$+(4-5{,}67)^2 + (5-5{,}67)^2 + (6-5{,}67)^2 +$$

$$(7-5{,}67)^2 + (8-5{,}67)^2 + (10-5{,}67)^2 + (9-5{,}67)^2 + (7-5{,}67)^2 = 76{,}72$$

Para el segundo elemento de la diagonal:

$$t_{22} = (3-5{,}75)^2 + (4-5{,}75)^2 + (4-5{,}75)^2 + (5-5{,}75)^2 +$$

$$+(8-5{,}75)^2 + (6-5{,}75)^2 + (7-5{,}75)^2 +$$

$$(6-5{,}75)^2 + (6-5{,}75)^2 + (8-5{,}75)^2 + (5-5{,}75)^2 + (6-5{,}75)^2 = 28{,}25$$

Nos quedaría calcular los productos cruzados de fuera de la diagonal, nos apoyaremos para ello en el cuadro 7.2. Vemos que para cada dato restamos de su valor la media de esa variable para el grupo (D_1 para y_1 y D_2 para y_2) y a continuación multiplicamos ambos productos $D_1 \times D_2$. La suma de $D_1 \times D_2$ es el elemento fuera de la diagonal de $\mathbf{T}$, que queda entonces como:

$$\mathbf{T} = \begin{bmatrix} 76{,}72 & 26{,}00 \\ 26{,}00 & 28{,}25 \end{bmatrix}$$

Como ocurría en el análisis de la varianza, esta matriz se debe descomponer en dos, por un lado, debe estar el equivalente a la varianza explicada por el factor, que denominaremos matriz $\mathbf{F}$ o **matriz suma de cuadrados y productos cruzados del factor (SCPCF),** y, por otro lado, la matriz residual, que denominaremos W o matriz **suma de cuadrados y productos cruzados residual (SCPCR)**. En buena lógica:

$$\mathbf{T} = \mathbf{F} + \mathbf{W} \tag{7.3}$$

Comencemos con la matriz residual que contiene las diferencias de cada caso con la media del grupo al que pertenece. Como tenemos tres grupos ($G = 3$), tendremos tres matrices:

$$\mathbf{W} = \mathbf{W}_1 + \mathbf{W_2} + \mathbf{W}_3$$

veamos, con los datos del ejemplo, los componentes de una de ellas:

Estadísticos e-Books & Papers

Cuadro 7.2.: Cálculo de los productos cruzados de $\mathbf{T}$

Grupo	Datos		$\overline{\boldsymbol{y}}_{\boldsymbol{g}}$		D_1	D_2	
	y_1	y_2	$\overline{y}_{1g}$	$\overline{y}_{2g}$	$y_1 - \overline{y}_{1g}$	$y_2 - \overline{y}_{2g}$	$D_1 \times D_2$
1	2	3	3	4	-3.67	-2.75	10.08
	3	4			-2.67	-1.75	4.67
	5	4			-0.67	-1.75	1.17
	2	5			-3.67	-0.75	2.75
2	4	8	5	7	-1.67	2.25	-3.75
	5	6			-0.67	0.25	-0.17
	6	7			0.33	1.25	0.42
3	7	6	8.2	6.4	1.33	0.25	0.33
	8	7			2.33	1.25	2.92
	10	8			4.33	2.25	9.75
	9	5			3.33	-0.75	-2.50
	7	6			1.33	0.25	0.33
$\overline{\boldsymbol{y}}$	5.67	5.75				$\sum =$	26.00

$$\mathbf{W}_1 = \begin{bmatrix} ss_1 & ss_{12} \\ ss_{21} & ss_2 \end{bmatrix}$$

donde los elementos de la diagonal son la suma de los cuadrados residuales de un análisis de la varianza de un factor, es decir, los calculados con la diferencia del individuo a la media de su grupo para cada una de las dos variables. Los elementos fuera de la diagonal son la suma de los productos cruzados de esas diferencias. Con los datos del ejemplo, para la matriz $\mathbf{W}_1$ del grupo 1:

$$ss_1 = (2-3)^2 + (3-3)^2 + (5-3)^2 + (2-3)^2 = 6$$

$$ss_2 = (3-4)^2 + (4-4)^2 + (4-4)^2 + (5-4)^2 = 2$$

$$ss_{12} = ss_{21} = (2-3)(3-4) + (3-3)(4-4) + (5-3)(4-4) + (2-3)(5-4) = 0$$

Luego:

$$\mathbf{W_1} = \begin{bmatrix} 6 & 0 \\ 0 & 2 \end{bmatrix}$$

y análogamente se podría obtener que:

$$\mathbf{W_2} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \quad \mathbf{W_3} = \begin{bmatrix} 6{,}8 & 1{,}6 \\ 2{,}6 & 5{,}2 \end{bmatrix}$$

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

por lo tanto, la variabilidad residual vendría dada por la matriz:

$$\mathbf{W} = \begin{bmatrix} 14{,}8 & 1{,}6 \\ 1{,}6 & 9{,}2 \end{bmatrix}$$

Veamos ahora cuál sería la variabilidad explicada por el factor que se recoge en la matriz **F**.

$$\mathbf{F} = \begin{bmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{bmatrix}$$

En la diagonal de esta matriz estarán las diferencias de lo que el centroide de cada grupo difiere del centroide de la muestra completa, lógicamente con un elemento para cada variable, es decir, la misma definición que en el análisis de la varianza de un factor. De nuevo fuera de la diagonal estarán los productos cruzados. En todos los casos se multiplica por el tamaño muestral de cada grupo.

$$f_{11} = 4(3-5{,}67)^2 + 3(5-5{,}67)^2 + 5(8{,}2-5{,}67) = 61{,}87$$

$$f_{22} = 4(4-5{,}75)^2 + 3(7-5{,}75)^2 + 5(6{,}4-5{,}75) = 19{,}05$$

$$f_{12} = f_{21} =$$

$$= 4(3-5{,}67)(4-5{,}75) + 3(5-5{,}67)(7-5{,}75) + 5(8{,}2-5{,}67)(6{,}4-5{,}75) = 24{,}4$$

Por lo tanto:

$$\mathbf{F} = \begin{bmatrix} 61{,}87 & 24{,}40 \\ 24{,}40 & 19{,}05 \end{bmatrix}$$

El lector puede comprobar fácilmente que la matriz **T** (SCPCT) que habíamos calculado al principio coincide con la suma de los elementos en que la hemos descompuesto, es decir:

$$\mathbf{T} = \mathbf{F} + \mathbf{W} = \begin{bmatrix} 61{,}87 & 24{,}40 \\ 24{,}40 & 19{,}05 \end{bmatrix} + \begin{bmatrix} 14{,}8 & 1{,}6 \\ 1{,}6 & 9{,}2 \end{bmatrix} = \begin{bmatrix} 76{,}72 & 26{,}00 \\ 26{,}00 & 28{,}25 \end{bmatrix}$$

Aunque no sea necesario, porque creemos que la intuición aportada de la descomposición de la matriz que recoge la variación total en el MANOVA es suficiente, a efectos exclusivamente de tener las expresiones generales de la descomposición, la matriz **T** tendría la expresión que recoge (7.4), la matriz **F**, la recogida en (7.5), y la **W**, la que se puede comprobar en (7.6).

Estadísticos e-Books & Papers

$$\mathbf{T} = \begin{bmatrix} \sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(Y_{1gi}-\overline{Y}_1\right)^2 & \cdots & \sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(Y_{1gi}-\overline{Y}_1\right)\left(Y_{Kgi}-\overline{Y}_K\right) \\ \vdots & \ddots & \vdots \\ \sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(Y_{Kgi}-\overline{Y}_K\right)\left(Y_{1gi}-\overline{Y}_1\right) & \cdots & \sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(Y_{Kgi}-\overline{Y}_K\right)^2 \end{bmatrix} \tag{7.4}$$

$$\mathbf{F} = \begin{bmatrix} \sum_{g=1}^{G} n_g\left(\overline{Y}_{1g}-\overline{Y}_1\right)^2 & \cdots & \sum_{g=1}^{G} n_g\left(\overline{Y}_{1g}-\overline{Y}_1\right)\left(\overline{Y}_{Kg}-\overline{Y}_K\right) \\ \vdots & \ddots & \vdots \\ \sum_{g=1}^{G} n_g\left(\overline{Y}_{Kgi}-\overline{Y}_K\right)\left(\overline{Y}_{1g}-\overline{Y}_1\right) & \cdots & \sum_{g=1}^{G} n_g\left(\overline{Y}_{Kgi}-\overline{Y}_K\right)^2 \end{bmatrix} \tag{7.5}$$

$$\mathbf{W} = \begin{bmatrix} \sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(Y_{1gi}-\overline{Y}_{1g}\right)^2 & \cdots & \sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(Y_{1gi}-\overline{Y}_{1g}\right)\left(Y_{Kgi}-\overline{Y}_{Kg}\right) \\ \vdots & \ddots & \vdots \\ \sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(Y_{Kgi}-\overline{Y}_{Kg}\right)\left(Y_{1gi}-\overline{Y}_{1g}\right) & \cdots & \sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(Y_{Kgi}-\overline{Y}_{Kg}\right)^2 \end{bmatrix} \tag{7.6}$$

Antes de comenzar a explicar los distintos estadísticos que podemos utilizar para comprobar la hipótesis nula, veamos cómo estimar en `R` un MANOVA, aunque solo nos fijemos, por ahora, en la descomposición de las matrices que hemos efectuado. La sintaxis es bastante sencilla. Utilizamos la función **`manova{stats}`**, que mantiene la estructura de **`variable dependiente ~ variable independiente`**. Solo hay que tener en cuenta dos cosas: que, como tenemos dos variables dependientes, y_1 e y_2, le hemos de señalar que las tenga en cuenta a la vez lo que hacemos con **`cbind(y1,y2)`** y, en segundo lugar, que, al igual que ocurría con el ANOVA, es necesario especificar primero que la variable grupo ha de actuar como factor, lo que pedimos con datos **`$grupo <- factor(datos$grupo)`**. Para que en la salida, en el resumen del objeto estimado **`summary(fit)`**, nos aparezcan las matrices **F** y **W** lo hemos de solicitar con `$SS`.

```
datos<-data.frame(Ilustracion7_1)
datos$grupo <- factor(datos$grupo)
```

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 7.3.: Matrices residual y factorial del MANOVA

```
$grupo
         y1    y2
y1 61.86667 24.40
y2 24.40000 19.05

$Residuals
     y1  y2
y1 14.8 1.6
y2  1.6 9.2
```

```
fit<-manova(cbind(y1,y2)~grupo,data=datos)
summary(fit)$SS
```

El cuadro 7.3 nos ofrece las matrices generadas por el factor (grupo) y la residual; el lector puede comprobar que coinciden con las calculadas manualmente.

7.2.2. Cálculo del estadístico de contraste

El lector recordará que en el tema 6 explicábamos la lógica del estadístico F que se construía como la ratio entre la suma de los cuadrados explicados por el factor y los residuales. Como ambos suman el total, cuanto más grande se hace el denominador —más explica el factor— a la vez más pequeño se hace el numerador y por ello tanto más crece F.

En el MANOVA, aunque de forma matricial, tenemos los mismos componentes en las matrices $\mathbf{T}$, $\mathbf{F}$ y $\mathbf{B}$ y, por ello, los estadísticos tendrán definiciones similares.

A. Estadístico Λ de Wilks

El estadístico Λ de Wilks se define como la ratio entre la varianza residual y la total. Es importante darse cuenta, que, a diferencia del estadístico F que crecía cuando el factor tenía más capacidad para explicar las diferencias entre los grupos, la Λ de Wilks se comporta de manera contraria, cuanto más pequeño es el estadístico, menos representan los residuos sobre el total de la varianza y, por ello, más explica el factor. Su expresión es la siguiente:

$$\Lambda = \frac{|\boldsymbol{W}|}{|\boldsymbol{T}|} \tag{7.7}$$

donde toda la notación es conocida salvo que, dado que la división —o la multiplicación por la inversa— de dos matrices es otra matriz, para obtener un

Estadísticos e-Books & Papers

escalar se calculan primero los determinantes de las matrices, lo que indicamos con el símbolo $|\ |$.

El problema que tiene el estadístico Λ de Wilks es que no es sencillo determinar su distribución. El estadístico F en una ANOVA se distribuye como una F de Snedecor, pero no es tan inmediato para la Λ de Wilks. Existen, fundamentalmente, dos aproximaciones a la distribución. La F de Rao y la χ^2 de Barlett. Veamos ambas, aunque hay bastante consenso en que la primera es una mejor aproximación (Lohnes, 1961).

Aproximación χ^2 de Barlett. Stevens (2009) señala que la siguiente función:

$$\chi^2 = -\left[(N-1) - ,5\,(K+G)\right] \ln \Lambda \tag{7.8}$$

se distribuye como una χ^2 con $K(G-1)$ grados de libertad y, por lo tanto, puede conocerse su significatividad. N es el tamaño muestral, K es el número de variables dependientes, y G, el número de grupos.

Aproximación F de Rao. El estadístico propuesto por Rao, al que denominaremos Ra, tiene la siguiente expresión:

$$Ra = \frac{1-\Lambda^{1/s}}{\Lambda^{1/s}} \times \frac{1+ts-K\,(G-1)\,/2}{K(G-1)} \tag{7.9}$$

donde t y s están definidos de la siguiente forma:

$$t = N-1-(K+G)/2 \tag{7.10}$$

$$s = \sqrt{\frac{K^2(G-1)^2-4}{K^2+(G-1)^2-5}} \tag{7.11}$$

En (7.11), si $K^2+(G-1)^2 < 5$ entonces s es un número infinito o imaginario. En ese caso, Rao proponer hacer $s=1$. El estadístico Ra, bajo la hipótesis nula de igualdad de los vectores de medias, se distribuye de manera aproximada como una F con $K(G-1)$ grados de libertad en el numerador y $1+ts-K(G-1)/2$ grados de libertad en el denominador. Sin embargo, como demuestra Tatsuoka (1971), cuando $G=2$ o $G=3$ la distribución es exactamente una F independientemente del número K de variables.

Calculando el estadístico Λ de Wilks para nuestro ejemplo:

$$\Lambda = \frac{\begin{vmatrix} 14{,}8 & 1{,}6 \\ 1{,}6 & 9{,}2 \end{vmatrix}}{\begin{vmatrix} 76{,}72 & 26 \\ 26 & 28{,}25 \end{vmatrix}} = \frac{14{,}8 \times 9{,}2 - 1{,}6^2}{76{,}72 \times 28{,}25 - 26^2} = 0{,}0897$$

y el estadístico χ^2 asociado según la fórmula de Barlett sería:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 7.4.: Contraste de hipótesis con la λ de Wilks

```
          Df    Wilks approx F num Df den Df    Pr(>F)
grupo      2 0.089674   9.3575      4     16 0.0004271 ***
Residuals  9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$\chi^2 = -[(12-1) - 0,5(2+3)]\ln(0{,}0897) = 20{,}4987 \; gl = 2(3-1) = 4$$

y, como el valor crítico para una $\chi^2(4) = 9{,}49$, para $\alpha = 0{,}05$ podríamos rechazar la hipótesis nula y concluir que existe un efecto significativo del factor.

Si usamos la aproximación de Rao, entonces:

$$t = 12 - 1 - (2+3)/2 = 8{,}5$$

$$s = \sqrt{\frac{2^2(3-1)^2 - 4}{2^2 + (3-1)^2 - 5}} = \sqrt{\frac{12}{3}} = 2$$

por lo que:

$$Ra = \frac{1 - 0{,}0897^{1/2}}{0{,}08971^{1/2}} \times \frac{1 + 8{,}5 \times 2 - 2\,(3-1)\,/2}{2(3-1)} = 9{,}357$$

y, como el valor crítico para una F con $2(3-1) = 4$ grados de libertad en el numerador y $1 + 8,5 \times 2 - 2(3-1)/2 = 16$ en el denominador es 3,0069, puede rechazarse la hipótesis nula de igualdad de medias como ocurría con la aproximación de Barlett.

`R` tiene implementada la aproximación de Rao, luego solicitando que nos muestre el contraste asociado a Wilks con la siguiente sintaxis, el cuadro 7.4 nos permite comprobar como todos los cálculos manuales se corresponden con la salida del MANOVA.

```
summary(fit, test="Wilks")
```

Field (2005) y Field *et al.* (2012) plantean otra interesante aproximación a la lógica de los estadísticos del contraste del MANOVA, cuya equivalencia queremos mostrar, sobre todo porque facilita mucho el cálculo del resto de estadísticos que nos quedan. Cuando vimos el estadístico F para el ANOVA, lo planteábamos como la ratio entre lo que explica el factor y los residuos:

$$F = \frac{MCF}{MCR} \tag{7.12}$$

Estadísticos e-Books & Papers

Cuadro 7.5.: $\mathbf{FW^{-1}}$

```
> F.InvW
         [,1]     [,2]
[1,] 3.968064 1.962076
[2,] 1.452096 1.818114
```

Lo mismo puede plantearse para el MANOVA, la diferencia es que hablamos de matrices y la división es una multiplicación por la inversa:

$$Estadístico = \mathbf{FW^{-1}} \tag{7.13}$$

Con los datos de nuestro ejemplo, en R, podríamos calcularlo como sigue y el resultado sería la matriz del cuadro 7.5. Esa matriz representa la ratio entre la varianza explicada por el modelo y la residual, pero, como vemos, contiene cuatro valores, no uno, con tres variables dependientes, tendría 9, y así sucesivamente. La cuestión es cómo convertir esa matriz en un valor único que tenga sentido y que sirva para aplicar el test.

```
F = matrix(c(61.86667, 24.40, 24.40, 19.05), nrow=2,ncol=2)
W = matrix(c(14.8,1.6,1.6,9.2), nrow=2,ncol=2)
# Invertimos la matriz de residuos Inv.W
Inv.W<-solve(W)
# Multiplicamos F por la inversa(Inv.W)
F.InvW=F%*%(Inv.W)
```

La forma de hacerlo sería convertir las variables dependientes en variables latentes o factores como una combinación de ellas, tal y como veremos en el capítulo 13. Con esto tendríamos una única variable (la latente) y esa matriz se convertiría en un único estadístico. De lo que se trata, en definitiva, es de encontrar las funciones discriminantes como combinación de las variables dependientes que mejor explicarían que un sujeto perteneciera al grupo 1, al 2 o al 3. Esas funciones discriminantes son simplemente combinaciones lineales de las variables dependientes y sus coeficientes lineales son los autovectores de la matriz $\mathbf{FW^{-1}}$. El primer autovector (y por ello la primera función) siempre será responsable de explicar un mayor porcentaje de la separación entre los grupos que el segundo. Si calculamos los autovectores sobre la matriz $\mathbf{FW^{-1}}$, estaremos encontrando la función que maximiza la ratio entre varianza sistemática y residual (primer autovector), mientras que el segundo tendrá menores valores de este ratio. Si tenemos dos variables y dos grupos, tendremos un único autovector y podremos calcular directamente el equivalente al estadístico F. En general para K variables y G grupos tendremos tantos autovectores como el menor valor de K y $G-1$. En nuestro ejemplo, $K = 2$ y $G - 1 = 2$, tenemos dos autovectores.

Los autovalores son una medida de la longitud de ese autovector. Al calcular

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 7.6.: Autovalores y autovectores de $\mathbf{FW}^{-1}$

```
> print(vec_y)
          [,1]       [,2]
[1,] 0.9043088 -0.5377583
[2,] 0.4268790  0.8430991

> print(val_y)
[1] 4.8942620 0.8919159
```

los autovectores de la matriz que contiene la ratio entre varianza explicada y residual, tenemos muchos menos datos para resumir el estadístico. Con los autovalores como indicadores son todavía menos. En esos autovalores son en los que se basan los estadísticos. Sabemos, además, que el primer autovalor resume mejor el ratio varianza sistemática sobre varianza residual, hecho que tienen en cuenta los autores de los estadísticos. El cuadro 7.6 muestra el cálculo de los autovectores y autovalores en nuestro ejemplo, que se obtiene en R fácilmente:

```
vec_y<-eigen(F.InvW)$vec
val_y<-eigen(F.InvW)$val
```

En el caso de la Λ de Wilks, una formulación alternativa al estadístico que hemos derivado con anterioridad, y que usa los autovalores como hemos explicado, es la siguiente:

$$\Lambda = \prod_{i=1}^{s} \frac{1}{1+\lambda_i} \tag{7.14}$$

donde λ_i es cada uno de los s autovalores fruto de la descomposición de la matriz. Aplicados a nuestro ejemplo:

$$\Lambda = \frac{1}{1+4{,}8942620} \times \frac{1}{1+0{,}8919159} = 0{,}0897$$

valor que, como puede comprobarse, coincide con el calculado bajo el enfoque anterior. Seguiremos este planteamiento de los autovalores para el resto de estadísticos propuestos para el contraste de la hipótesis nula.

B. Traza de Pillai-Bartlett (V)

La traza de Pillai-Barlett es la suma de la proporción de varianza explicada por cada una de las funciones discriminantes que obtenemos con los autovectores:

$$V = \sum_{i=1}^{s} \frac{1}{1+\lambda_i} \tag{7.15}$$

Estadísticos e-Books & Papers

Cuadro 7.7.: Traza V de Pillai-Bartlett

```
> summary(fit, test="Pillai")
          Df Pillai approx F num Df den Df    Pr(>F)
grupo      2 1.3018   8.3899      4     18 0.0005283 ***
Residuals  9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

La V sigue una distribución aproximada a una F. Para nuestro ejemplo tomaría el siguiente valor:

$$V = \frac{1}{1+4{,}8942620} + \frac{1}{1+0{,}8919159} = 1{,}301779$$

que como puede comprobarse en el cuadro 7.7 coincide con el valor obtenido mediante `manova{stats}` a través de la instrucción que se añadiría a las que hemos mostrado para la Λ de Wilks:

```
summary(fit, test="Pillai")
```

C. T^2 de Hotelling

La traza T^2 de Hotelling-Lawley es simplemente la suma de los autovalores de cada función discriminante, por lo tanto, la suma de la ratio entre la varianza explicada por el modelo y la residual para cada función y se interpreta como la ratio F del ANOVA.

$$T = \sum_{i=1}^{s} \lambda_i \tag{7.16}$$

que aplicado a nuestro ejemplo tomaría el valor:

$$T = 4{,}8942620 + 0{,}8919159 = 5{,}786178$$

y que, igual que en el caso anterior y como puede comprobarse en el cuadro 7.8, coincide con el valor obtenido mediante `manova{stats}` a través de la instrucción que se añadiría a las que hemos mostrado para la Λ de Wilks:

```
summary(fit, test="Hotelling")
```

D. Raíz mayor de Roy

Es, simplemente, el autovalor correspondiente al primer autovector, es decir, el mayor de todos los s obtenidos:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 7.8.: Traza T^2 de Hotelling

```
> summary(fit, test="Hotelling")
          Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
grupo      2           5.7862   10.126      4     14 0.0004576 ***
Residuals  9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cuadro 7.9.: Raíz mayor de Roy

```
          Df    Roy approx F num Df den Df    Pr(>F)
grupo      2 4.8943   22.024      2      9 0.0003412 ***
Residuals  9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$\textit{Raíz Roy} = \text{máx}\,(\lambda_s) \tag{7.17}$$

siendo, por tanto, la ratio varianza explicada sobre residual de la primera función discriminante. Para nuestro ejemplo se corresponde con $\lambda = 4{,}8942620$ que es lo que muestra la función `manova{stats}` cuando le añadimos a la estimación la sintaxis, tal como vemos en el cuadro 7.9:

```
summary(fit, test="Roy")
```

Si quisiéramos obtener el cálculo manual con R, cosa innecesaria porque aplicamos la función `manova{stats}` pero útil didácticamente, bastaría con operativizar las fórmulas del siguiente modo:

```
lambdaW=(1/(1+val_y[1:1])*(1/(1+val_y[2:2])))
Pillai=(val_y[1:1]/(1+val_y[1:1]))+(val_y[2:2]/(1+val_y[2:2]))
HotelingT=val_y[1:1]+val_y[2:2]
Roy=max(val_y[1:1],val_y[2:2])
```

E. Elección del estadístico adecuado

Como señala Field (2005), solo cuando se genera una única función discriminante el valor de los estadísticos presentados será el mismo, por lo que para cualquier otra situación es necesario tener algún criterio para elegir. Tanto Olson (1974;1976;1979) como Stevens (1979) han analizado la potencia de prueba de los distintos estadísticos y concluyen que difieren poco para tamaños muestrales pequeños y medios.

Estadísticos e-Books & Papers

Cuando los grupos están bien separados por solo una función discriminante, la raíz de Roy es la de mayor potencia porque se centra solo en su autovalor, seguida por la T de Hotelling, la λ de Wilks y la traza de Pillai. Sin embargo, cuando más de una función discriminante explica las diferencias, entonces el orden de preferencia es el contrario.

Todos los estadísticos son bastante robustos ante violaciones de la normalidad, y la raíz de Roy es muy sensible a violaciones de la hipótesis de igualdad de la matriz de covarianzas (Stevens, 1979). Bray y Maxwell (1985) concluyen que, cuando las muestras son iguales por grupo, la traza de Pillai es el estadístico más robusto ante violaciones de los supuestos.

En nuestro ejemplo, vemos que, se aplique el test que se aplique, los resultados coinciden en rechazar la hipótesis nula de igualdad del vector de medias en los tres grupos que forma la variable independiente. Al igual que ocurría en el ANOVA, será necesario establecer si los tres grupos tienen vectores de medias distintos o solo uno difiere de los otros dos (pruebas *post hoc*), pero, previamente, hemos de comprobar las condiciones de aplicabilidad del MANOVA.

7.2.3. Comprobación de los supuestos en los que se basa el MANOVA

A. Normalidad multivariante

La normalidad multivariante es una exigencia mucho más limitativa que la univariante. La normalidad univariante es condición necesaria para la multivariante, pero no suficiente. Pero un primer paso es, por tanto, comprobar la normalidad univariante. Dado que se cumple que cualquier par de variables extraídas de una distribución normalmente multivariante también lo serán, una prueba parcial de normalidad multivariante pasaría por analizar los gráficos de dispersión entre cada par de variables. Como señala Stevens (2009), la normalidad bivariante implica que los gráficos de dispersión para cada par de variables deberían ser elípticos y cuanto mayor sea la correlación, más delgada debería ser la elipse.

Al trabajar en R, sí que disponemos de un estadístico para analizar la normalidad multivariante que no contamos en la mayoría de programas, lo que nos puede evitar los pasos anteriores. Está disponible mediante la función `mshaphiro.test{mvnormtest}`. Esta función se aplica grupo a grupo, por lo que primero es necesario segregar la base de datos en los grupos, tres en nuestro ejemplo. Otro detalle es que el test necesita que las variables estén en filas, por lo que es necesario transponer la base de datos. Estos detalles se sintetizan en esta sintaxis:

```
library(mvnormtest)
#desagregamos la base por grupos
grupo1<-datos[1:4,2:3]
grupo2<-datos[5:7,2:3]
```

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 7.10.: Pruebas de normalidad multivariante

```
> mshapiro.test(grupo1)

     Shapiro-Wilk normality test

data:  Z
W = 0.82743, p-value = 0.1612

> mshapiro.test(grupo2)

     Shapiro-Wilk normality test

data:  Z
W = 0.75, p-value < 2.2e-16

> mshapiro.test(grupo3)

     Shapiro-Wilk normality test

data:  Z
W = 0.66338, p-value = 0.003809
```

```
grupo3<-datos[8:12,2:3]
#El test necesita las variables en filas, transponemos
grupo1<-t(grupo1)
grupo2<-t(grupo2)
grupo3<-t(grupo3)
#Ejecutamos el test
mshapiro.test(grupo1)
mshapiro.test(grupo2)
mshapiro.test(grupo3)
```

Aunque estamos trabajando con datos simulados y muy pocos casos para poder ilustrar los cálculos que hemos realizado con anterioridad y, por tanto, no cabe esperar resultados de normalidad, el cuadro 7.10 muestra las salidas correspondientes al test de normalidad multivariante, donde solo habría normalidad en el primer grupo. Para el resto deberían plantearse transformaciones como las que mostrábamos en el capítulo 2.

B. Homocedasticidad

La validez de los estadísticos construidos para el contraste está condicionada al cumplimiento, entre otras condiciones, de que la matriz de covarianzas sea la misma para todas las poblaciones o grupos. Por lo tanto, la hipótesis nula

Estadísticos e-Books & Papers

para contrastar la homocedasticidad sería:

$$H_0: \ \mathbf{\Sigma_1} = \mathbf{\Sigma_1} = \cdots = \mathbf{\Sigma_G} \tag{7.18}$$

Para determinar si la matriz de covarianzas es la misma para los distintos grupos se puede utilizar el contraste de Barlett-Box[1], que utiliza el estadístico M. Este estadístico se define de la siguiente forma:

$$M = \frac{\prod_{g=1}^{K} |\mathbf{S_g}|^{(n_g-1)/2}}{|\overline{\mathbf{S}}|^{(n-K)/2}} \tag{7.19}$$

donde:

$$\mathbf{S_g} = \frac{\mathbf{W_g}}{n_g - 1}$$

$$\overline{\mathbf{S}} = \frac{\sum_{g=1}^{G} \mathbf{W_g}}{n-G} = \frac{\sum_{g=1}^{G} (n_g - 1)\mathbf{S_g}}{n-G}$$

La matriz $\mathbf{S}_g$ es una estimación de la matriz de covarianzas correspondiente a la celda g-ésima $\mathbf{\Sigma_g}$, mientras que $\overline{\mathbf{S}}$ es una estimación de la matriz de covarianzas global $\mathbf{\Sigma}$. Cuando el numerador de (7.19), donde aparecen los determinantes de las estimaciones de la matriz de covarianzas para cada grupo, sea muy superior al denominador, donde aparece el determinante de la estimación global de la matriz de covarianzas, será indicativo de que existe heterocedasticidad, es decir, de que no existe homogeneidad entre las matrices de covarianzas de cada grupo.

Desgraciadamente el estadístico M no tiene una distribución exacta. Se han obtenido, sin embargo, aproximaciones a las distribuciones F por Box y ji-cuadrado por Barlett. De las dos aproximaciones en general es mejor la primera, pero la implementada en `R` en la funcion `boxM{biotools}` es la segunda. En el caso que estamos utilizando como ilustración, se solicitaría de este modo:

```
library(biotools)
boxM(datos[2:3],datos[,1])
```

El resultado, como muestra el cuadro 7.11, confirmaría la igualdad en las matrices de covarianzas al no poder rechazarse la hipótesis nula.

[1]Este contraste fue propuesto inicialmente por Barlett en 1947. Posteriormente Box (1949) desarrolló un procedimiento para aproximarlo a una distribución F.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 7.11.: Test M de Box para analizar la igualdad de las matrices de covarianzas

```
        Box's M-test for Homogeneity of Covariance Matrices

data:  datos[2:3]
Chi-Sq (approx.) = 1.2282, df = 6, p-value = 0.9755
```

C. Test de esfericidad

Con el contraste de esfericidad, propuesto por Barlett en 1950, se pretende dar respuesta a la cuestión de si existe o no una relación significativa entre las variables analizadas. De no existir esta correlación, la secuencia de varios ANOVA sería una aproximación igual de razonable. La matriz de correlación poblacional $\mathbf{R_p}$ recoge la relación existente entre cada par de variables. La diagonal principal de una matriz de correlación está formada por 1, mientras que los elementos fuera de la diagonal principal ρ_{ij} son coeficientes de correlación entre cada par de variables. Si todos los coeficientes ρ_{ij} son nulos (es decir, si no existe ninguna relación entre las K variables), la matriz $\mathbf{R_p}$ será igual a la matriz identidad, con lo que su determinante será igual a la unidad. Consecuentemente con la argumentación anterior, la hipótesis nula en el contraste de esfericidad es la siguiente:

$$H_0: \ |\mathbf{R_p}| = 1 \tag{7.20}$$

El estadístico para contrastar la hipótesis anterior está basado en la matriz de correlación muestral de los residuos, a la que denominaremos $\mathbf{R}$:

$$\chi^2_{\frac{1}{2}(K^2-K)} = -\left[n - 1 - \frac{1}{6}(2K+5)\right] \ln|\mathbf{R}| \tag{7.21}$$

En el caso de que se acepte la hipótesis nula (7.20) se debería abandonar el enfoque multivariante y aplicar el análisis de la varianza por separado a cada una de las variables dependientes.

`R` no tiene implementado para el MANOVA el test de esfericidad de Barlett, pero sí de manera general para cualquier técnica en la función `cortest.bartlett {psych}`. Basta con que la matriz residual $\mathbf{W}$, que contiene la sumas de los cuadrados y productos cruzados residuales para los tres grupos, la transformemos en una matriz de correlación para poder aplicar directamente el test. Primero habría que pasarla a una matriz de varianzas y covarianzas, pero esto es sencillo. Las varianzas[2] son simplemente la diagonal de $\mathbf{W}$ (que contiene los cuadrados

[2]Aunque es estadística básica, recordemos que la varianza de una variable x es $\sum_i (x_i - \overline{x})^2/(N-1)$, es decir, la suma de los cuadrados de las diferencias respecto a la media dividida por el tamaño muestral menos 1, y la covarianza entre x e y, la suma de los productos cruzados dividido por ese mismo valor: $\sum_i (x_i - \overline{x})(y_i - \overline{y})/(N-1)$. Los numeradores están en $\mathbf{W}$, luego solo falta dividir por los denominadores. Pero recuérdese que $\mathbf{W} = \mathbf{W}_1 + \mathbf{W_2} + \mathbf{W_3}$

Estadísticos e-Books & Papers

Cuadro 7.12.: Test de esfericidad de Barlett

```
> cortest.bartlett(R,n=9)
$chisq
[1] 0.1233726

$p.value
[1] 0.725405

$df
[1] 1
```

de las diferencias respecto a la media) dividido el número de grados de libertad de los residuos en la estimación del MANOVA que, como se comprueba en el cuadro 7.9, son $df = 9$. Las covarianzas son simplemente los elementos fuera de la diagonal que contiene los productos cruzados, dividido también por $N-1$ y, al ser suma de G grupos, por $G(N-1)$, que, de nuevo, es equivalente al número de grados de libertad de los residuos en la estimación del MANOVA. Luego esa matriz de varianzas y covarianzas se pasa a matriz de correlaciones mediante la función `cov2cor{stats}` y se le aplica el test de Barlett, cuyo resultado se aprecia en el cuadro 7.12 y que no permitiría rechazar la hipótesis nula de matriz identidad, lo que no es de extrañar en un contexto simulado de tan solo 12 casos, posteriormente veremos una aplicación más realista.

```
W = matrix(c(14.8,1.6,1.6,9.2), nrow=2,ncol=2)
COV.W=W/(9)
R<-cov2cor(COV.W)
cortest.bartlett(R,n=9)
```

7.2.4. Bondad de ajuste y potencia de prueba

Para analizar la bondad del ajuste se utiliza como medida de bondad del ajuste un estadístico que es una generalización del estadístico (7.7). Antes de examinarlo conviene recordar que el estadístico Λ puede interpretarse como la razón entre la suma de cuadrados generalizada residual y la suma de cuadrados generalizada total. Por lo tanto, la ratio es la proporción de la suma de cuadrados generalizada total no explicada por el factor, mientras que $1-\Lambda$ será la proporción explicada por el factor. Esta proporción, a la que se denomina también eta cuadrado (η^2), se toma como medida de bondad del ajuste, siendo una generalización del coeficiente de determinación al caso multivariante. Su expresión es la siguiente:

porque tenemos $K = 3$ grupos, de tal forma que habrá que dividir por $K(N-1)$, siendo N el tamaño medio de los grupos o, dicho de otra forma, el tamaño muestral divido por el número de grupos.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

$$\eta^2 = 1 - \Lambda = 1 - \frac{|\mathbf{W}|}{|\mathbf{T}|} \tag{7.22}$$

Un valor próximo a 1 indica que la mayor parte de la variabilidad total puede atribuirse al factor, mientras que un valor próximo a 0 significa que el factor explica muy poco de esa variabilidad total.

En el caso de ejemplo ya obtuvimos que $\Lambda = 0{,}08967445$, por lo que es inmediato que $\eta^2 = 1 - 0.08967445 = 0.9103255$. De todos modos, como ejercicio elemental de sintaxis de `R`, puede repetirse su cálculo a partir de las matrices originales:

```
F = matrix(c(61.86667, 24.40, 24.40, 19.05), nrow=2,ncol=2)
W = matrix(c(14.8,1.6,1.6,9.2), nrow=2,ncol=2)
T=F+W
detW=det(W) detT=det(T)
eta2=1-detW/detT
```

Aunque muchos trabajos han abordado el cálculo de la potencia de la prueba para MANOVA de más de dos grupo (Ito, 1962; Pillai y Jayachandian, 1967; Olson, 1974; Lauter, 1978), las tablas que ofrecen tienen muchos huecos en cuanto al número de variables dependientes y/o grupos. Cuando los cuadros son bastante exhaustivos, como es el caso de Lauter (1978), se centran solo en alguno de los estadísticos, en este caso la T^2 de Hotelling. El hecho de que las diferencias de potencia entre los distintos estadísticos sean reducidas para muestras medias y pequeñas (Olson, 1974) convierte a las tablas de Lauter (1978) en la mejor opción para abordar los análisis de potencia del MANOVA.

7.2.5. Pruebas *post hoc*

Hay bastante controversia acerca del mejor enfoque para realizar el análisis *post hoc* de un MANOVA, es decir, una vez establecido que el vector de medias es estadísticamente diferente en los grupos del factor, cómo determinar si esas diferencias son significativas en todos los grupos o solo en algunos pares de ellos. Algunos autores como Field (2005) y Field *et al.* (2012) abogan por la realización de un análisis discriminante, es decir, evaluar qué combinación de las variables dependientes explica la pertenencia a los grupos. Sin embargo, este enfoque no evalúa la significatividad de las diferencias, aunque sí establece la contribución relativa de las variables dependientes al efecto de los grupos.

Otros autores como Stevens (2009) optan por un planteamiento más ortodoxo en el sentido de más similar el enfoque seguido en el ANOVA. El enfoque de este autor es el siguiente: realizar un MANOVA para cada par de grupos que genera el factor (en nuestro ejemplo serían tres MANOVA 1:2, 1:3 y 2:3). Estos MANOVA vendrían seguidos de pruebas t para establecer la contribución relativa de cada variable dependiente a explicar la diferencia en los dos grupos de cada uno de los tres MANOVA.

Estadísticos e-Books & Papers

Cuadro 7.13.: MANOVA para el contraste *post hoc*

```
> summary.manova(fit12)
          Df  Pillai approx F num Df den Df  Pr(>F)
grupo      1 0.84686    11.06      2      4 0.02345 *
Residuals  5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary.manova(fit13)
          Df  Pillai approx F num Df den Df   Pr(>F)
grupo      1 0.84117   15.888      2      6 0.004007 **
Residuals  7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary.manova(fit23)
          Df  Pillai approx F num Df den Df  Pr(>F)
grupo      1 0.71917   6.4021      2      5 0.04179 *
Residuals  6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Aunque ya hemos señalado que la ilustración que estamos utilizando para seguir el funcionamiento interno del MANOVA está basada en pocos casos y los análisis son meramente ilustrativos, apliquemos el procedimiento de Stevens (2009) a la misma. En primer lugar, separamos la base de datos para que solo contengan dos grupos del factor en cada una de ellas:

```
# Generamos datos con pares de factores
datos12<-datos[ which(datos$grupo==1 | datos$grupo==2), ]
datos13<-datos[ which(datos$grupo==1 | datos$grupo==3), ]
datos23<-datos[ which(datos$grupo==2 | datos$grupo==3), ]
```

Encadenamos ahora los tres MANOVA sobre los factores con los pares de grupos. El cuadro 7.13 ofrece los resultados que confirman diferencias significativas en las tres comparaciones.

```
fit12<-manova(cbind(y1,y2)~grupo,data=datos12)
fit13<-manova(cbind(y1,y2)~grupo,data=datos13)
fit23<-manova(cbind(y1,y2)~grupo,data=datos23)
summary.manova(fit12)
summary.manova(fit13)
summary.manova(fit23)
```

Ahora realizamos las pruebas t para ver qué variables y_1 y/o y_2 son las que mayores diferencias muestran en los dos grupos comparados en cada MANOVA

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 7.14.: Pruebas t para el contraste *post hoc*

```
> t.test(y1~grupo,data=datos12)

     Welch Two Sample t-test

data:  y1 by grupo
t = -2.1909, df = 5, p-value = 0.08001
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -4.3466094  0.3466094
sample estimates:
mean in group 1 mean in group 2
              3               5

> t.test(y2~grupo,data=datos12)

     Welch Two Sample t-test

data:  y2 by grupo
t = -4.2426, df = 3.8571, p-value = 0.0143
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -4.992255 -1.007745
sample estimates:
mean in group 1 mean in group 2
              4               7
```

```
t.test(y1~grupo,data=datos12) t.test(y2~grupo,data=datos12)
t.test(y1~grupo,data=datos13) t.test(y2~grupo,data=datos13)
t.test(y1~grupo,data=datos23) t.test(y2~grupo,data=datos23)
```

El cuadro 7.14 muestra los resultados —una selección para el primer MANOVA que comparaba los grupos 1 y 2—. Vemos que, aunque las diferencias del vector eran significativas tomadas como conjunto, es la variable y_2 quien más tiene que ver en esas diferencias que genera el grupo. Este procedimiento, como apunta Stevens (2009), tiene un gran control sobre el error tipo I en la primera etapa (MANOVA) pero mucho menor en la segunda (pruebas t).

Estadísticos e-Books & Papers

7.3. Análisis multivariante de la varianza con dos factores

En el análisis de la varianza con dos factores, a los que denominaremos A y B, el modelo teórico lo expresamos de la siguiente forma:

$$\begin{bmatrix} Y_{1gj} \\ Y_{2gj} \\ \vdots \\ Y_{KgJ} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_K \end{bmatrix} + \begin{bmatrix} \alpha_{1g} \\ \alpha_{2g} \\ \vdots \\ \alpha_{Kg} \end{bmatrix} + \begin{bmatrix} \beta_{1j} \\ \beta_{2j} \\ \vdots \\ \alpha_{Kj} \end{bmatrix} + \begin{bmatrix} (\alpha\beta)_{1gj} \\ (\alpha\beta)_{2gj} \\ \vdots \\ (\alpha\beta)_{Kgj} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1gj} \\ \varepsilon_{2gj} \\ \vdots \\ \varepsilon_{Kgj} \end{bmatrix} \tag{7.23}$$

o, en notación matricial condensada:

$$\mathbf{y_{gj}} = \boldsymbol{\mu_g} + \boldsymbol{\beta_j} + (\boldsymbol{\alpha\beta})_{gj} + \boldsymbol{\varepsilon}_{gj} \tag{7.24}$$

donde, en la formulación anterior:

- μ es la media general.
- α_g es el efecto diferencial del nivel g del factor A.
- β_j es el efecto del nivel j del factor B.
- $(\alpha\beta)_g j$ es el efecto interacción de los niveles g y j.

Para referirse a una variable determinada en el caso de dos factores se requieren tres subíndices. Así, el primer subíndice de Y_{2gj} se refiere a la variable 2 (existen K variables), el segundo subíndice hace referencia al grupo g (existen G grupos del factor A) y el tercer subíndice hace referencia al grupo j (existen J grupos del factor B).

En el análisis multivariante de la varianza con dos factores, se mantienen las hipótesis de poblaciones con distribución normal multivariante e igual matriz de covarianzas para cada combinación de niveles de los dos factores.

En el modelo (7.23) se pueden formular hipótesis, a efectos de realizar contrastes, sobre el factor A, sobre el factor B y sobre la interacción entre ambos factores.

Para determinar si el factor A tiene o no efecto sobre la variable Y, las hipótesis nulas que se formulan son las siguientes:

$$H_0 : \boldsymbol{\alpha}_g = \mathbf{0} \tag{7.25}$$

$$H_0 : \boldsymbol{\beta_j} = \mathbf{0} \tag{7.26}$$

$$H_0 : (\boldsymbol{\alpha\beta})_{gj} = \mathbf{0} \tag{7.27}$$

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 7.15.: Medias muestrales para cada combinación de los factores A y B

Niveles del factor A	Niveles del factor B				Marginal A
	1	2	$\cdots$	J	
1	$\overline{y}_{11}$	$\overline{y}_{12}$	$\cdots$	$\overline{y}_{1J}$	$\overline{y}_{1\bullet}$
2	$\overline{y}_{21}$	$\overline{y}_{22}$	$\cdots$	$\overline{y}_{2J}$	$\overline{y}_{2\bullet}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
G	$\overline{y}_{G1}$	$\overline{y}_{G2}$	$\cdots$	$\overline{y}_{GJ}$	$\overline{y}_{G\bullet}$
Marginal B	$\overline{y}_{\bullet 1}$	$\overline{y}_{\bullet 2}$	$\cdots$	$\overline{y}_{\bullet J}$	$\overline{y}$

Es decir, no hay efecto principal del factor A, no lo hay de B y no hay efecto de interacción.

7.3.1. Descomposición de la matriz de covarianzas

Antes de realizar la descomposición de la matriz de covarianzas, en el cuadro 7.15 se ha construido una tabla de doble entrada que recoge el vector de medias muestrales para cada combinación de los factores A y B. En la última columna se indican los vectores de medias muestrales del factor A y en la última fila los vectores de medias muestrales del factor B. El vector de medias para el conjunto de la muestra, denominado $\overline{\mathbf{y}}$, aparece en la celda de la esquina inferior derecha.

La descomposición de la matriz SCPC total, o matriz $\mathbf{T}$, correspondiente al análisis multivariante de la varianza con dos factores se puede presentar de la siguiente forma:

$$\boldsymbol{T} = \boldsymbol{F_A} + \boldsymbol{F_B} + \boldsymbol{F_{A\times B}} + \boldsymbol{W} \tag{7.28}$$

7.3.2. Cálculo del estadístico de contraste

Antes de definir el estadístico de Wilks correspondiente al caso de dos factores, vamos a presentar el estadístico de Wilks para un factor de una forma alternativa. Teniendo en cuenta (7.7), entonces se verifica que:

$$\Lambda = \frac{|\boldsymbol{W}|}{|\boldsymbol{F} + \boldsymbol{W}|} \tag{7.29}$$

En el denominador de (7.29) aparece la matriz residual intragrupos ($\boldsymbol{W}$) y la matriz correspondiente al factor ($\boldsymbol{F}$). A esta última se la denomina también matriz SCPC de la *hipótesis* debido a que se está contrastando la hipótesis de si el factor considerado tiene o no influencia sobre el vector de variables dependientes.

En el caso de dos factores se pueden contrastar tres hipótesis relativas al factor A, al factor B y a la interacción entre ambos. A estas tres hipótesis les corresponden las matrices $\boldsymbol{F_A}$, $\boldsymbol{F_B}$ y $\boldsymbol{F_{AB}}$, respectivamente. Pues bien, en el caso de dos factores, el estadístico de Wilks se define de la siguiente forma:

Estadísticos e-Books & Papers

$$\Lambda = \frac{|\boldsymbol{W}|}{|\boldsymbol{F_H} + \boldsymbol{W}|} \; H = A, B, AB \tag{7.30}$$

Desarrollaremos todos los cálculos y la interpretación de los mismos con el caso 7.1 donde nos enfrentamos a un MANOVA con dos factores.

Caso 7.1. Promoción de dos productos por la cadena Mercanova
La cadena de supermercados Mercanova desea analizar los efectos de una campaña de promoción conjunta de dos productos: la bebida gaseosa Toca-cola y el ron Morenita. En la campaña de promoción, que ha tenido una duración de dos semanas, se ha llevado a cabo en todos los supermercados de la cadena una promoción de ventas basada en el "precio por paquete", es decir, un precio inferior por la compra conjunta a lo que supondría la compra separada de los dos productos. La acción ha venido apoyada por degustaciones en el punto de venta. Atendiendo a las características de las zonas donde están enclavados, se han agrupado los 57 supermercados de la cadena en tres grupos: urbano comercial (UC), urbano residencial (UR) y rural (R). El potencial general de ventas de cada uno de estos grupos es muy similar. La campaña de promoción comenzó a finales de mes. Concretamente, la primera semana fue del lunes 24 al sábado 29 de octubre. La segunda semana estaba situada prácticamente en los primeros días de mes. Los resultados de las ventas, expresadas en decenas de miles de euros, durante la campaña de promoción aparecen en el cuadro 7.16.

Las preguntas que se plantea el director de marketing de Mercanova son las siguientes:

- ¿Han sido análogos los resultados de la campaña en los tres tipos de supermercado?
- ¿Es relevante la distinción entre semana fin de mes y semana principio de mes en la adquisición de productos de esta clase?
- ¿Existe interacción entre semana y tipo de supermercado?

Del examen de los datos del cuadro 7.16 se desprende que las ventas son más elevadas en el tipo de supermercado UC que en los otros dos tipos. También las ventas son más elevadas en la segunda semana que en la primera.

Estimamos ahora el modelo MANOVA, donde la sintaxis es inmediata y solo debe notarse de qué modo indicamos que hay dos factores ahora. Primero cargamos los datos e indicamos qué dos variables son los factores:

```
datos<-data.frame(Datos_7_1_Caso)
datos$semana <- factor(datos$semana)
datos$tipsuper <-
factor(datos$tipsuper, levels=c("UC", "UR", "R"))
```

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 7.16.: Datos sobre las ventas de dos productos de la cadena Mercanova

Semana promoción		Urbano comercial		Urbano residencial		Rural		Totales		Medias	
Semana	Día	Cola	Ron	Cola	Ron	Cola	Ron	Cola	Ron	Cola	Ron
Primera	Lunes	11	5	9	4	6	2	180	96	10	5.33
Primera	Martes	8	7	8	3	5	2				
Primera	Miércoles	10	5	8	4	6	1				
Primera	Jueves	14	6	9	6	7	4				
Primera	Viernes	17	9	12	9	9	4				
Primera	Sábado	18	10	14	10	9	5				
Segunda	Lunes	17	7	8	10	6	2	252	156	14	8.67
Segunda	Martes	19	10	10	8	7	7				
Segunda	Miércoles	22	9	12	6	10	6				
Segunda	Jueves	21	11	10	11	9	4				
Segunda	Viernes	23	14	15	12	10	5				
Segunda	Sábado	24	15	17	13	12	6				
Totales		204	108	132	96	96	48	432	252	–	–
Medias		17	9	11	8	8	4	–	–	12	7

© Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 7.17.: Matrices suma de cuadrados y productos cruzados

```
$semana
     cola ron
cola  144 120
ron   120 100

$tipsuper
     cola ron
cola  504 252
ron   252 168

$`semana:tipsuper`
     cola ron
cola   72  12
ron    12   8

$Residuals
     cola ron
cola  240 143
ron   143 172
```

Solicitamos inmediatamente el MANOVA y pedimos en la salida que se nos proporcionen las matrices con las sumas de los cuadrados como hicimos en el MANOVA de un factor:

```
fit<-manova(cbind(cola,ron)~semana*tipsuper,data=datos)
summary.manova(fit,intercept=TRUE)
summary.manova(fit,intercept=TRUE)$SS
```

En el cuadro 7.17 se ha recogido la salida relativa a los contrastes del análisis de la varianza con los dos factores considerados. La primera información que aparece en este cuadro son las matrices de suma de productos y productos cruzados de los factores, del efecto interacción —matrices que habíamos denominado $\mathbf{F_H}$— y residual, que es, como ya hemos visto anteriormente, la nomenclatura que utiliza R para referirse a la matriz $\mathbf{W}$. Recordemos que esta última es necesaria para el cálculo de la prueba de esfericidad de Barlett. Todos los datos que requiere el cálculo de la Λ de Wilks aparecen en el cuadro 7.17. En efecto, aplicando (7.30) se obtiene de manera ilustrativa para el efecto interacción:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 7.18.: Estimación de la significatividad de los efectos

```
                Df   Wilks approx F num Df den Df    Pr(>F)
(Intercept)      1 0.04423  313.335      2     29 < 2.2e-16 ***
semana           1 0.59046   10.057      2     29 0.0004812 ***
tipsuper         2 0.21490   16.779      4     58 3.438e-09 ***
semana:tipsuper  2 0.64823    3.510      4     58 0.0123884 *
Residuals       30
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$\Lambda = \frac{|\boldsymbol{W}|}{|\boldsymbol{F}_{\boldsymbol{A}\times\boldsymbol{B}} + \boldsymbol{W}|} = \frac{\begin{vmatrix} 240 & 143 \\ 143 & 172 \end{vmatrix}}{\left|\begin{bmatrix} 72 & 12 \\ 12 & 8 \end{bmatrix} + \begin{bmatrix} 240 & 143 \\ 143 & 172 \end{bmatrix}\right|} = \frac{20831}{32135} = 0{,}64823$$

Valor que coincide con la salida del cuadro 7.18, donde se recoge la significatividad de todos los efectos. El nivel de significación crítico para este contraste es 0,012, lo que implica que se puede rechazar la hipótesis nula para un nivel de significación del 5 %, pero no para un nivel del 1 %. Existe pues una evidencia débil de interacción entre el tipo de supermercado y la semana.

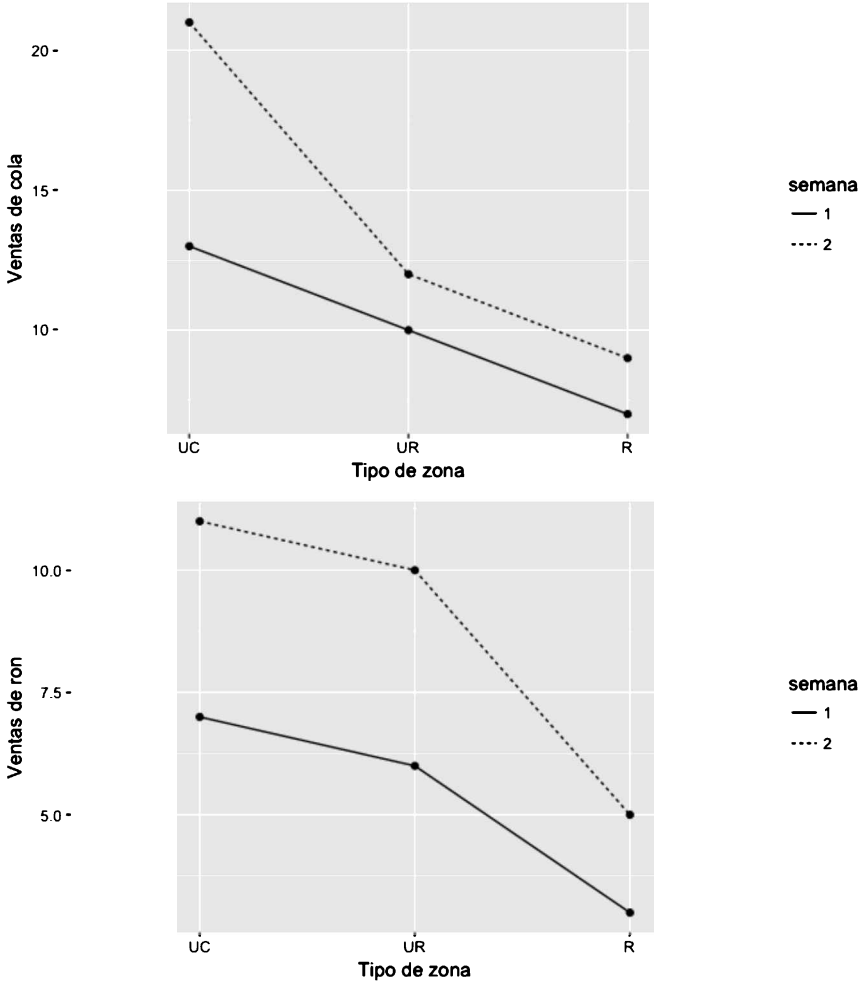
Para tratar de localizar el efecto de interacción, en la figura 7.1, se ha reflejado el gráfico de interacción entre tipo de supermercado y semana para cada uno de los dos productos. En lo que respecta al ron Morenita, el efecto de interacción es prácticamente inexistente, ya que la línea de ventas medias de la primera semana es paralela a la línea de ventas medias de la segunda semana. También sucede lo mismo en el producto Toca-cola para los tipos de supermercado UR y R. Sin embargo, se observa que la segunda semana (semana-principio-de-mes) va asociada a unas ventas más elevadas de lo normal en el tipo de establecimientos UC.

Los efectos principales del tipo de establecimiento ($\Lambda = 0{,}21490, F = 16{,}779$; $p < 0{,}01$) y de la semana de la promoción ($\Lambda = 0{,}59046$, $F = 10{,}057; p < 0{,}01$) son claramente significativos. Con la ayuda de la figura 7.1, podemos concluir que en los entornos rurales las ventas han sido inferiores a los entornos urbanos para ambos tipos de bebida y que en la segunda semana las ventas también han sido superiores. La caída de ventas solo ha sido más intensa cuando se ha pasado en la segunda semana de centros en zonas urbanas comerciales a zonas urbanas residenciales (efecto interacción).

Estos resultados, sin embargo, son preliminares hasta que no se verifiquen las hipótesis en las que se basa el MANOVA. La hipótesis de normalidad multivariante la comprobamos con el test multivariante de Shapiro, para lo que, en primer lugar, es necesario segregar la base de datos en cada uno de los grupos (2 grupos para semana y 3 para tipo de superficie):

Estadísticos e-Books & Papers

Figura 7.1.: Ventas de ron y cola por semana y tipo de zona



© Ediciones Paraninfo

Estadísticos e-Books & Papers

```
#desagregamos la base por grupos por grupos: tipo de tienda
grupoUC<-datos[1:12,3:4]
grupoUR<-datos[13:24,3:4]
grupoR<-datos[25:36,3:4]
#El test necesita las variables en filas, transponemos
grupoUC<-t(grupoUC)
grupoUR<-t(grupoUR)
grupoR<-t(grupoR)

#desagregamos la base por grupos por grupos: Semana
#ordenamos la base por semana
datos<-datos[order(datos$semana),]
grupoSem1<-datos[1:18,3:4]
grupoSem2<-datos[19:36,3:4]
#El test necesita las variables en filas, transponemos
grupoSem1<-t(grupoSem1)
grupoSem2<-t(grupoSem2)
```

A continuación aplicamos el test multivariante de Shapiro, cuyos resultados se ofrecen en el cuadro 7.19, que muestra que no se puede rechazar la hipótesis de normalidad multivariante de las distribuciones de las variables con las ventas de cola y ron para ninguno de los grupos, salvo para las ventas en tiendas ubicadas en entornos rurales, variables para las que podría realizarse alguna transformación como las ilustradas en el tema 2.

```
mshapiro.test(grupoUC)
mshapiro.test(grupoUR)
mshapiro.test(grupoR)
mshapiro.test(grupoSem1)
mshapiro.test(grupoSem2)
```

En cuanto a la igualdad de las matrices de covarianzas, no hay implementado en R un procedimiento para aplicar el test M de Box simultáneamente en los 6 grupos que genera el diseño experimental, por ello la mejor aproximación es realizar el test que presentamos en el caso de un factor para los grupos generados por un factor y posteriormente por el otro. El cuadro 7.20 muestra como en ambos casos no puede descartarse la igualdad de las matrices de covarianzas para $p < 0{,}01$.

```
library(biotools)
boxM(datos[3:4],datos[,1])
boxM(datos[3:4],datos[,2])
```

El último paso es asegurarnos de que la matriz de correlaciones no es la identidad, porque tener variables dependientes incorrelacionadas nos llevaría a

Estadísticos e-Books & Papers

Cuadro 7.19.: Test de normalidad multivariante de Shapiro

```
> mshapiro.test(grupoUC)

     Shapiro-Wilk normality test

data:  Z
W = 0.91279, p-value = 0.2316

> mshapiro.test(grupoUR)

     Shapiro-Wilk normality test

data:  Z
W = 0.88022, p-value = 0.08822

> mshapiro.test(grupoR)

     Shapiro-Wilk normality test

data:  Z
W = 0.78631, p-value = 0.006574

> mshapiro.test(grupoSem1)

     Shapiro-Wilk normality test

data:  Z
W = 0.96142, p-value = 0.6294

> mshapiro.test(grupoSem2)

     Shapiro-Wilk normality test

data:  Z
W = 0.94767, p-value = 0.3896
```


© Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 7.20.: Test M de Box de igualdad de las matrices de covarianzas

```
> boxM(datos[3:4],datos[,1])

     Box's M-test for Homogeneity of Covariance Matrices

data:  datos[3:4]
Chi-Sq (approx.) = 10.094, df = 3, p-value = 0.01779

> boxM(datos[3:4],datos[,2])

     Box's M-test for Homogeneity of Covariance Matrices

data:  datos[3:4]
Chi-Sq (approx.) = 13.049, df = 6, p-value = 0.04226
```

que lo más sensato sería encadenar ANOVA de dos factores. Es improbable que esto ocurra en la medida en que se ha detectado un efecto de interacción, pero hay que aplicar el test de esfericidad de Barlett sobre la matriz de residuos. El test de Barlett, como vimos para el caso del MANOVA de un factor, se aplica sobre la matriz de correlaciones, mientras que el cuadro 7.17 nos ofrece la matriz con los cuadrados y productos cruzados. Recordemos el procedimiento para obtener a partir de ella la matriz de covarianzas y luego la matriz de correlaciones y poder aplicar el test:

```
W = matrix(c(240,143,143,172), nrow=2,ncol=2)
COV.W=W/30 # df=30 para los residuos
R<-cov2cor(COV.W)
cortest.bartlett(R,n=30)
```

Recordemos que para obtener la matriz de covarianzas basta dividir por el número de grados de libertad de los residuos en el proceso de estimación del MANOVA que, como se observa en el cuadro 7.18, es $df = 30$. La matriz de covarianzas se convierte en matriz de correlaciones mediante la función `cov2cor{stats}` y ya puede aplicarse el test. El cuadro 7.20 confirma que puede rechazarse la hipótesis nula de ausencia de correlaciones entre las variables del MANOVA, por lo que las conclusiones obtenidas pueden considerarse definitivas.

Estadísticos e-Books & Papers

Cuadro 7.21.: Test de esfericidad de Barlett

```
> cortest.bartlett(R,n=30)
$chisq
[1] 18.80824

$p.value
[1] 1.445416e-05

$df
[1] 1
```


© Ediciones Paraninfo

Estadísticos e-Books & Papers

8. Regresión lineal múltiple

8.1. Introducción

Este capítulo está dedicado al modelo de regresión múltiple. Este modelo es uno de los más conocidos y aplicados del análisis multivariante. Por otra parte, constituye el núcleo a partir del cual se ha desarrollado la econometría. El modelo de regresión lineal múltiple se aplica tanto a datos de corte transversal (es decir, a observaciones referidas a un mismo momento del tiempo, como pueden datos de encuestas a familias, empresas, etc.) como a datos de series temporales. En el contexto del análisis multivariante los datos que se utilizan suelen ser predominantemente de corte transversal. Por ello, en la exposición de este tema vamos a prescindir de las implicaciones derivadas de la utilización de datos de series temporales. Al lector que esté interesado en estos aspectos le remitimos a los manuales de econometría y/o series temporales.

La estructura del capítulo es la siguiente. En el epígrafe 8.2 se formula el modelo de regresión simple y se procede a su estimación utilizando el método de mínimos cuadrados. El epígrafe 8.3 tiene el mismo contenido que el epígrafe 8.2 pero referido al modelo de regresión múltiple. Los epígrafes 8.4 y 8.5 están dedicados, respectivamente, a la formulación de las hipótesis básicas y a la evaluación de la bondad de ajuste del modelo. El epígrafe 8.6 está dedicado a presentar la comprobación de las hipótesis básicas en que se sustenta el modelo de regresión lineal múltiple. El epígrafe 8.7 se centra en analizar las particularidades que supone el uso de variables ficticias en estos modelos.

8.2. El modelo de regresión lineal simple y su estimación por mínimos cuadrados

En el modelo de regresión simple se trata de analizar el comportamiento de una variable a la que denominaremos dependiente como función lineal de una variable explicativa. El *modelo de regresión poblacional*, o teórico, se puede representar de la siguiente forma:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{8.1}$$

La variable y designa a la variable dependiente o endógena, mientras que x designa a la variable explicativa, exógena o independiente. En (8.1) se ha introducido la variable ε para recoger todos aquellos factores distintos de x que afectan a y. Es denominada error o perturbación aleatoria. La perturbación

Estadísticos e-Books & Papers

Cuadro 8.1.: Datos simulados para el caso 8.1

	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
	1	1,5	-4,5	-4,2	18,9	20,3
	2	2,5	-4,5	-3,2	11,2	12,3
	3	2,5	-2,5	-3,2	8,0	6,3
	4	3,9	-1,5	-1,8	2,7	2,3
	5	6,1	-0,5	0,4	-0,2	0,3
	6	7,1	0,5	1,4	0,7	0,3
	7	6,1	1,5	0,4	0,6	2,3
	8	7,5	2,5	1,8	4,5	6,3
	9	9,5	3,5	3,8	13,3	12,3
	10	10,2	4,5	4,5	20,3	20,3
Suma	55,0	56,9	0,0	0,0	80,0	82,5
Media	5,5	5,7	0,0	0,0	8,0	8,3

es una variable no observable. Finalmente, β_0 y β_1 son los parámetros del modelo que son desconocidos. El objetivo principal del modelo de regresión es la determinación o estimación de β_0 y β_1 a partir de una muestra dada. La *función de regresión muestral* o modelo ajustado es la contrapartida del modelo poblacional y se expresa de la siguiente forma:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{8.2}$$

y permite calcular el valor ajustado $\hat{y}_i$ para y cuando $x = x_i$. En (8.2) $\hat{\beta}_0$ y $\hat{\beta}_1$ son los estimadores de los parámetros β_0 y β_1. Por tanto, para cada x_i tenemos un valor observado y_i y un valor ajustado $\hat{y}_i$. A la diferencia entre y_i y $\hat{y}_i$ se la denomina residuo $\hat{\varepsilon}_i$:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \tag{8.3}$$

Caso 8.1. La estimación de la función de consumo

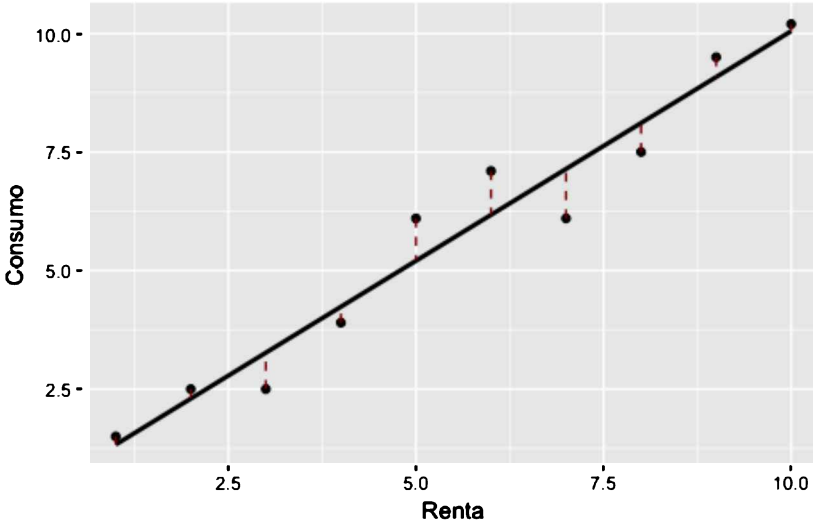
Para ilustrar el desarrollo del proceso de estimación de una regresión lineal, así como los elementos fundamentales que componen una regresión, lo ilustraremos, como siempre, con un ejemplo muy sencillo para luego extenderlo a un caso más general. Comenzaremos por una regresión simple, esto es, aquella en la que solo contamos con una variable dependiente y un regresor. En este primer caso vamos a ilustrar la estimación de la función de consumo keynesiana utilizando unos hipotéticos datos de 10 hogares, que aparecen en el cuadro 8.1. En esta función el consumo en términos reales se explica en función de la renta disponible también expresada en términos reales. La figura 8.1 ilustra gráficamente los datos.

En otras palabras, el residuo $\hat{\varepsilon}_1$ es la diferencia entre el valor muestral y_i y el valor ajustado de $\hat{y}_i$, según hemos ilustrado en la figura 8.1 con la línea discontinua. Para cada observación muestral puede hacerse la siguiente des-

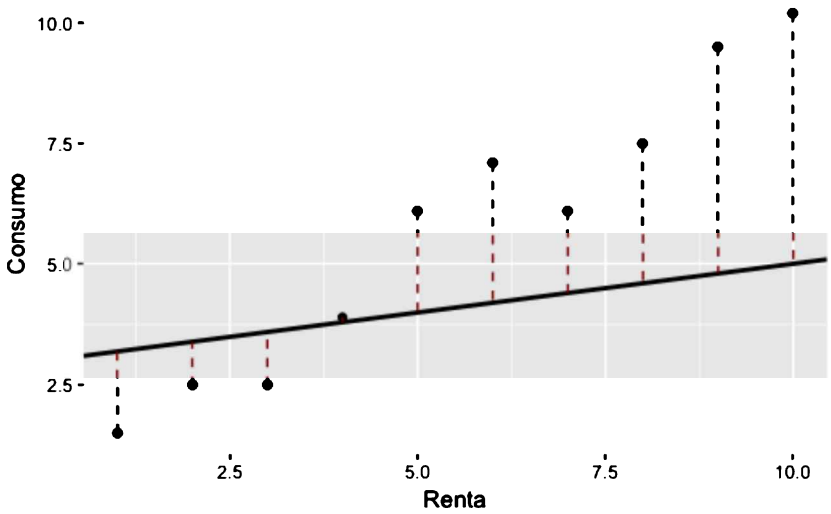
 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 8.1.: Ejemplo de regresión lineal simple



(a) Regresión lineal simple. Función ajustada



(b) Función alternativa $y = 3 + 0,2x$

Estadísticos e-Books & Papers

composición:

$$y_i = \hat{y}_i + \hat{\varepsilon}_i$$

El problema que tenemos que resolver ahora es la obtención de los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$. A la vista de la figura 8.1 se trataría de seleccionar la recta de regresión de forma que los residuos en su conjunto sean lo más pequeños posible. Pero como se observa, existen muchas rectas (panel b) que pueden utilizarse. En este sentido, para determinar la recta adecuada, el método más adecuado es el de **mínimos cuadrados (MC)**, que consiste en elegir $\hat{\beta}_0$ y $\hat{\beta}_1$ de forma que minimice la suma de los cuadrados de los residuos (SSR). Para ello, en primer lugar, vamos a expresar SSR como una función de los estimadores, utilizando (8.3):

$$\text{mín } SSR = \text{mín} \sum_{i=1}^{n} (\hat{\varepsilon}_i)^2 = \text{mín} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \text{mín} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{8.4}$$

Para minimizar SSR, derivamos parcialmente con respecto a $\hat{\beta}_0$ y $\hat{\beta}_1$ e igualando a 0 se obtiene el siguiente sistema de ecuaciones que se denominan *ecuaciones normales* o *condiciones de primer orden de MC*:

$$\begin{aligned} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) &= 0 \\ \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) x_i &= 0 \end{aligned} \tag{8.5}$$

Resolviendo este sistema de ecuaciones se obtienen los estimadores para $\hat{\beta}_0$ y $\hat{\beta}_1$, que vienen dados por:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned} \tag{8.6}$$

El cuadro 8.1 ofrece los cálculos necesarios para que obtengamos que:

$$\begin{aligned} \hat{\beta}_1 &= \frac{8{,}0}{8{,}3} = 0{,}97 \\ \hat{\beta}_0 &= 5{,}7 - 0{,}97 \times 5{,}5 = 0{,}36 \end{aligned}$$

siendo, por tanto, la recta estimada y representada en el panel (a) de la figura 8.1:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

$$y = 0{,}36 + 0{,}97x \tag{8.7}$$

donde y es el consumo, y x, la renta de los hogares. Lo relevante es la **interpretación de los coeficientes de regresión en el modelo de regresión lineal simple.** En el modelo (8.7), el coeficiente $\hat{\beta}_1$ mide el incremento que se producirá en la variable y (en las unidades en que esté medida la variable y) al incrementarse en una unidad la variable x (en las unidades en que esté medida la variable x).

El coeficiente $\hat{\beta}_0$ es el valor que predice el modelo $\hat{y}$ cuando x toma el valor 0. En muchas ocasiones este término no tiene un significado económico claro.

En el modelo lineal de regresión simple, el coeficiente $\hat{\beta}_1$ es la derivada de la variable endógena estimada con respecto a la variable explicativa. En el caso concreto del modelo (8.7), el coeficiente $\hat{\beta}_1$ es precisamente la propensión marginal al consumo:

$$\frac{\partial y}{\partial x} = \frac{\partial cons_i}{\partial renta_i} = \hat{\beta}_1 \tag{8.8}$$

Evidentemente no vamos a estimar la regresión realizando siempre los cálculos de manera manual. Utilizaremos a lo largo del tema la función `lm{stats}`, cuya sintaxis es elemental, basta indicarle los datos y el modelo que queremos estimar. El cuadro 8.2 muestra la salida, de la cual, de momento, solo nos interesa comprobar como los parámetros estimados coinciden con los que hemos deducido. En el resto de la información profundizaremos inmediatamente.

```
fit<-lm(data=datos,y~x)
```

Caso 8.2. Cantidad de café vendido como una función de su precio

Dado que los datos del caso 8.1 eran ficticios para facilitar la didáctica del desarrollo del modelo y el procedimiento de estimación, veamos un caso con datos reales para asegurarnos la interpretación correcta de los parámetros del modelo. En el trabajo de Bemmaor y Mouchoux (1991) se formula el siguiente modelo para explicar la cantidad de café vendido por semana (*cafeqt*) en función del precio del café (*cafepr*).

$$cafeqt = \beta_0 + \beta_1 cafepr + \varepsilon$$

La variable *cafepr* toma el valor 1, el precio habitual, pero se realizan distintas promociones de ventas que bajan el precio del café a 0,95 y 0,85 en dos acciones para analizar el efecto de la promoción de ventas sobre el consumo. El experimento duró 12 semanas, *cafeqt* está expresado en miles de unidades y *cafepr* en FF (francos franceses). Los datos aparecen en el cuadro 8.3. La estimación se realiza del siguiente modo y los resultados aparecen en el cuadro 8.4.

Estadísticos e-Books & Papers

Cuadro 8.2.: Parámetros estimados para la regresión lineal

```
Call:
lm(formula = y ~ x, data = datos)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0436 -0.5436  0.1600  0.3641  0.9254

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.36000    0.48846   0.737    0.482
x            0.96909    0.07872  12.310 1.76e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.715 on 8 degrees of freedom
Multiple R-squared:  0.9499,    Adjusted R-squared:  0.9436
F-statistic: 151.5 on 1 and 8 DF,  p-value: 1.765e-06
```

```
fit<-lm(data=datos,cafeqt~cafepr)
```

El modelo ajustado es el siguiente:

$$cafeqt = 774,9 - 693,33cafepr \tag{8.9}$$

Interpretación del coeficiente $\hat{\beta}_1$: si el precio del café se incrementa en 1 FF, la cantidad vendida de café se reducirá en 693,33 miles de unidades. En la medida que el precio del café es una magnitud pequeña, es preferible dar la siguiente interpretación: si aumenta el precio del café en 1 céntimo de FF, la cantidad vendida de café se reducirá en 6,93 miles de unidades.

8.3. El modelo de regresión lineal múltiple y su estimación por mínimos cuadrados

El modelo de regresión lineal simple no es adecuado para modelizar muchos fenómenos económicos, ya que para explicar una variable económica se requiere en general tener en cuenta más de un factor. Veamos algunos ejemplos.

En la función keynesiana clásica el consumo se hace depender, como en (8.7), de la renta disponible como única variable relevante. Sin embargo, hay otros factores que pueden considerarse relevantes en el comportamiento del consumidor. Uno de esos factores podría ser la riqueza. Con la inclusión de este factor se tendrá un modelo con dos variables explicativas:

$$cons = \beta_0 + \beta_1 inc + \beta_2 riqueza + \varepsilon \tag{8.10}$$

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 8.3.: Datos sobre cantidades y precios del café

semana	cafepr	cafeqt
1	1,00	89
2	1,00	86
3	1,00	74
4	1,00	79
5	1,00	68
6	1,00	84
7	0,95	139
8	0,95	122
9	0,95	102
10	0,85	186
11	0,85	179
12	0,85	187

Cuadro 8.4.: Parámetros estimados para el caso 8.2 y análisis de la significatividad global

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   774.92      47.67   16.26 1.61e-08 ***
cafepr       -693.33      50.07  -13.85 7.52e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 10.62 on 10 degrees of freedom
Multiple R-squared:  0.9504, Adjusted R-squared:  0.9455
F-statistic: 191.7 on 1 and 10 DF,  p-value: 7.523e-08
```


Estadísticos e-Books & Papers

Los salarios se determinan por diferentes factores. Un modelo relativamente simple para explicar los salarios en función de los años de educación y de los años de experiencia es el siguiente:

$$salarios = \beta_0 + \beta_1 educ + \beta_2 exper + \varepsilon \tag{8.11}$$

De todos modos, otros factores importantes para explicar los salarios pueden ser variables cuantitativas tales como el tiempo de formación y la edad, o variables cualitativas, como el sexo, la rama de actividad, etc.

En el modelo de regresión múltiple que vamos a presentar se considera que la variable dependiente es una función lineal de k regresores y de un término de error. Designando por y_i al regresando, por $x_{1i}, x_{2i}, \cdots, x_{ki}$ a los regresores y por ε_i a la perturbación aleatoria (error), el modelo teórico de regresión lineal viene dado, para la observación genérica i-ésima, por la siguiente expresión:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i \qquad i = 1, 2, \cdots, n \tag{8.12}$$

siendo n el tamaño muestral. Para cada una de las observaciones, el sistema de ecuaciones generado sería:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{k1} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{k2} + \varepsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{kn} + \varepsilon_n \end{aligned} \tag{8.13}$$

que puede expresarse de manera matricial definiendo las matrices de la siguiente forma:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{k1} \\ 1 & x_{21} & x_{22} & \cdots & x_{k2} \\ 1 & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{8.14}$$

El modelo de regresión lineal (8.12) múltiple expresado en notación matricial es el siguiente:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{8.15}$$

y si tenemos en cuenta las denominaciones dadas a vectores y matrices, el modelo de regresión lineal múltiple se puede expresar de forma compacta de la siguiente forma:

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon} \tag{8.16}$$

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

donde $\mathbf{y}$ es un vector $n \times 1$ que contiene las puntuaciones de cada uno de los n casos en la variable, dependiente, $\mathbf{X}$ es una matriz $N \times (k+1)$ que contiene los valores de cada uno de los n casos en las k variables y $\boldsymbol{\beta}$ es un vector $(k+1) \times 1$ con los k coeficientes de regresión más el intercepto (punto de corte con el eje de abscisas en una regresión simple).

Si el procedimiento aplicado lo generalizamos a la regresión múltiple, es decir, con k regresores en lugar de uno, el desarrollo para (8.12) en lugar de para (8.1) sería el siguiente. El modelo ajustado sería:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \tag{8.17}$$

El vector con los residuos sería la diferencia entre el vector de valores observados y los ajustados:

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \tag{8.18}$$

Y con todo ello, la suma de los cuadrados de los residuos que habría que minimizar sería:

$$SSR = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \begin{bmatrix} \hat{\varepsilon}_1 & \hat{\varepsilon}_2 & \cdots & \hat{\varepsilon}_n \end{bmatrix} \begin{bmatrix} \hat{\varepsilon}_1 \\ \hat{\varepsilon}_2 \\ \vdots \\ \hat{\varepsilon}_n \end{bmatrix} = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \tag{8.19}$$

y teniendo en cuenta (8.18), se obtiene:

$$SSR = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{8.20}$$

Minimizar implica derivar (8.20) con respecto al vector de coeficientes $\hat{\boldsymbol{\beta}}$ e igualar a 0. Si lo hiciéramos, obtendríamos el siguiente sistema de ecuaciones:

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \tag{8.21}$$

Al sistema anterior se le denomina genéricamente **sistema de ecuaciones normales del hiperplano**. Cuando $k = 2$, se obtiene el sistema de ecuaciones normales de la recta; cuando $k = 3$, se obtiene el sistema de ecuaciones normales del plano; finalmente, cuando $k > 3$, se obtiene específicamente el sistema de ecuaciones normales del hiperplano.

Para poder resolver el sistema (8.21) respecto a $\hat{\boldsymbol{\beta}}$ unívocamente se debe cumplir que el rango de la matriz $\mathbf{X}'\mathbf{X}$ sea igual a k. Si se cumple esta condición, se pueden premultiplicar ambos miembros de (8.21) por $[\mathbf{X}'\mathbf{X}]^{-1}$:

$$[\mathbf{X}'\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{X}]\hat{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{y} \tag{8.22}$$

con lo cual se obtiene la expresión del vector de estimadores mínimo - cuadráticos:

Estadísticos e-Books & Papers

$$\hat{\boldsymbol{\beta}} = [\mathbf{X'X}]^{-1}\,\mathbf{X'y} \tag{8.23}$$

que sería la expresión equivalente a (8.6) que obtuvimos para la regresión simple. Tan es así que podemos reproducir los resultados simplemente multiplicando las matrices en R:

```
library(MASS)
y=c(1.5,2.5,2.5,3.9,6.1,7.1,6.1,7.5,9.5,10.2)
x=matrix(c(1,1,1,1,1,1,1,1,1,1,1,2,3,4,5,6,7,8,9,10),10,2)
beta<-ginv(t(x)%*%x)%*%t(x)%*%y
beta

##           [,1]
## [1,] 0.3600000
## [2,] 0.9690909
```

Como hicimos con la regresión lineal simple, veamos cuál sería la **interpretación de los coeficientes** en el caso de la regresión lineal múltiple.

El coeficiente β_j mide el efecto parcial del regresor x_i, manteniendo los otros regresores fijos. Vamos a ver el significado de esta expresión. El modelo estimado para la observación i-ésima viene dado por:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ji} + \cdots + \beta_k x_{ki} + \varepsilon_i \tag{8.24}$$

Consideremos ahora el modelo estimado para la observación h-ésima, en el que los valores de las variables explicativas y, en consecuencia, de y habrán cambiado con respecto a la (8.24):

$$y_h = \beta_0 + \beta_1 x_{1h} + \beta_2 x_{2h} + \cdots + \beta_k x_{jh} + \cdots + \beta_k x_{kh} + \varepsilon_h \tag{8.25}$$

Restando (8.25) de (8.24) tenemos que:

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2 + \cdots + \hat{\beta}_k \Delta x_k \tag{8.26}$$

donde $\Delta\hat{y} = \hat{y}_i - \hat{y}_h$; $\Delta\hat{x}_k = \hat{x}_{ki} - \hat{x}_{kh}$. La expresión anterior capta la variación de $\hat{y}$ debida a cambios en todos los regresores. Si solo se cambia x_j, tendremos que:

$$\Delta\hat{y} = \hat{\beta}_j \Delta x_j \tag{8.27}$$

y si x_j se incrementa en una unidad, tenemos

$$\Delta\hat{y} = \hat{\beta}_j \tag{8.28}$$

En consecuencia, el coeficiente β_j mide el cambio en y cuando x_j aumenta en una unidad, *manteniendo fijos los regresores* $x_1, x_2, \cdots, x_{j-1}, x_{j+1}, \cdots, x_k$. Es muy importante en la interpretación de los coeficientes tener en cuenta esta

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 8.5.: Parámetros estimados para el caso 8.3

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.41330    1.60303   8.991 1.60e-11 ***
edad        -0.09598    0.04785  -2.006    0.051 .
antigue     -0.07757    0.06720  -1.154    0.255
salario     -0.03645    0.00734  -4.966 1.08e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 2.164 on 44 degrees of freedom
Multiple R-squared:  0.6942, Adjusted R-squared:  0.6733
F-statistic: 33.29 on 3 and 44 DF,  p-value: 2.169e-11
```

claúsula *ceteris paribus.* Esta interpretación no es válida, lógicamente, para el término independiente.

Caso 8.3. Cuantificando la influencia de la edad y del salario sobre el absentismo en la empresa Buenosaires

Buenosaires es una empresa dedicada a la fabricación de ventiladores, habiendo tenido resultados relativamente aceptables en los últimos años. Los directivos consideran, sin embargo, que los resultados habrían sido mejores si el absentismo en la empresa no fuera tan alto. Para explicar las razones del absentismo, el modelo que se propone es el siguiente:

$$absen = \beta_0 + \beta_1 edad + \beta_2 antigue + \beta_3 salario + \varepsilon \tag{8.29}$$

donde la ausencia en el trabajo, *absent*, se mide en días por año; el *salario,* en miles de euros al año; los años trabajados en la empresa, *antigue*, y la *edad* se expresan en años.

Estimando el modelo con la siguiente sintaxis, los resultados aparecen en el cuadro 8.5.

```
fit<-lm(data=datos,absen~edad+antigue+salario)
```

Utilizando una muestra de tamaño 48, se ha estimado la siguiente ecuación con los datos del cuadro 8.5:

$$absen = 14,413 - 0,096 \times edad - 0,078 \times antigue - 0,036 \times salario \tag{8.30}$$

La interpretación de β_1 es la siguiente: manteniendo fijo el salario y los años trabajados en la empresa, si la edad se incrementa en un año, el absentismo laboral se reducirá en 0,096 días al año. La interpretación de β_2 es como sigue:

Estadísticos e-Books & Papers

manteniendo fijo el salario y la edad, si los años trabajados en la empresa se incrementan en un año, el absentismo laboral se reducirá en 0,078 días al año. Finalmente, la interpretación de β_3 es la siguiente: manteniendo fija la edad y los años trabajados en la empresa, si el salario se incrementa en 1000 euros al año, el absentismo laboral se reducirá en 0,036 días por año.

8.4. Contraste de hipótesis

Estimado el modelo, esto es, obtenidos los parámetros β_i que lo definen, el paso siguiente es establecer la significatividad de dichos parámetros. Como veremos inmediatamente, eso significa calcular un estadístico que ha de ser capaz de evaluar cuánto del total de la información que contienen las variables originales es capaz de recoger el modelo estimado y cuánta información ha quedado fuera. Ese estadístico será el estadístico F, pero lo que es crucial es entender su lógica más que simplemente interpretar sus resultados.

8.4.1. Contraste para el conjunto de parámetros

Como veíamos en el apartado anterior, nuestro modelo se ha estimado de tal forma que minimiza la suma de los cuadrados de los residuos SSR. La pregunta es si los residuos resultantes suponen una parte más o menos importante de la información original. Field (2005) ofrece un razonamiento muy sencillo e intuitivo para explicar la construcción del estadístico F. Preguntémonos en primer lugar cuál es esa información total que contiene el modelo. Si tuviéramos que predecir el valor de las ventas de una nueva empresa que no está en nuestra base de datos y no contáramos con ninguna información adicional (es decir, no tuviéramos información sobre la inversión en publicidad), ¿cuál sería la estimación más sensata? Lo lógico sería asignarle la media del gasto en publicidad de las empresas de nuestro modelo, es decir, 5,7 (cuadro 8.1). Si utilizáramos esa recta ($y = 5{,}7$) como recta de regresión de nuestro modelo en lugar de (8.7) ¿cuál habría sido el error que hubiéramos cometido? Pues el que se ilustra en el panel (a) de la figura 8.2, la suma de los cuadrados de la diferencia entre el valor original y la distancia a la recta ($y = 5{,}7$). Como esta estimación es la más sensata en ausencia de información, esa suma de cuadrados de residuos para este modelo *naïf* pero racional se puede considerar el *benchmark* que cualquier otro modelo debería mejorar. A ese valor le llamaremos suma de cuadrados totales o SST:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{8.31}$$

¿Qué explica entonces el modelo? Lógicamente el total menos el residuo, es decir, la distancia que hay del valor original a la media $(y_i - \bar{y})$ (total) menos la parte que va del valor original a la recta de regresión $(y_i - \hat{y}_i)$ (residuo), dicho de otra forma (cuadrados aparte):

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

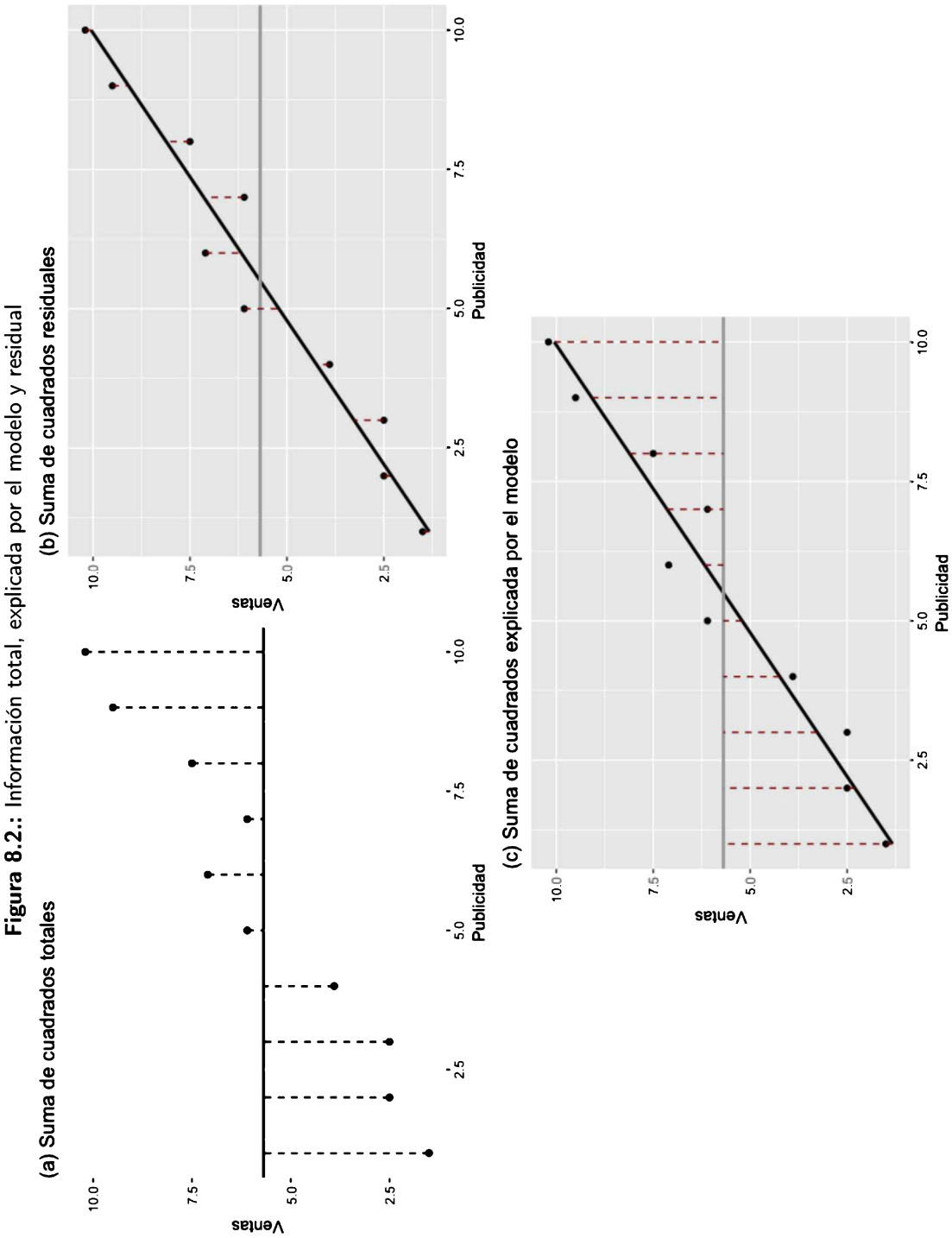


Figura 8.2.: Información total, explicada por el modelo y residual

(a) Suma de cuadrados totales

(b) Suma de cuadrados residuales

(c) Suma de cuadrados explicada por el modelo

Estadísticos e-Books & Papers

$$(y_i - \bar{y}) - (y_{i-}\hat{y}_i) = (\hat{y}_i - \bar{y})$$

o sea, la distancia entre el valor estimado y la recta que representa la media. De una manera algo más formal, incorporando los cuadrados:

$$\begin{aligned} SST &= \sum_{i=1}^{n}(y_i - \bar{y})^2 \\ SSR &= \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \\ SSM &= \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 \end{aligned} \tag{8.32}$$

debiendo cumplirse, lógicamente, que:

$$SST = SSM + SSR \tag{8.33}$$

Bien, veamos cómo podemos construir un estadístico que se comporte de una manera razonable. Esa "manera razonable" es que cuanto más explique el modelo del total, más grande debería ser el estadístico para ayudarnos a rechazar la hipótesis nula de que nuestro modelo no aporta nada. Hay, básicamente, dos formas de hacerlo: dividiendo lo que explica el modelo sobre el total o, para hacer todavía más sensible la ratio:

$$F = \frac{SSM}{SSR} \tag{8.34}$$

¿Por qué es más sensible un estadístico así definido? Porque, como vemos en la ecuación (8.33), si la suma del modelo y los residuos es constante, cuánto mayor sea el modelo (numerador), más pequeño será el residuo (denominador) y ambas cosas contribuyen a hacer más grande F. Hay, sin embargo, un sesgo. Para construir las sumas de cuadrados hemos utilizado un número diferente de elementos (diferencias). Remitimos al lector al apartado 6.2.1 donde ilustramos el concepto de grados de libertad, pero es bastante intuitivo que para los cuadrados totales solo hemos estimado un parámetro, la media global $\bar{y}$, con lo que nos quedarán $n-1$ grados de libertad, mientras que para estimar los cuadrados de los residuos, además del ya consumido para los cuadrados totales, hemos tenido que estimar el coeficiente de regresión β_1 (si tuviéramos k variables explicativas, hubiéramos tenido que estimar k de estos parámetros, por lo tanto, sus grados de libertad son $n-1-k$). Por lo tanto, por diferencia, los grados de libertad de la regresión serán, teniendo en cuenta (8.33), $(n-1)-(n-1-k) = k$.

Para corregir el sesgo del estadístico F si lo definiéramos con la expresión (8.34) de tener los cuadrados construidos con diferente número de elementos, dividimos la suma de cuadrados por los grados de libertad, lo que denominamos medias cuadráticas y, ahora sí, el estadístico F que utilizaremos será:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

$$F = \frac{MCM}{MCR} = \frac{SCM/k}{SCR/n-k-1} = \frac{\sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2 / k}{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 / (n-k-1)} \tag{8.35}$$

estadístico que se distribuye como una F con k grados de libertad en el numerador y $n-k-1$ en el denominador.

Tenemos un estadístico, pero no tenemos una hipótesis nula que contrastar. Decíamos anteriormente que lo hemos de utilizar para evaluar en qué medida nuestro modelo aporta valor. Ese valor ha de ser que la relación entre lo que explica y entre lo que deja de explicar sea importante. Si no lo hiciera, sería tanto como decir que ninguna de las variables explicativas que hemos usado en su construcción tiene ninguna relación con la dependiente. Esa es, precisamente, la hipótesis nula que hay que contrastar:

$$H_o: \quad \beta_1 = \beta_2 = \cdots \beta_k = 0 \tag{8.36}$$

De no poder rechazarla usando el estadístico F de la expresión (8.35) no tendría sentido ningún análisis posterior. Téngase en cuenta que, de rechazar esa hipótesis, no estaremos diciendo que todos los β sean no nulos, sino que alguno lo es, y será necesaria una prueba específica para cada uno de ellos para determinar cuál. Si nos fijamos en la información del cuadro 8.2, vemos cómo, en nuestra ilustración, el estadístico $F(1,8) = 151{,}5; p < 0{,}01$ nos permite rechazar la hipótesis (8.36), lo que implica que el modelo es capaz de explicar una cantidad significativa de la varianza de la variable dependiente. Al tener solo un regresor, este debe ser significativo, por lo que la prueba individual no tiene sentido, pero deberemos explicarla a continuación para el caso general. Nótese que el número de grados de libertad es $k = 1$ en el numerador (un único regresor) y $n-k-1 = 10-1-1 = 8$ en el denominador.

Veamos el contraste para el conjunto de parámetros aplicado al caso 8.3 que analizaba los determinantes del absentismo. La sintaxis de estimación ya se presentó y los resultados estaban ya ofrecidos en el cuadro 8.4. En el mismo se comprueba que puede rechazarse la hipótesis nula de que todos los parámetros sean cero, puesto que $F(3,44) = 33{,}29; p < 0{,}01$.

8.4.2. Contraste para un parámetro individual

Una vez rechazada la hipótesis nula de que todos los parámetros sean cero, surge todo un abanico de posibilidades de contrastes para los parámetros individuales. La más habitual es la de analizar la significatividad de un parámetro concreto contrastando la hipótesis nula de que ese β sea nulo (esa variable no ejercería una influencia significativa). Pero hay más posibilidades que pueden

Estadísticos e-Books & Papers

tener sentido para el investigador en función de las hipótesis generales de su investigación, por ejemplo:

- Si un parámetro es estadísticamente menor que cero (negativo), por ejemplo para analizar si, en el caso 8.2, podemos tener la seguridad de que crecimientos en el precio vendrán acompañados de caídas en la demanda.
- Si un parámetro es significativamente superior a 1, por ejemplo si ese parámetro es la elasticidad/renta convertiría al bien en un bien de lujo.
- Si un subconjunto de parámetros (no uno individualmente) es nulo, con el fin de valorar la exclusión de ese conjunto de variables del modelo.

Comenzaremos por el caso más habitual, que es imprescindible en toda investigación (contrastar si un β concreto es nulo), y veremos a continuación el resto de contrastes menos habituales.

En el caso concreto de contrastes sobre un solo coeficiente o de una hipótesis nula con una sola restricción se puede utilizar un estadístico con distribución t de Student. Así, cuando se trata de contrastar una hipótesis nula, con una sola restricción, tal como:

$$H_0 : \beta_j = d \tag{8.37}$$

entonces se puede utilizar el siguiente estadístico t (que es la raíz cuadrada del estadístico F correspondiente):

$$t = \frac{\hat{\beta}_j - d}{SE_{\beta_j}} \tag{8.38}$$

donde SE_{β_j} es el error estándar en la estimación del parámetro β_j. Este estadístico se distribuye como una t de Student con $n-k$ grados de libertad, siendo n el tamaño muestral y k el número de regresores.

Para contrastar la significatividad de que un coeficiente individual sea nulo, basta con hacer $d = 0$ y el estadítico t tomaría la forma:

$$t = \frac{\hat{\beta}_j}{SE_{\beta_j}} \tag{8.39}$$

Ante una hipótesis nula $H_0 : \beta_j = 0$ existen varias hipótesis alternativas, todas con sentido en función del objetivo del investigador: cola derecha, cola izquierda y dos colas.

En el caso de la **hipótesis alternativa de dos colas**, esta es la siguiente:

$$H_1 : \beta_j \neq 0 \tag{8.40}$$

es decir, el signo de β_j no está bien determinado por la teoría ni el sentido común, solo nos interesa saber si su valor absoluto es distinto de cero, pero

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

conceptualmente nos sirve exactamente igual que sea positivo o negativo. Se rechazaría la hipótesis nula cuando:

$$\left|t_{\hat{\beta}_j}\right| \geq t_{n-1-k}^{\alpha/2} \tag{8.41}$$

El valor de t que aparece en las salidas que hemos venido mostrando se corresponde con el contraste de dos colas. Por ejemplo, para el caso 8.2 si la pregunta que nos hacemos es si el precio del café influye en su demanda —y no tenemos una hipótesis a priori acerca de cómo lo hace— entonces viendo el cuadro 8.4, vemos como $\left|t_{\hat{\beta}_j}\right| = 13.85$. Si buscáramos en tablas[1] $t_{10}^{\alpha/2}$ veríamos que toma el valor 3,17 para $\alpha = 0{,}01$, por lo que podemos rechazar la hipótesis nula, aunque este cálculo era innecesario en la medida en que la propia salida ya nos da esa significatividad bajo la etiqueta `Pr(>|t|) = 7.52e-08`.

Cuando no podemos leer directamente la salida es cuando tenemos hipótesis sobre el signo del parámetro. Si nuestra hipótesis es que el precio del café influye significativa y *negativamente* en la demanda, entonces la hipótesis alternativa es la siguiente:

$$H_1 : \beta_1 < 0 \tag{8.42}$$

En esta situación la hipótesis nula se rechazará cuando[2]:

$$t_{\hat{\beta}_j} \leq -t_{n-1-k}^{\alpha} \tag{8.43}$$

El valor de $-t_{n-1-k}^{\alpha}$ es $-2,764$ para α =0,01, por lo tanto, como $-13{,}85 <$ 2,764, podemos confirmar que el coeficiente para el precio del café no solo es significativamente distinto de cero, sino que es negativo.

8.4.3. Contraste para un subconjunto de parámetros

Otro contraste que suele ser habitual es el **contraste para un subconjunto de parámetros**. Lo aplicamos cuando deseamos saber si tiene sentido la exclusión de un subconjunto de variables del modelo en respuesta a una hipótesis teórica. Consideremos, para ello el siguiente caso.

Caso 8.4. Determinantes de los salarios en la empresa

Se tienen datos de una muestra de 935 observaciones de empleados de una empresa de los que se conocen las siguientes variables: *salario,* que es la remu-

[1] Las tablas dan normalmente el valor para una cola, al mirar en tablas para $\alpha/2$ simplemente estamos repartiendo esa probabilidad entre las dos colas. Por ejemplo en `R` podemos obtener los valores mediante la función `qt`. Si señalamos `qt(0.01,df=10)` entonces nos va a dar -2,76 que es el valor que deja un 1 % de probabilidad en la cola inferior, sin embargo si le señalamos `qt(0.005,df=10)` nos dará -3,16 que deja un 0,5 % en la cola inferior mientras que 3,16 dejará otro 0,5 % en la superior. Luego es `qt(0.005,df=10)` lo que tendríamos que indicar para el cálculo.

[2] Es el contraste para signo negativo, si el contraste fuera para un signo positivo su equivalente para el rechazo sería $t_{\hat{\beta}_j} < t_{n-1-k}^{\alpha}$

Estadísticos e-Books & Papers

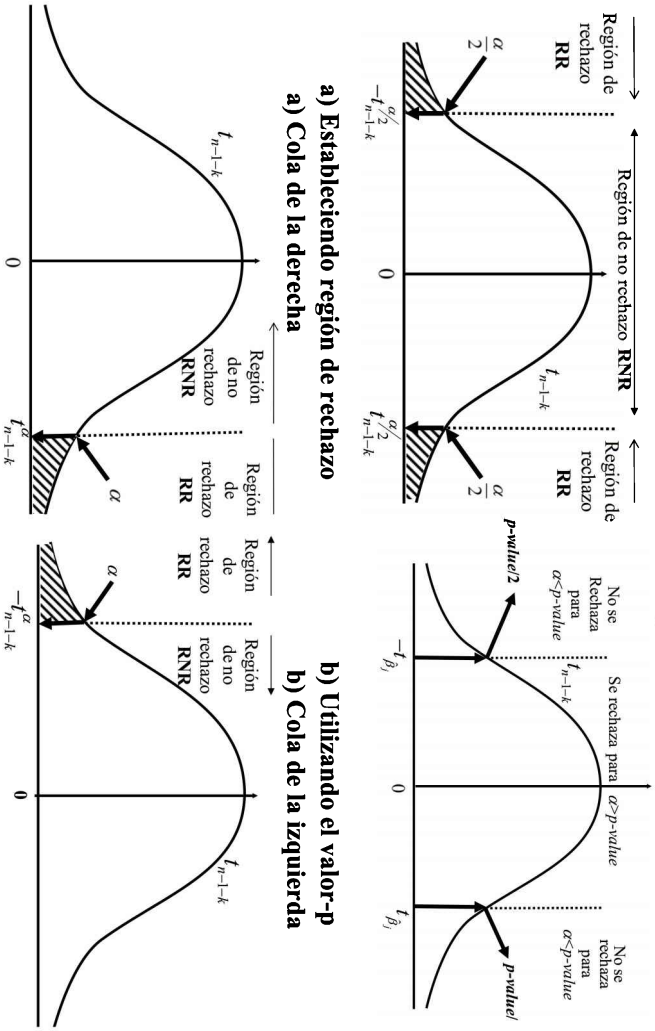


Figura 8.3.: Contraste de una y dos colas

a) Estableciendo región de rechazo

a) Cola de la derecha

b) Utilizando el valor-p

b) Cola de la izquierda

© Ediciones Paraninfo

Estadísticos e-Books & Papers

neración mensual, *educ,* que es el nivel educativo medido en años de formación, *antigue,* que son los años trabajados en la empresa y *edad,* que es la edad en años del trabajador. Con ello se plantea este modelo:

$$\ln(salario) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 antigue + \beta_4 edad + \varepsilon \tag{8.44}$$

pero el investigador tiene la intención de excluir *antigue* del modelo, ya que en muchos casos es igual a la experiencia, y también la *edad*, ya que está altamente correlacionada con la experiencia. ¿Es aceptable la exclusión de ambas variables?

Veamos el planteamiento de este tipo de contrastes. Partimos del siguiente modelo general:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{8.45}$$

Supongamos que hay r restricciones de exclusión a contrastar. Si (8.45) es el modelo general G, entonces el *modelo restringido* R será:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-r} x_{k-r} + \varepsilon \tag{8.46}$$

donde se han eliminado las r variables que van desde $k-r$ hasta k. Por lo tanto, las hipótesis nula y alternativa en que se sustancia lo anterior serán:

$$\begin{array}{c} H_0: \ \beta_{k-r+1} = \beta_{k-r+2} = \beta_k = 0 \\ H_1: \ H_0 \text{ no es cierta} \end{array} \tag{8.47}$$

Para realizar el contraste de la hipótesis (8.47) se utiliza el siguiente estadístico:

$$F = \frac{(SSR_R - SSR_G)/r}{SSR_G/(n-1-k)} \tag{8.48}$$

donde SSR_G es la suma de los cuadrados de los residuos del modelo no restringido y SSR_R la del restringido. El estadístico F se distribuye como una F con r grados de libertad en el numerador y $r-1-k$ en el denominador, siendo la regla de decisión la siguiente. Se rechazará la hipótesis nula para un nivel de significación α si:

$$F \geq F^{\alpha}_{r,n-1-k} \tag{8.49}$$

Este contraste por sí solo no nos permite decir cuáles de las variables tienen un efecto parcial sobre y, ya que todas ellas pueden afectar a y, o tal vez solo una afecta a y. Si no se rechaza H_0, entonces decimos que no son estadísticamente significativas conjuntamente, o simplemente que no son significativas conjuntamente, lo que a menudo justifica su eliminación del modelo. El estadístico F es a menudo útil para contrastar la exclusión de un grupo de variables cuando las variables del grupo están altamente correlacionadas entre sí.

Estadísticos e-Books & Papers

Cuadro 8.6.: Modelos general y restringido para el caso 8.4

```
=================================================================
                                Dependent variable:
                        -------------------------------------------
                                     log(salario)
                              (1)                     (2)
-----------------------------------------------------------------
educ                      0.072***                0.078***
                           (0.007)                 (0.007)

exper                     0.012***                0.020***
                           (0.004)                 (0.003)

antigue                   0.013***
                           (0.003)

edad                       0.009*
                           (0.005)

Constant                  5.296***                5.503***
                           (0.158)                 (0.112)

-----------------------------------------------------------------
Observations                935                     935
R2                         0.158                   0.131
Adjusted R2                0.154                   0.129
Residual Std. Error    0.387 (df = 930)        0.393 (df = 932)
F Statistic         43.621*** (df = 4; 930) 70.162*** (df = 2; 932)
=================================================================
Note:                                   *p<0.1; **p<0.05; ***p<0.01
```

Veamos la aplicación a nuestro caso. Estimamos ambos modelos (cuadro 8.6) que visualizamos en formato de tabla mediante `stargazer{stargazer}` (Hlavac, 2015) y, mediante la opción `anova{stats}`, extraemos del objeto donde hemos guardado la estimación la suma de los cuadrados de los residuos (cuadro 8.7).

```
fit.g<-lm(log(salario)~educ+exper+antigue+edad,data=datos)
fit.r<-lm(log(salario)~educ+exper,data=datos)
stargazer(fit.g,fit.r,type="text")

aov.g<-anova(fit.g)
aov.r<-anova(fit.r)
```

Con el valor de la suma de los cuadrados de los residuos del modelo general SSR_G y del restringido SSR_R podemos aplicar directamente la expresión (8.48), solo hemos de ser conscientes de que hemos introducido $r = 2$ restricciones, que el modelo general tiene $k = 4$ regresores y que contamos con $n = 935$ casos. Así:

$$F = \frac{(143{,}979 - 139{,}486)/2}{139{,}486/(935 - 1 - 4)} = 14{,}97$$

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 8.7.: Análisis de la varianza de los modelos general y restringido para el caso 8.4

```
> aov.g
Analysis of Variance Table

Response: log(salario)
           Df  Sum Sq Mean Sq  F value    Pr(>F)
educ        1  16.138 16.1377 107.5955 < 2.2e-16 ***
exper       1   5.540  5.5400  36.9367 1.778e-09 ***
antigue     1   4.018  4.0177  26.7871 2.783e-07 ***
edad        1   0.475  0.4749   3.1663    0.0755 .
Residuals 930 139.486  0.1500
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> aov.r
Analysis of Variance Table

Response: log(salario)
           Df  Sum Sq Mean Sq F value    Pr(>F)
educ        1  16.138 16.1377 104.462 < 2.2e-16 ***
exper       1   5.540  5.5400  35.861 3.022e-09 ***
Residuals 932 143.979  0.1545
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

y como el valor crítico viene determinado para $\alpha = 0{,}01$ por $F_{2;930}^{0,01} = 4{,}62$, el estadístico calculado supera ese valor, por lo que podemos rechazar la hipótesis nula de que el subconjunto de coeficientes de regresión es nulo y el investigador no puede prescindir de ellos sin un empeoramiento significativo del modelo.

8.5. Bondad de ajuste del modelo

Una vez que se ha realizado el ajuste por mínimos cuadrados, conviene disponer de algún indicador que permita medir el grado de ajuste entre el modelo y los datos. En el caso de que se hayan estimado varios modelos alternativos podrían utilizarse medidas de este tipo, a las que se denomina medidas de la bondad del ajuste, para seleccionar el modelo más adecuado.

Existen en la literatura econométrica numerosas medidas de la bondad del ajuste. Las más conocidas son el **coeficiente de determinación**, al que se designa por R^2 o R cuadrado, y el **coeficiente de determinación corregido**, al que se designa $\overline{R}^2$ o R cuadrado corregido. Dadas que estas medidas tienen algunas limitaciones, se expondrá también el **estadístico AIC**, introducido por Akaike (1974) y cuyas siglas corresponden a la expresión *Akaike Information Criterium.*

Estadísticos e-Books & Papers

8.5.1. Coeficiente de determinación

Si el lector siguió con claridad la presentación del estadístico F para el contraste de la significatividad global del modelo, le resultará muy sencillo seguir el sentido del coeficiente de determinación. Si el estadístico F era una ratio entre lo que el modelo explicaba (SSM) y lo que le restaba por explicar (SSR), el coeficiente de determinación es simplemente el porcentaje de la información que contiene el modelo (SST) que es explicada por él (SSM), en síntesis: $R^2 = SSM/SST$. Si se recuperan las expresiones (8.32) y (8.33), entonces es inmediato:

$$R^2 = \frac{SSM}{SST} = \frac{\sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{8.50}$$

Una forma alternativa de expresar lo anterior y dado que si:

$$SST = SSM + SSR \tag{8.51}$$

entonces:

$$R^2 = \frac{SSM}{SST} = \frac{SST - SSR}{SST} = 1 - \frac{SSR}{SST}$$

sería escribir:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{8.52}$$

Los valores extremos del coeficiente de determinación son: 0, cuando la varianza explicada es nula, y 1, cuando la varianza residual es nula, es decir, cuando el ajuste es perfecto.

Para interpretar adecuadamente el coeficiente de determinación conviene tener en cuenta las siguientes observaciones:

1. A medida que en un modelo se añaden nuevas variables explicativas, el coeficiente de determinación aumenta su valor o, en el peor de los casos, queda con el mismo valor. Esto ocurre aunque la variable o variables añadidas no tengan ninguna relación con la variable endógena.

2. Si el modelo no tiene término independiente, el coeficiente de determinación no tiene una interpretación clara debido a que entonces no se cumple la descomposición dada en (8.51). Además, las dos formas de cálculo se-

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

ñaladas —(8.50) y (8.52)— conducirán en general a resultados diferentes, que en algunos casos pueden caer fuera del intervalo [0,1].

3. Cuando se estima un modelo con datos de series temporales se obtienen en general coeficientes de determinación elevados, aun en casos en que las variables explicativas tengan una relación causal débil. Esto es debido a que las variables de la ecuación estimada pueden estar sometidas a una evolución tendencial más o menos similar.

4. El coeficiente de determinación no se puede utilizar para comparar modelos en los que la forma funcional con que aparece la variable endógena es distinta. Por ejemplo, el R^2 no se puede aplicar para comparar dos modelos en los que los regresandos son y y $\ln(y)$, respectivamente.

Para distinguirlo del corregido, al coeficiente de determinación que se acaba de exponer se le da el calificativo de *ordinario*.

8.5.2. Coeficiente de determinación corregido

El coeficiente de determinación corregido $\overline{R}^2$ permite comparar modelos con distinto número k de regresores. Analíticamente viene dado por:

$$\overline{R}^2 = 1 - \frac{n-1}{n-1-k}\left(1 - R^2\right) \tag{8.53}$$

El $\overline{R}^2$ toma el valor 1 cuando el ajuste es perfecto. En cambio, no está acotado por la parte inferior, pudiendo tomar valores negativos cuando el ajuste realizado es muy malo.

Cuando en un modelo se añade una nueva variable explicativa $\overline{R}^2$ puede aumentar, quedar igual o disminuir su valor. Para que aumente es necesario que la variable añadida tenga un cierto poder explicativo. Por el contrario, si la variable añadida tiene un poder pequeño o nulo, el coeficiente de determinación corregido disminuirá de valor, penalizándose de esta forma su introducción.

Las observaciones 2), 3) y 4) sobre el R^2 siguen siendo válidas para el $\overline{R}^2$.

8.5.3. Estadístico AIC

El estadístico AIC, basado en la teoría de la información, tiene la siguiente expresión:

$$AIC = n \ln\left(\frac{SSR}{n}\right) + 2\,(k+1) \tag{8.54}$$

En el estadístico AIC, a diferencia de los coeficientes de determinación ordinario y corregido, cuanto mejor es el ajuste, más pequeño es el valor que toma el estadístico. El estadístico AIC penaliza la introducción de nuevas variables

Estadísticos e-Books & Papers

explicativas, ya que, como puede verse, en el segundo término del segundo miembro de (8.54) figura el número de parámetros $2(k+1)$. Su lógica es bastante sencilla. Si añadimos a un modelo determinado un regresor adicional pero no somos capaces de reducir los residuos SSR, entonces el valor original del AIC se verá incrementado en +4. En general si, cuando añadimos un regresor, ese +2 se ve compensado por la caída del SSR —y por tanto del $n\ln(SSR/n)$— el AIC puede reducirse.

El estadístico AIC no está acotado a diferencia del R^2 y el $\overline{R}^2$, aunque este último está parcialmente acotado. Por otra parte, el estadístico AIC no es una medida de carácter relativo, como lo son el R^2 y el $\overline{R}^2$. Por ello, no puede decirse que un valor obtenido en un modelo sea en sí mismo elevado o bajo.

La utilidad del estadístico AIC se manifiesta cuando se comparan los valores obtenidos en modelos alternativos, ya que permite comparar todo tipo de modelos. En concreto, para ver cómo se pueden realizar comparaciones nos vamos a referir al caso más usual en que se comparan dos modelos en los que los regresandos son y y $\ln(y)$, respectivamente.

Cuando el regresando es y se aplica directamente la fórmula (8.54), ya que el logaritmo de verosimilitud va referido directamente al vector original de variables endógenas. Cuando el regresando es $\ln(y)$ y, además, se desea comparar con otro modelo en el que el regresando es directamente y, la fórmula que se emplea es la siguiente:

$$AIC_L = AIC + 2\ln\bar{y} \tag{8.55}$$

8.5.4. Error estándar de la estimación

También es útil analizar la desviación estándar de los residuos. La lógica es que esta desviación estándar debería ser menor que la desviación típica de la variable dependiente, es decir, parece lógico esperar que las observaciones varíen menos alrededor de la recta de regresión que alrededor de su media, como se aprecia intuitivamente en los paneles (a) y (b) de la figura 8.2.

La salida estándar de una estimación mediante `lm{stats}` ofrece toda esta información salvo el AIC. Si nos fijamos en la estimación del caso 8.3 en el cuadro 8.5, podemos comprobar fácilmente como:

$$R^2 = 0{,}6942$$
$$\overline{R}^2 = 0{,}6733$$
$$SE_R = 2{,}164$$

Para tener una referencia para evaluar el error estándar de los residuos hemos de conocer la desviación de la variable dependiente que, como se observa en el cuadro 8.8, es claramente superior (3,786).

```
stat.desc(datos$absen,basic=FALSE,desc=TRUE)
```

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 8.8.: Estadísticos descriptivos de *absen*

```
                          x
median        5.0000000
mean          4.5000000
SE.mean       0.5465884
CI.mean.0.95  1.0995940
var          14.3404255
std.dev       3.7868754
coef.var      0.8415279
```

Cuadro 8.9.: Análisis de la varianza en la estimación del caso 8.3

```
Analysis of Variance Table

Response: absen
          Df  Sum Sq Mean Sq F value    Pr(>F)
edad       1 307.366 307.366 65.6108 2.899e-10 ***
antigue    1  44.985  44.985  9.6025  0.003381 **
salario    1 115.522 115.522 24.6596 1.079e-05 ***
Residuals 44 206.126   4.685
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La única información que no proporciona directamente `lm{stats}` es el AIC. No la proporciona en la salida estándar, sí que la calcula. Para que la ofrezca basta pedírselo mediante `extractAIC(fit)`, donde `fit` es el objeto en el que se han guardado los resultados. El resultado para el caso 8.3 sería: 77,9498. Como ya hemos señalado que es una información que solo tiene valor cuando se compara con otros modelos alternativos, no se ofrece por defecto.

Como ejercicio meramente ilustrativo, podemos calcular manualmente los indicadores anteriores solicitando a `lm{stats}` que nos ofrezca los resultados de análisis de la varianza en que se basa el cálculo del estadístico F (cuadro 8.9).

```
anova(fit)
```

Así, teniendo en cuenta sus respectivas expresiones (8.50), (8.53) y (8.54):

$$R^2 = \frac{SSM}{SST} = \frac{307,36 + 44,98 + 115,52}{307,36 + 44,98 + 115,52 + 206,12} = 0{,}6942$$

$$\overline{R}^2 = 1 - \frac{48-1}{48-1-3}(1 - 0{,}6942) = 0{,}6733$$

Estadísticos e-Books & Papers

$$AIC = 48 \ln \left(\frac{206,126}{48} \right) + 2 \times 4 = 77{,}9498$$

8.6. Supuestos del análisis de regresión múltiple

Antes de interpretar los resultados de un análisis de regresión múltiple, es necesario evaluar si se cumplen los supuestos en que se basa el mismo. No hemos querido entrar en detalles a lo largo de la presentación de los modelos para no hacer más compleja la misma, pero ha llegado el momento de hacer explícitos estos supuestos y evaluar el mecanismo para su comprobación y, llegado el caso, para introducir correcciones si algún supuesto no se cumple. Las hipótesis fundamentales son la siguientes:

- No multicolinealidad. No puede haber una relación exacta entre los regresores porque la matriz $\mathbf{X}$ no sería invertible y, como vimos en la expresión (8.23), esta inversión era necesaria. La relación exacta, que derivaría en un determinante nulo de la matriz, se denomina multicolinealidad perfecta, pero no es necesario que la multicolinealidad sea perfecta para que existan problemas. Si no es perfecta pero es muy alta, los parámetros podrían obtenerse pero la fiabilidad de los mismos quedaría afectada.

- Normalidad, los residuos siguen una distribución normal.

- Homocedasticidad. Esta palabra viene del griego: *homo* (igual) y *scedasticidad* (variabilidad). Esto significa que la variabilidad en torno a la línea de regresión es la misma en toda la muestra de x; es decir, que no aumenta o disminuye cuando x varía, como puede verse en la figura 8.4, parte a), donde las perturbaciones son homocedásticas.

- Linealidad. La relación entre el regresando, los regresores y el error es lineal. El regresando y los regresores pueden ser cualquier función de la variable endógena o de las variables predeterminadas, respectivamente, siempre que entre regresando y regresores se mantenga una relación lineal, es decir, el modelo sea lineal en los parámetros.

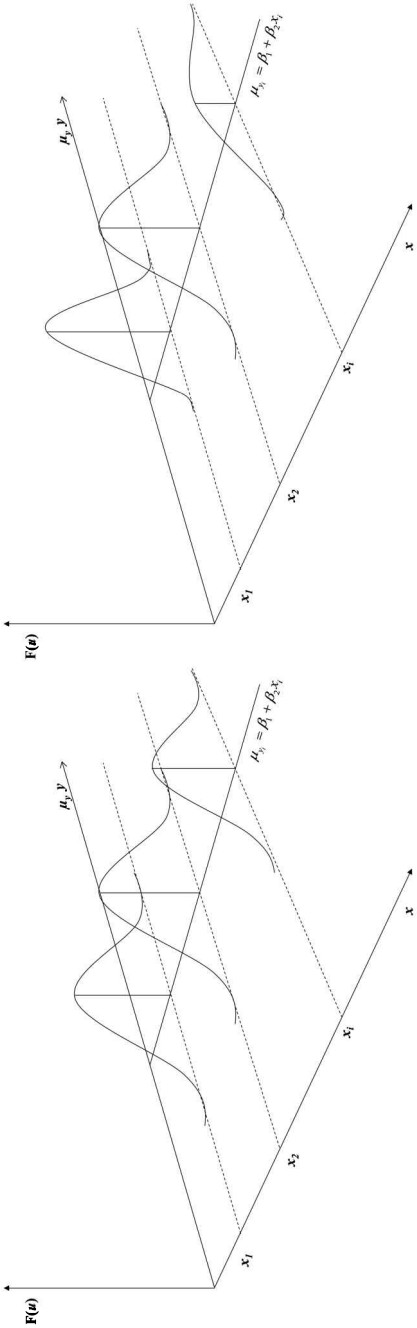
- Independencia de los términos de error. Es decir, los errores correspondientes a diferentes individuos o a diferentes momentos de tiempo no están correlacionados entre sí. Este supuesto de no autocorrelación o no correlación serial, al igual que en el caso de homocedasticidad, es contrastable a posteriori. La transgresión de este supuesto se produce con bastante frecuencia en los modelos que utilizan datos de series temporales.

- *Outliers*. No es estrictamente un supuesto en el que se base la regresión, pero es necesario evaluar, como veremos inmediatamente, si existen casos influyentes que puedan estar sesgando los resultados del análisis.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 8.4.: Ejemplo de homocedasticidad y heteroscedasticidad



Estadísticos e-Books & Papers

Caso 8.5. Determinantes del desempeño del trabajador de una empresa[3]. El servicio de recursos humanos de una empresa ha pasado a los 60 componentes de su equipo de ventas un *test* para medir las variables que a continuación se detallan. También tiene una medida de su desempeño en el trabajo (*perf*) obtenida como el volumen de pedidos conseguidos en decenas de miles de euros corregidos por la dificultad de la zona asignada.

- *perf.* Desempeño en el puesto de trabajo (ventas conseguidas en decenas de miles de euros corregidos por el factor de dificultad de zona).
- *iq.* Resultados del test de inteligencia.
- *mot.* Resultados del test de motivación.
- *soc.* Resultados del test de sociabilidad.

El objetivo es evaluar mediante una regresión lineal múltiple si existe relación entre el desempeño en el puesto de trabajo y la inteligencia, motivación y sociabilidad, pero para ello es necesario, previamente, evaluar si se cumplen los supuestos necesarios para la estimación de ese modelo.

8.6.1. Multicolinealidad

Como vimos con anterioridad, para que el modelo sea estimable es necesario que no exista relación lineal exacta entre los regresores, o, en otras palabras, que no exista multicolinealidad perfecta en el modelo. Esta hipótesis es necesaria para el cálculo del vector de estimadores mínimo cuadráticos. La multicolinealidad perfecta no se suele presentar en la práctica, salvo que se diseñe mal el modelo. En cambio, sí es frecuente que entre los regresores exista una relación aproximadamente lineal, en cuyo caso los estimadores que se obtengan serán en general poco precisos, aunque siguen conservando la propiedad de lineales, insesgados y óptimos. En otras palabras, la relación entre regresores hace que sea difícil cuantificar con precisión el efecto que cada regresor ejerce sobre el regresando, lo que determina que las varianzas de los estimadores sean elevadas. Cuando se presenta una relación aproximadamente lineal entre los regresores, se dice que existe multicolinealidad no perfecta. Es importante señalar que el problema de multicolinealidad, en mayor o menor grado, se plantea porque no existe información suficiente para conseguir una estimación precisa de los parámetros del modelo.

Para analizar este problema, vamos a examinar la varianza de un estimador. En el modelo de regresión lineal múltiple, el estimador de la varianza de un coeficiente cualquiera —por ejemplo, de $\hat{\beta}_j$— se puede formular de la siguiente forma:

[3]El fichero de datos está tomado de los tutoriales de SPSS https://www.spss-tutorials.com/linear-regression-in-spss-example/ y forma parte de los datos de ejemplo para el desarrollo de casos de este *software*.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

$$Var\left(\hat{\beta}_j\right) = \frac{\hat{\sigma}^2}{n\left(1 - R_j^2\right) S_j^2}$$

Como se deduce de la expresión anterior, el estimador de la varianza viene afectado por los siguientes factores:

- Cuanto mayor es $\hat{\sigma}^2$, es decir, cuanto mayor es la dispersión de los datos en el modelo ajustado, mayor será la varianza del estimador.
- Al aumentar el tamaño de la muestra se reduce la varianza del estimador.
- Cuanto menor sea la varianza muestral del regresor x_j, es decir, cuanto menor sea la variabilidad muestral del regresor, mayor será la varianza del correspondiente coeficiente.
- Cuanto mayor sea R_j^2, es decir, cuanto mayor sea la correlación del regresor con el resto de los regresores, mayor será la varianza.

De los cuatro factores señalados es el último el que se refiere a la multicolinealidad. Cuando se presenta multicolinealidad de una cierta gravedad, es decir, cuando uno o más de los R_j^2 se aproximan a 1, se presentan los siguientes problemas al realizar inferencias con el modelo:

- Las varianzas de los estimadores son muy grandes.
- Se puede aceptar con frecuencia la hipótesis nula de que un parámetro es cero, aun cuando la correspondiente variable sea relevante.
- Los coeficientes estimados serán muy sensibles ante pequeños cambios en los datos.

Veamos cómo se puede detectar el nivel de multicolinealidad que existe entre los regresores o variables explicativas del modelo. Como la multicolinealidad es un problema muestral, ya que va asociada a la configuración concreta de la matriz $\mathbf{X}$, no existen contrastes estadísticos, propiamente dichos, que sean aplicables para su detección. En cambio, se han desarrollado numerosas reglas prácticas que tratan de determinar en qué medida la multicolinealidad afecta gravemente a la estimación y contraste de un modelo. Estas reglas no son siempre fiables, siendo en algunos casos muy discutibles. A continuación se van a exponer dos procedimientos (el factor de agrandamiento de la varianza y el número de condición) que son los que gozan de mayor soporte —especialmente el segundo— en la literatura econométrica actual.

Veamos la lógica de la construcción del **factor de inflación de la varianza** o de agrandamiento de la varianza, que nos sirve, además, para explicar otro de los indicadores que se utilizan, que es la **tolerancia.** Parece lógico que si el problema de la multicolinealidad se da entre las variables independientes

Estadísticos e-Books & Papers

regresemos cada una de ellas sobre las demás y analicemos la R_j^2 de cada una de esas regresiones como una medida del nivel en que una está explicada por las demás. Si ese R_j^2 es alto —luego determinaremos qué se puede entender por alto—, esa variable estaría tan relacionada con las demás que podría causarnos un problema.

Si nuestro modelo es:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{8.56}$$

las regresiones que se plantearían entre las variables dependientes serían:

$$\begin{aligned} x_1 &= \alpha_2 x_2 + \alpha_3 x_3 + \cdots + \alpha_k x_k \rightarrow R_1^2 \\ x_2 &= \alpha_1 x_1 + \alpha_3 x_3 + \cdots + \alpha_k x_k \rightarrow R_2^2 \\ &\vdots \\ x_k &= \alpha_2 x_2 + \alpha_3 x_3 + \cdots + \alpha_{k-1} x_{k-1} \rightarrow R_k^2 \end{aligned}$$

Digamos que tenemos un problema de multicolinealidad cuando el $R_j^2 \geq 0{,}75$. El valor es arbitrario pero es el recomendado por Huber y Stephens (1993). Otros autores recomiendan considerar un problema $R_j^2 \geq 0{,}80$ (Menard, 1995) y otros a partir de $R_j^2 \geq 0{,}90$ (Tabachnick y Fidell, 2001). Esto sería lo mismo que decir que tendríamos un problema de multicolinealidad cuando la **tolerancia** fuera $TOL_j \leq 0{,}25$, dado que esta se define como:

$$TOL_j = 1 - R_j^2 \tag{8.57}$$

Pues bien, el **índice de inflación de la varianza (VIF)** se define como:

$$VIF_j = \frac{1}{TOL_j} = \frac{1}{1 - R_j^2} \tag{8.58}$$

cuya interpretación es la siguiente: la $\sqrt{VIF_j}$ es el grado en que el intervalo de confianza para el coeficiente de regresión β_j se agranda —de ahí su nombre— en relación con un modelo en que los regresores fueran ortogonales, estuvieran incorrelacionados. En otras palabras, cuánto se "agranda" el error estándar del estimador como consecuencia de la no ortogonalidad de los regresores. Recordemos que la significatividad individual del parámetro se estima dividiendo el estimador por el error estándar, con lo que, si este se incrementa, disminuiría la probabilidad de que fuera significativo. Al estar conectado a la definición de la R_j^2 el punto de corte viene implícito. Si se considera problemática una $R_j^2 \geq 0{,}75$, se estaría considerando problemática un $VIF_j \geq 4$. Este es el punto de corte habitual (Kabacoff, 2015) en la medida en que significa que se está doblando el error estándar de la estimación $\left(\sqrt{VIF_j} \geq 2\right)$.

Para el caso 8.5, la estimación del modelo y el cálculo del VIF se realizaría de este modo mediante la función `vif{car}`. Los resultados (cuadro 8.10) muestran que, de acuerdo con este criterio, no parece existir un problema de multicolinealidad entre los regresores.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 8.10.: Análisis de la multicolinealidad con el VIF

```
> vif(fit)
      iq      mot      soc
1.016101 1.160834 1.168128
```

```
library(car)
fit<-lm(data=datos,perf~iq+mot+soc)
vif(fit)
```

Otro procedimiento para el diagnóstico de la multicolinealidad es el llamado **índice de condición** (Belsey, 1991; Belsey *et al.*, 1980). Veamos su lógica. La matriz $\mathbf{X}$ está formada por $k+1$ elementos (k regresores más la constante). Si realizáramos un análisis de componentes principales PCA, obtendríamos tantos componentes principales como elementos de esa matriz, es decir, $k+1$. Si hubiera regresores muy correlacionados, se agruparían en un componente principal con un autovalor muy alto (remitimos al lector al tema 11 para un análisis más en profundidad de este concepto), y las sucesivas componentes principales tendrían autovalores muy bajos. En el caso ideal de que los regresores fueran ortogonales, cada uno de los $k+1$ autovalores tomaría el valor 1.

El número de condición se define como la raíz cuadrada de la razón entre el autovalor más grande y el más pequeño, es decir:

$$\kappa(\mathbf{X}) = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} \tag{8.59}$$

mientras que el **índice de condición** para cada uno de los $k+1$ autovalores se define análogamente como:

$$ic(\lambda_i) = \sqrt{\frac{\lambda_{max}}{\lambda_i}} \tag{8.60}$$

Nótese que, en el caso de ortogonalidad, tanto el número de condición como el índice de condición sería 1. Como los autovalores están ordenados de mayor a menor, el número de condición coincidirá con el índice de condición del autovalor más pequeño, que siempre será el último. Se considera que hay un problema de multicolinealidad —aunque hay distintos criterios en la literatura, usaremos el más frecuente— cuando el índice de condición de un componente principal sea superior a 30 (Belsey, 1991). Pero como un componente principal es una combinación lineal de todos los regresores es necesario determinar cuáles son los más implicados. Para ello se ofrece el cuadrado de la correlación entre cada regresor y la componente principal, (normalmente bajo el nombre de "proporciones de la varianza". Siguiendo a Hair *et al.* (2014a), el problema lo generan aquellos regresores que tengan proporciones de varianza en el componente principal problemático superiores a 0,90.

Estadísticos e-Books & Papers

Cuadro 8.11.: Análisis de la multicolinealidad con el índice de condición

```
Condition
Index     Variance Decomposition Proportions
          intercept iq    mot   soc
1   1.000 0.001     0.001 0.002 0.002
2  11.298 0.026     0.247 0.345 0.134
3  13.209 0.002     0.017 0.640 0.642
4  24.437 0.971     0.735 0.013 0.222
```

El índice de condición lo podemos obtener en R mediante la función `colldiag {perturb}`, cuya sintaxis es elemental. El resultado, cuadro 8.11, nos muestra que no hay problemas de multicolinealidad, pero aunque el último componente tuviera un índice de condición más alto o bajáramos el nivel de referencia de 30 a 20 (con un ic de 24,437 lo superaría), la proporción de varianza tampoco supera el valor de 0,90 tomado como referencia.

```
library(perturb)
colldiag(fit)
```

Estos valores vienen generalmente referidos a regresores medidos con escala de longitud unidad (es decir, con los regresores divididos por la raíz cuadrada de la suma de los valores de las observaciones), pero no centrados. No es conveniente centrar los datos (es decir, restarles sus correspondientes medias), ya que esta operación oscurece cualquier dependencia lineal que implique al término independiente. El no centrado es la opción por defecto de `colldiag {perturb}`.

¿Cuáles son las **soluciones ante problemas de multicolinealidad**? En principio, el problema de la multicolinealidad está relacionado con deficiencias en la información muestral. El diseño muestral no experimental es, a menudo, el responsable de estas deficiencias. Sin embargo, la aproximación cuantitativa a los conceptos teóricos puede ser inadecuada, haciendo que en el término de perturbación se absorban errores de especificación. Veamos a continuación algunas de las soluciones propuestas para resolver el problema de la multicolinealidad.

Eliminación de variables. La multicolinealidad puede atenuarse si se eliminan los regresores que son más afectados por la multicolinealidad. El problema que plantea esta solución es que los estimadores del nuevo modelo serán sesgados en el caso de que el modelo original fuera el correcto.

Aumento del tamaño de la muestra. Teniendo en cuenta que un cierto grado de multicolinealidad acarrea problemas cuando aumenta ostensiblemente la varianza muestral de los estimadores, se puede modificar el tamaño de la misma introduciendo observaciones adicionales. Esta solución no siempre es

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

viable, puesto que los datos utilizados en las contrastaciones empíricas proceden generalmente de fuentes estadísticas diversas, interviniendo en contadas ocasiones el investigador en la recogida de información.

Utilización de información extramuestral. Otra posibilidad es la utilización de información extramuestral, bien estableciendo restricciones sobre los parámetros del modelo, bien aprovechando estimadores procedentes de otros estudios.

El establecimiento de restricciones sobre los parámetros del modelo reduce el número de parámetros a estimar y, por tanto, palia las posibles deficiencias de la información muestral. En cualquier caso, para que estas restricciones sean útiles deben estar inspiradas en el propio modelo teórico o, al menos, tener un significado económico.

En general, un inconveniente de esta forma de proceder es que el significado atribuible al estimador obtenido con datos de corte transversal es muy diferente del obtenido con datos temporales. A veces, estos estimadores pueden resultar realmente "extraños" o ajenos al objeto de estudio. Por otra parte, al estimar las varianzas de los estimadores obtenidos en la segunda regresión hay que tener en cuenta la estimación previa.

Utilización de ratios. Si en lugar del regresando y de los regresores del modelo original se utilizan ratios con respecto al regresor que tenga mayor colinealidad, puede hacer que la correlación entre los regresores del modelo disminuya. Una solución de este tipo resulta muy atractiva por su sencillez de aplicación. Sin embargo, las transformaciones de las variables originales del modelo utilizando ratios pueden provocar otro tipo de problemas. Suponiendo admisibles las hipótesis básicas con respecto a las perturbaciones originales del modelo, esta transformación modificaría implícitamente las propiedades del modelo, de tal manera que las perturbaciones del modelo transformado utilizando ratios ya no serían perturbaciones homocedásticas, sino heterocedásticas.

8.6.2. Normalidad

Si no se cumple el supuesto de normalidad de la variable dependiente o, lo que es lo mismo, la normalidad de los residuos, entonces los contrastes t y F que hemos examinado tienen una validez aproximada, que será tanto mayor cuanto mayor sea el tamaño de la muestra. Sin embargo, no es usual realizar contrastes de normalidad, quizás debido a que la mayoría de las veces no se dispone de una muestra lo suficientemente grande —por ejemplo, 50 o más observaciones— que es necesaria para realizar contrastes sobre esta hipótesis. De todas formas, recientemente los contrastes sobre normalidad están recibiendo un interés creciente tanto en los estudios teóricos como aplicados.

Vamos a examinar a continuación uno de los contrastes más aplicados para verificar la hipótesis de normalidad de las perturbaciones en un modelo econométrico, el **test de Jarque y Bera**. Este contraste fue propuesto por Jarque y

Estadísticos e-Books & Papers

Bera (1980), y está basado en los estadísticos de asimetría y curtosis de los residuos. El estadístico de *asimetría* es un momento de tercer orden estandarizado que, aplicado a los residuos, toma el siguiente valor:

$$\gamma_{1(\hat{\varepsilon})} = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^3}{\left[\sum_{i=1}^{n} \hat{\varepsilon}_i^2 / n\right]^{\frac{3}{2}}} \tag{8.61}$$

En una distribución simétrica, como es el caso de la distribución normal, el coeficiente de asimetría es 0.

El estadístico de *curtosis*, que es un momento de cuarto orden estandarizado, toma el siguiente valor cuando se aplica a los residuos:

$$\gamma_{2(\hat{\varepsilon})} = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^4 / n}{\left[\sum_{i=1}^{n} \hat{\varepsilon}_i^2 / n\right]^{2}} \tag{8.62}$$

En una distribución normal estandarizada, es decir, en una distribución $N(0,1)$, el coeficiente de *curtosis* es igual a 3.

El estadístico de Jarque y Bera (1980) (JB) está construido sobre los coeficientes de asimetría y curtosis de los residuos, y viene dado por:

$$JB = \frac{n-k+1}{6}\left(\gamma_{1(\hat{\varepsilon})} + \frac{1}{4}\left[\gamma_{2(\hat{\varepsilon})} - 3\right]^2\right) \tag{8.63}$$

En una distribución normal teórica, la anterior expresión tomará un valor nulo, ya que los coeficientes de asimetría y curtosis toman, respectivamente, los valores de 0 y 3. El estadístico JB tomará valores elevados en la medida que el coeficiente de asimetría se aleje de 0 y el coeficiente de curtosis se aleje de 3. Bajo la hipótesis nula de normalidad, el estadístico JB tiene la siguiente distribución:

$$JB \underset{n\to\infty}{\longrightarrow} \chi_2^2 \tag{8.64}$$

Con la indicación de $n \to \infty$, se quiere señalar que es un contraste asintótico, es decir, que tiene validez cuando la muestra sea suficientemente grande. Sin embargo en `R` existen versiones en las que, como vamos a ver, la significatividad puede calcularse mediante *bootstrapping,* lo que haría no necesitar estas muestras amplias. Aplicado a nuestro caso y recordando que los residuos están guardados en el objeto `fit` y que los extraemos mediante `fit$resid`, podemos calcular el test mediante `jarqueberaTest{fBasics}` y mediante `ajb.norm.test{normtest}`. El primer caso nos da el test mediante

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 8.12.: Test de Jarque y Bera de normalidad multivariante

```
> jarqueberaTest(fit$resid)

Title:
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 0.2725
  P VALUE:
    Asymptotic p Value: 0.8726

> ajb.norm.test(fit$resid, nrepl=2000)

     Adjusted Jarque-Bera test for normality

data:  fit$resid
AJB = 0.19835, p-value = 0.908
```

el cálculo de la significatividad descrito al presentarlo y, en el segundo caso, lo obtiene mediante *bootstrapping*. Los resultados (cuadro 8.10) coinciden en no descartar la hipótesis nula de normalidad multivariante.

```
jarqueberaTest(fit$resid)
ajb.norm.test(fit$resid, nrepl=2000)
```

Otro procedimiento estándar para evaluar la normalidad a través de los residuos es aplicar los test de normalidad que presentamos en el tema 2, los más habituales son el test de **Shapiro-Wilk** (Shapiro y Wilk, 1965), el test de **Kolmogorov-Smirnov** (Chakravarti *et al.*, 1967), el test de **Anderson-Darling** (Stephens, 1974) y el test de **Cramer-von Mises** (Marsaglia y Marsaglia, 2004). Su desarrollo teórico va más allá del alcance de este libro pero, en general, todos ellos plantean la hipótesis nula de que la variable sigue una distribución normal. Es muy importante señalar que, mientras que el test de Shapiro-Wilk es recomendable para muestras inferiores a 2.000 casos, el resto exige muestras superiores a esta cifra. Los resultados del cuadro 8.13 confirman la normalidad para los dos primeros test.

```
jshapiro.test(fit$resid)
lillie.test(fit$resid)
```

Aunque es siempre recomendable recurrir a test estadísticos, es muy habitual el uso del gráfico de comparación de cuantiles Q-Q que presentamos en el tema 2 y sobre el que no vamos a volver respecto a su construcción, salvo que

Estadísticos e-Books & Papers

Cuadro 8.13.: Test de Shapiro y de Kolmogorov para analizar la normalidad de los residuos

```
> shapiro.test(fit$resid)

     Shapiro-Wilk normality test

data:  fit$resid
W = 0.99231, p-value = 0.97

> lillie.test(fit$resid)

     Lilliefors (Kolmogorov-Smirnov) normality test

data:  fit$resid
D = 0.096325, p-value = 0.1805
```

representa los valores de los residuos estandarizados respecto a los valores que deberían tener en caso de normalidad. Si se cumple la normalidad, los residuos deberían caer sobre la línea de 45 grados sin grandes desviaciones. Obtenemos ese gráfico junto con otros que nos servirán para fases ulteriores de diagnóstico mediante `autoplot{ggfortify}`. El gráfico Q-Q, que aparece en la esquina superior derecha de la figura 8.5, no muestra, tampoco, desviaciones respecto a la normalidad en los residuos.

```
autoplot(fit, label.size = 3)
```

Cuando se viola la hipótesis de normalidad, una aproximación razonable es transformar la variable dependiente, aunque esto dificulta la interpretación de los coeficientes. La función `powerTransform{car}` calcula el factor λ que con más probabilidad logrará normalizar la variable dependiente, esto es, y^{λ}.

8.6.3. Homocedasticidad

En la etapa de verificación de los modelos estimados se deben realizar contrastes para determinar si es, o no, admisible la hipótesis de homocedasticidad, una de las hipótesis básicas postuladas al formular el modelo lineal.

En primer lugar, se examinará el test de Breusch-Pagan y, en segundo lugar, una variante del mismo, que es el contraste de White, el más utilizado actualmente. Posteriormente, se examinarán posibles soluciones para el caso de que las perturbaciones del modelo sean heterocedásticas.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

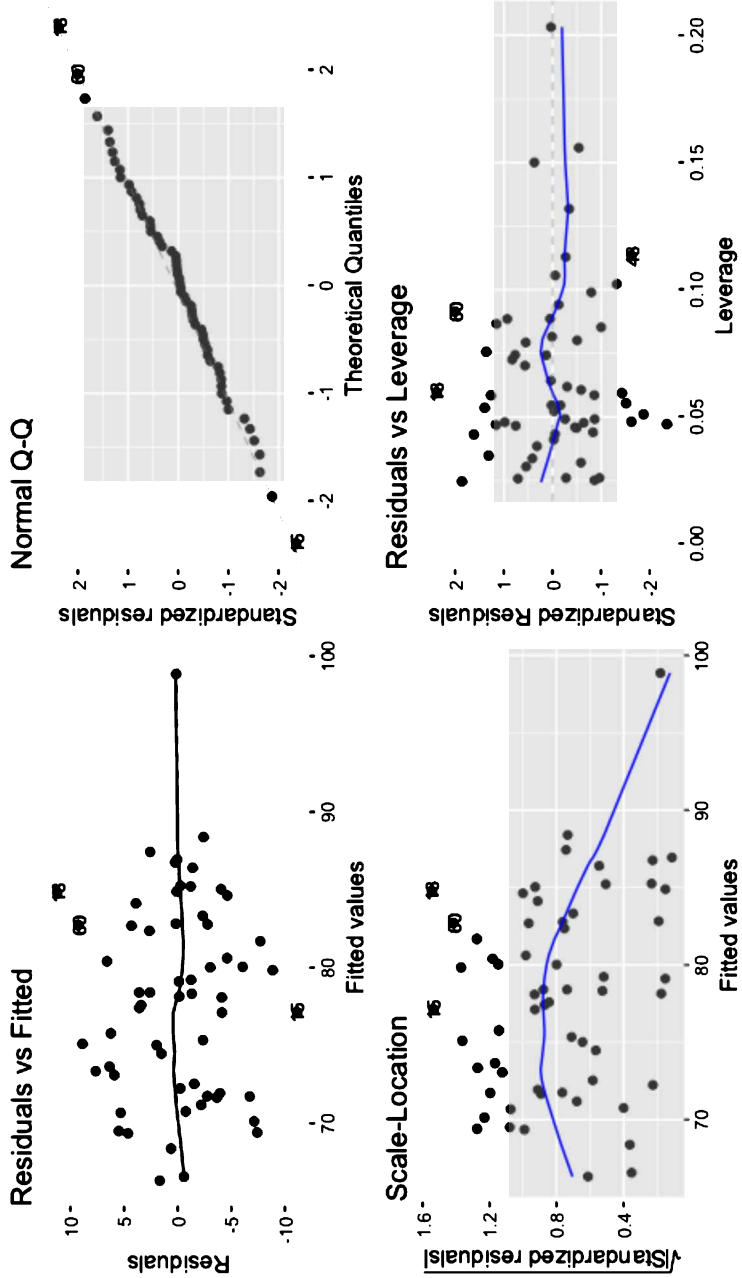


Figura 8.5.: Gráficos de diagnóstico

Estadísticos e-Books & Papers

A. Test de Breusch-Pagan

La lógica de este test, propuesta de manera independiente por Breusch y Pagan (1979) y por Cook y Weisberg (1983) con ligeras modificaciones es bastante sencilla. Se contrapone a la hipótesis nula de que las varianzas son constantes respecto a las variables independientes la hipótesis alternativa de que la varianza depende de las variables independientes. Para contrastarlo se regresan los cuadrados de los residuos sobre los regresores de la función estimada, en el ejemplo que estamos siguiendo para el caso 8.5, y llamando x_1 a la variable *iq*, x_2 a la variable *mot* y x_3 a la variable *soc*:

$$\hat{\varepsilon}_i^2 = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} + u_i \tag{8.65}$$

Pues bien, esta regresión tendrá una R^2 que denominaremos R^2_{BP}, construiremos el estadístico de contraste como:

$$n \times R^2_{BP} \tag{8.66}$$

que, bajo la hipótesis nula sigue una distribución χ^2_k, siendo k el número de regresores sin incluir el intercepto de la nueva regresión efectuada que, en este caso, coincide con el número de regresores originales, no como en el siguiente test.

B. Contraste de White

El contraste propuesto por White (1980) es una variante del test de Breusch-Pagan, donde los residuos al cuadrado no solo se regresan sobre los regresores de la función original, sino también sobre todas sus interacciones y sus términos al cuadrado. Volviendo al ejemplo del caso 8.5 y con la misma notación que hemos empleado con anterioridad:

$$\begin{gathered}\hat{\varepsilon}_i^2 = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \\ + \alpha_4 x_{1i} x_{2i} + \alpha_5 x_{1i} x_{3i} + \alpha_6 x_{2i} x_{3i} + \\ + \alpha_7 x_{1i}^2 + \alpha_8 x_{2i}^2 + \alpha_9 x_{3i}^2 + u_i\end{gathered} \tag{8.67}$$

De nuevo esta regresión tendrá una R^2 que denominaremos R^2_W, construiremos el estadístico de contraste como:

$$n \times R^2_W \tag{8.68}$$

que, bajo, la hipótesis nula, sigue una distribución χ^2_p, siendo p el número de regresores, no los originales (3), sino los 9 de la expresión (8.67), donde tampoco se considera el intercepto.

C. Test de Goldfeld-Quandt

El test propuesto por Goldfeld y Quandt (1965; 1972) es también de una lógica radical. Se ordenan los datos por la variable dependiente y a continuación se

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

divide la muestra en tres submuestras, la central debe ser inferior a un tercio y superior al 15 % de la muestra total. Las submuestras extremas deben ser aproximadamente iguales.

Se estima a continuación el modelo para cada una de las submuestras extremas (obviando la central). Cada una de las estimaciones tendrá una suma de cuadrados de los residuos, lo que venimos denominando SSR. Bajo la hipótesis nula de homocedasticidad se aplica el siguiente contraste, donde m_1 es el tamaño de la submuestra inferior, m_2, el de la superior, y k, el número de regresores:

$$F_{GQ} = \frac{SCR_1/(m_1 - k)}{SCR_2\,(m_2 - k)} \tag{8.69}$$

estadístico que se distribuye como una $F(m_1 - k, m_2 - k)$.

Estos tres estadísticos están implementados en `R`. El primero y el segundo mediante la función **`bptest{lmtest}`** y el segundo mediante la función **`gqtest{lmtest}`**. La sintaxis y los resultados se ofrecen a continuación. Solo hemos de llamar la atención sobre que, en cuanto que caso particular del test de Breusch-Pagan, el test de White se indica en la sintaxis explicitando adicionalmente las variables con los cuadrados y las interacciones que se han de añadir al modelo original que se recogía en el objeto **`fit`**, como se aprecia en la sintaxis. El resultado (cuadro 8.14) no permite rechazar en ninguno de los casos la hipótesis de homoscedasticidad.

```
#Breusch Pagan Test
bptest(fit)
#Test de White,
#se aplica con el anterior pero hay que anadir todas las
#interacciones entre los regresores y sus cuadrados
bptest(fit,~iq*mot+iq*soc+soc*mot+I(iq^2)+I(mot^2)+I(soc^2),
data=datos)
#Test de Goldfeld-Quandt
gqtest(fit)
```

Una vez más, aunque mucho más subjetivo que los estadísticos, es habitual analizar el gráfico que relaciona los residuos estandarizados (su raíz cuadrada) en función de los valores pronosticados. En caso de homocedasticidad los puntos deberían repartirse aleatoriamente alrededor de una línea horizontal. Si no es así, deberían seguir un patrón creciente (triangular o con forma de diamante) al aumentar el valor de la variable pronosticada. En el gráfico inferior izquierdo de la figura 8.5 está el gráfico al que hacemos referencia y no se aprecia ese patrón creciente que denotaría heterocedasticidad.

Si al realizar los contrastes se rechazara la hipótesis nula de homocedasticidad, eso significa que el método de mínimos cuadrados ordinarios (MCO) no es el más adecuado, ya que en ese caso dichos estimadores obtenidos no son óptimos.

Estadísticos e-Books & Papers

Cuadro 8.14.: Contrastes de homocedasticidad

```
> bptest(fit)

      studentized Breusch-Pagan test

data:  fit
BP = 3.8522, df = 3, p-value = 0.2779

bptest(fit,~iq*mot+iq*soc+soc*mot+I(iq^2)+I(mot^2)+I(soc^2),data=datos)

      studentized Breusch-Pagan test

data:  fit
BP = 8.7157, df = 9, p-value = 0.4639

>
> gqtest(fit)

      Goldfeld-Quandt test

data:  fit
GQ = 0.94096, df1 = 26, df2 = 26, p-value = 0.5611
alternative hypothesis: variance increases from segment 1 to 2
```

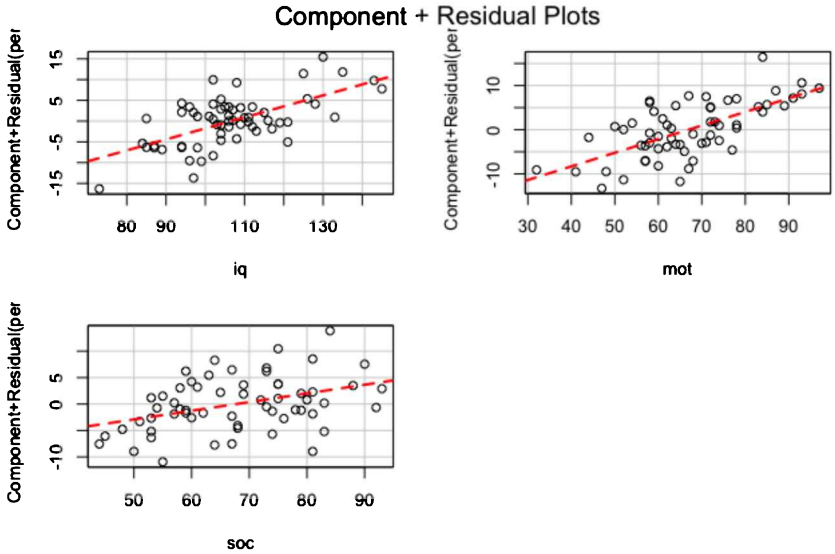
Cuando las perturbaciones son heterocedásticas se debe aplicar el método de mínimos cuadrados generalizados (MCG) para obtener unos estimadores óptimos. A la aplicación específica de MCG al caso concreto de que las perturbaciones sean heterocedásticas se le denomina mínimos cuadrados ponderados (MCP), debido a que su aplicación implica ponderar de forma distinta a cada una de las observaciones. Conviene advertir que en algunas ocasiones en que existen problemas de escala con las observaciones, la heterocedasticidad de las perturbaciones se puede corregir realizando transformaciones (por ejemplo, transformaciones logarítmicas) en las variables.

La aplicación de MCP exige el conocimiento previo del esquema de heterocedasticidad concreto de las perturbaciones. Sin embargo, en buena parte de los casos no se tiene ese conocimiento, por lo que no es posible aplicar estos métodos. Por otra parte, los estimadores obtenidos por MCO bajo el supuesto de heterocedasticidad, además de no ser óptimos, presentan el siguiente problema. La estimación de la matriz de covarianzas de los estimadores obtenida aplicando la fórmula usual no es válida cuando existe heterocedasticidad (y/o autocorrelación). Consecuentemente, los estadísticos t basados en dicha estimación de la matriz de covarianzas darán lugar a inferencias erróneas. Afortunadamente, se han desarrollado métodos para estimar de forma adecuada la matriz de covarianzas de los estimadores obtenidos por MCO, bajo el supuesto de existencia de heterocedasticidad. Utilizando estos métodos, se podrán realizar inferencias correctas a partir de las estimaciones por MCO.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 8.6.: Gráficos de componentes y residuos para el diagnóstico de la linealidad



8.6.4. Linealidad

Si la variable dependiente está linealmente relacionada con las independientes, no debería haber ninguna relación funcional aparente entre los residuos y los valores predichos de esa variable dependiente, es decir, el modelo debería capturar toda la varianza sistemática dejando solo ruido aleatorio. Si en este gráfico que relaciona residuos con valores predichos (gráfico superior izquierdo de la figura 8.5) se apreciara, por ejemplo, una relación curva, el investigador debería añadir un término cuadrático a la regresión.

Otra forma de buscar evidencia de relaciones no lineales es realizar gráficos denominados **gráficos de componentes y residuos,** donde cada regresor x_i se representa frente a los residuos del modelo general más el modelo sin ese regresor, es decir:

$$\varepsilon_i + \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$$

Estos gráficos se obtienen directamente con la función `crPlots{car}` mediante la sintaxis siguiente. El resultado se ofrece en la figura 8.6 y no se aprecian relaciones no lineales.

```
crPlots(fit)
```


Estadísticos e-Books & Papers

Cuadro 8.15.: Test de Durbin-Watson para el contraste de la autocorrelación

```
lag Autocorrelation D-W Statistic p-value
   1     -0.03476846      2.002737    0.99
 Alternative hypothesis: rho != 0
```

8.6.5. Independencia de los términos de error

La mejor forma de asegurarse de que la variable dependiente y los errores son independientes es conocer el modo en que los datos se han recogido. Así, por ejemplo, en los datos de serie temporal cabe esperar que los datos recogidos en momentos cercanos del tiempo estén más correlacionados entre sí que los datos obtenidos en momentos distantes.

La función `durbinWatsonTest{car}` permite aplicar el test de Durbin-Watson con la ventaja, respecto a otros paquetes estadísticos, de que calcula la significatividad del mismo mediante *bootstrapping* para el contraste de la hipótesis nula de que la correlación es nula. Aplicado al caso 8.5 mediante la siguiente sintaxis, el nivel de significación $p = 0{,}99$ no nos permite rechazar la hipótesis nula de ausencia de autocorrelación.

```
durbinWatsonTest(fit)
```

8.6.6. Valores atípicos

Aunque no estamos ante la verificación de un supuesto del modelo lineal, el diagnóstico previo antes de interpretar los resultados pasa, necesariamente, por evaluar la existencia de observaciones atípicas, también conocidas como *outliers,* que tienen un comportamiento distinto al resto de casos. También es importante evaluar si alguna observación atípica ejerce o no influencia muy significativa sobre los parámetros estimados. No siempre una observación atípica será influyente en el sentido de afectar gravemente a los parámetros estimados, por lo tanto, vamos a separar su análisis.

A. *Outliers*

Un *outlier* es una observación que no está bien predicha por el modelo y que, por lo tanto, tendrá un residuo elevado la mayor parte de las veces. La mayor parte de programas estadísticos utilizan los residuos para detectar estos casos clasificando como tales aquellos cuyo residuo estandarizado es superior en valor absoluto a 2. En el tema 2 vimos procedimientos más sofisticados basados en la distancia de Mahalanobis pero, como allí se detallaron con precisión, no repetiremos el diagnóstico basándonos en ella. Sí que utilizaremos la función `outlierTest{car}` para analizar los datos de nuestro caso 8.5 Este procedimiento calcula los errores estudentizados que siguen una distribución

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 8.16.: Test de Durbin-Watson para el contraste de la autocorrelación

```
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
   rstudent unadjusted p-value Bonferonni p
18  2.49044           0.015808      0.94847
```

t con $n-k-2$ grados de libertad, lo que permite tener un criterio objetivo para clasificar a un residuo elevado de *outlier* o no. Así el cuadro 8.16 nos muestra como el caso más extremo —con el residuo mayor— pese a ser superior a 2 ($p = 0{,}015808$) no puede considerarse un *outlier* cuando se corrige la significatividad mediante el criterio de Bonferroni que multiplica el nivel de significación al que puede considerarse un *outlier* por el número de observaciones ($p = 0{,}94847$). Si el caso más extremo no puede considerarse un *outlier*, el resto tampoco podrá serlo.

```
outlierTest(fit)
```

B. Casos influyentes

Los residuos estudentizados son muy útiles para determinar lo bien que el modelo es capaz de predecir ese caso, pero ayudan poco para determinar cuánto influye ese caso en los resultados del modelo —esto es, cómo afecta a la capacidad del modelo para predecir a *todos* los casos—.

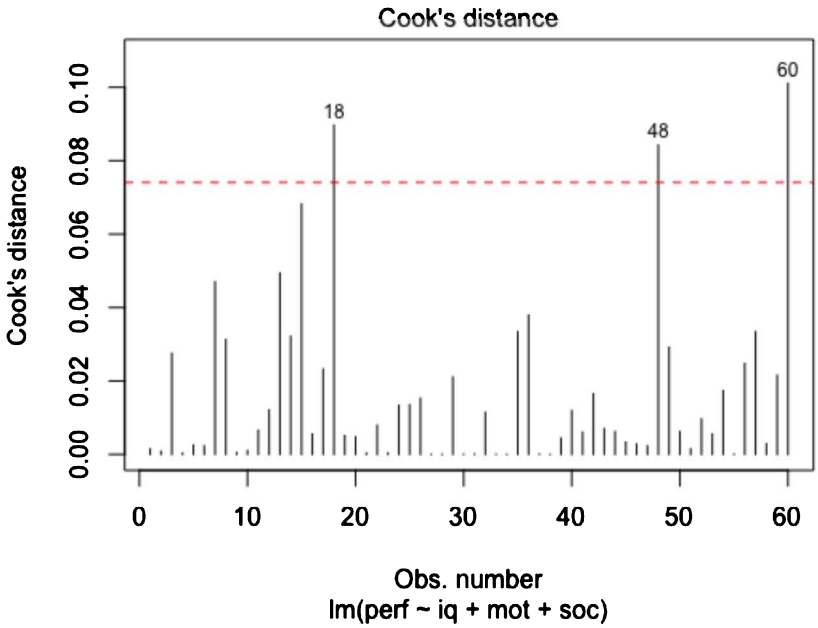
Un procedimiento para analizar este hecho es la denominada **distancia de Cook (D)**, (Cook y Weisberg, 1982). Valores de la distancia de Cook asociados a un caso que sean superiores a $D > 4/(n-k-1)$, donde, como siempre n es el tamaño muestral y k el número de regresores, serían un indicador de que ese caso está influyendo significativamente en la capacidad predictiva del modelo. Otros autores como Field *et al.* (2012) o Kabacoff (2015) ven más recomendable clasificarlo como caso influyente cuando $D > 1$. Podemos obtener fácilmente el gráfico (figura 8.7) que representa la distancia de Cook para cada caso en el modelo estimado y marcar el punto de corte en $D > 4/(n-k-1)$ con la siguiente sintaxis. Vemos que con ese criterio tres casos podrían ser influyentes, pero, con el criterio más generalizado de $D > 1$, ninguno lo sería.

```
cutoff <- 4/(nrow(datos)-length(fit$coefficients)-2)
plot(fit, which=4, cook.levels=cutoff)
abline(h=cutoff, lty=2, col="red")
```

Otro método de detección de casos influyentes es —y no nos atrevemos a traducirlo— los ***hat values*** también conocidos como **puntos de apalancamiento** (*leverage points*). Las observaciones con un elevado *hat value* son *outliers*

Estadísticos e-Books & Papers

Figura 8.7.: Test de Cook para los casos influyentes



© Ediciones Paraninfo

Estadísticos e-Books & Papers

con respecto al conjunto de predictores, es decir, tienen una combinación de valores en los predictores. Para una muestra determinada el *hat value* promedio es $(k+1)/n$, siendo la notación la que venimos siguiendo. Hay distintos criterios para considerar que el *hat value* de una observación es elevado. Unos autores recomiendan investigar los casos que superen dos veces el promedio, es decir, $2(k+1)/n$ (Hoaglin y Welsch, 1978), mientras que Stevens (2009) recomienda investigar los que superen en tres veces ese promedio, o sea, $3(k+1)/n$. Kabacoff (2015) ofrece una sintaxis muy sencilla para representar los *hat values* y los dos puntos de corte en un mismo gráfico (figura 8.8). Solo el caso 37 superaría el nivel de 3 veces el promedio.

```
hat.plot <- function(fit)
{
 p <- length(coefficients(fit))
 n <- length(fitted(fit))
 plot(hatvalues(fit), main="Index Plot of Hat Values")
  abline(h=c(2,3)*p/n, col="red", lty=2)
  identify(1:n, hatvalues(fit), names(hatvalues(fit)))
}
hat.plot(fit)
```

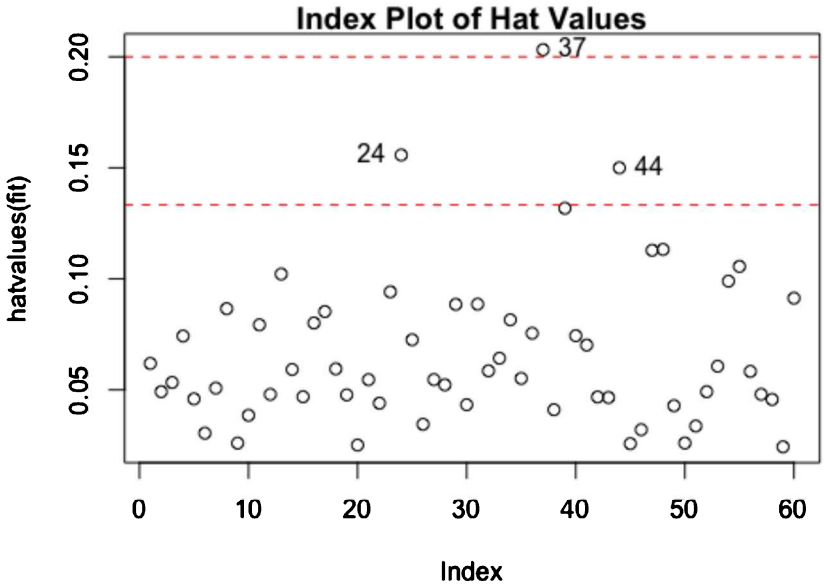
Una forma de evaluar cómo determinadas observaciones pueden afectar al modelo son los llamados ***added variable plots,*** que, para cada regresor, representan los residuos resultantes de regresar la variable dependiente sobre los restantes regresores. De esta forma podemos intuir cómo eliminar un caso cambiaría la recta de regresión parcial. Los podemos obtener mediante la función `avPlots{car}`. La figura 8.9 nos muestra cómo eliminar un caso no nos haría intuir un cambio en la recta.

```
avPlots(fit, ask=FALSE, id.method="identify")
```

Una manera de intentar poner juntos todos los criterios son los **gráficos de influencia.** En estos gráficos se representan los residuos estudentizados marcando los valores ± 2 que planteábamos como niveles a partir de los cuales deberíamos valorarlos como *outliers* potenciales, los *hat values*, marcando 2 y 3 veces el promedio como señalábamos y dibujando para cada residuo el tamaño de su círculo proporcional a su valor de la D de Cook. Se obtienen con la función `influencePlot{car}`. Si nos fijamos en la figura 8.10 no parece haber ningún caso especialmente peligroso. La zona de *hat values* superiores a los dos niveles de corte que a la vez son superiores a residuos estudentizados superiores a ± 2 está vacía. Los casos 44 y 24 tienen *hat values* superiores a 2 veces el promedio pero no tienen residuos superiores a ± 2. Los casos 18 y 60 tiene residuos superiores a 2, pero con *hat values* inferiores al umbral. Como son los que tienen la mayor distancia de Cook son, quizás, los que deberían investigarse con más cuidado.

Estadísticos e-Books & Papers

Figura 8.8.: *Hat values* para detectar casos influyentes



© Ediciones Paraninfo

Estadísticos e-Books & Papers

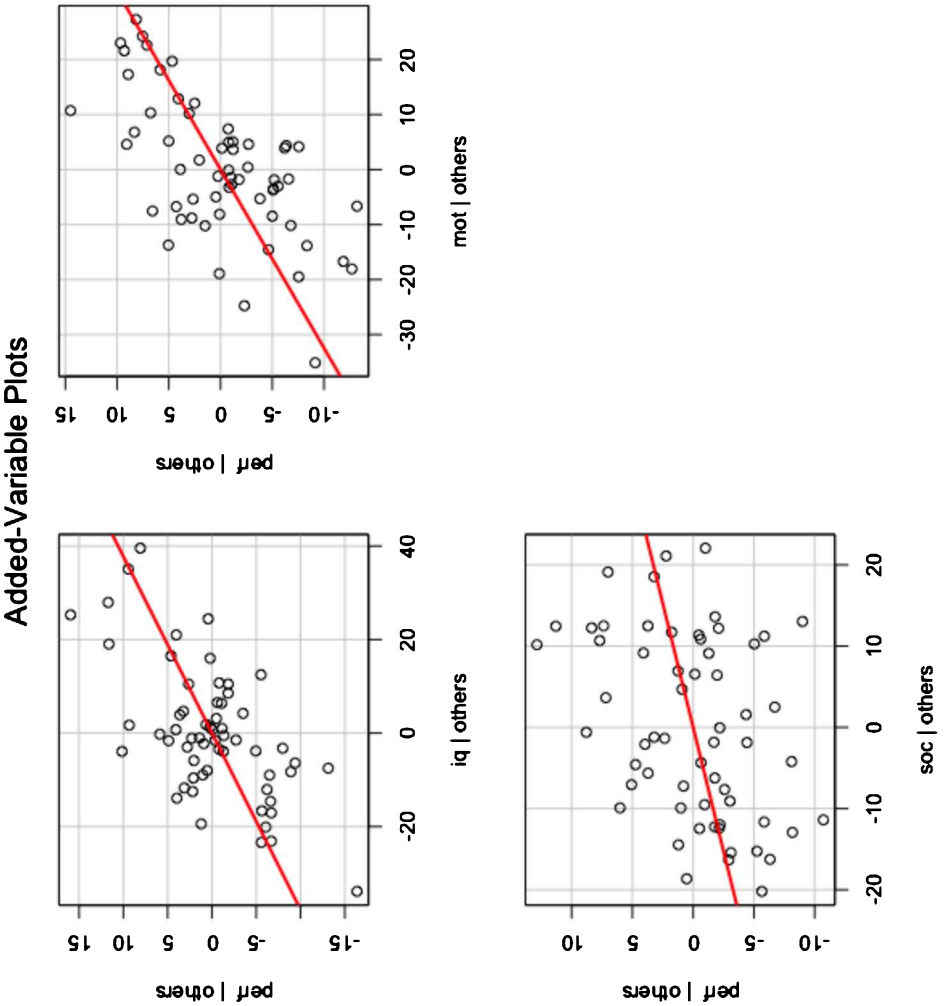
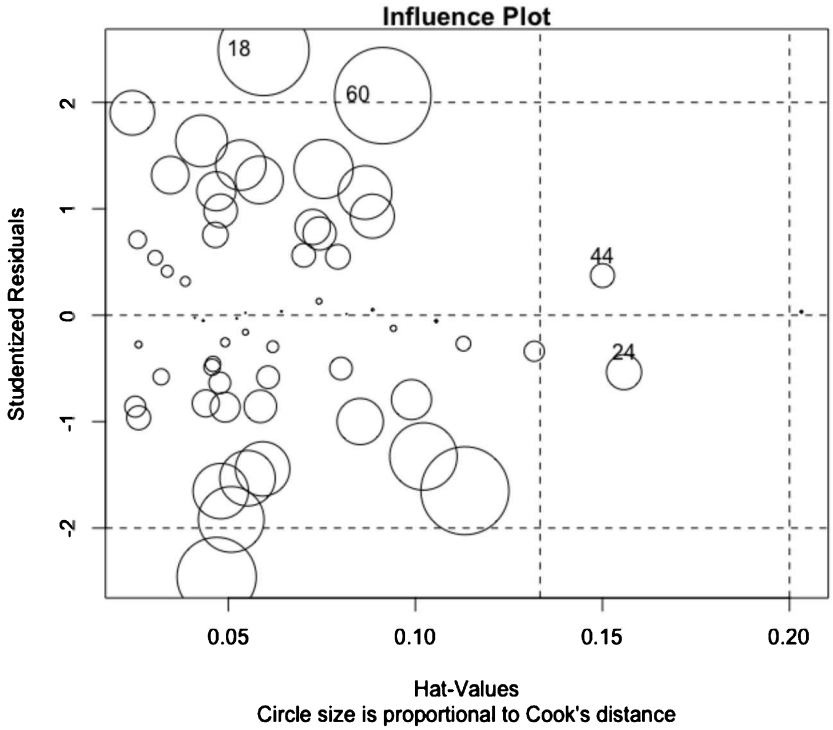


Figura 8.9.: Gráficos *Added variables*

Estadísticos e-Books & Papers

Figura 8.10.: Gráficos de influencia




```
influencePlot(fit, id.method="identify", main="Influence Plot",
sub="Circle size is proportional to Cook's distance")
```

8.7. Modelos con variables ficticias

Las variables ficticias son creadas por el investigador para introducir en un modelo aspectos cualitativos como el sexo, la estacionalidad, el estado civil, tipo de estudios, etc.

Para el tratamiento de las variables ficticias en un modelo econométrico conviene tener en cuenta las siguientes observaciones, que se agruparán en tres puntos: variables cualitativas con dos modalidades, con tres modalidades y contraste del cambio estructural.

Caso 8.6. ¿Existe discriminación salarial hacia la mujer en el mercado

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

laboral español?
Se utilizarán los datos de la *Encuesta de Estructura Salarial de España* del año 2002 en la que se han medido las siguientes variables:

- *salario* en miles de euros por hora.
- *educ*, formación en número de años de estudios.
- *mujer*, variable ficticia para el sexo que toma el valor 1=mujer, 0=hombre.
- *pequeña*, tamaño de la empresa. Variable ficticia que toma el valor 1 si la empresa tiene hasta 49 trabajadores y 0 en caso contrario.
- *mediana*, tamaño de la empresa. Variable ficticia que toma el valor 1 si la empresa tiene de 50 a 199 trabajadores y 0 en el resto de casos.
- *grande*, tamaño de la empresa. Variable ficticia que toma el valor 1 si la empresa tiene 200 trabajadores o más y 0 en caso contrario.

8.7.1. Variable ficticia con dos modalidades

Vamos a analizar cómo se puede incorporar la información dicotómica en los modelos de regresión. Considere el siguiente modelo para la determinación del salario por hora, en función de los años de educación (*educ*):

$$salario = \beta_0 + \beta_1 educ + \varepsilon \tag{8.70}$$

Para medir la discriminación salarial debida al género se introduce una variable ficticia (*mujer*) como variable independiente en el modelo definido anteriormente,

$$salario = \beta_0 + \delta_0 mujer + \beta_1 educ + \varepsilon \tag{8.71}$$

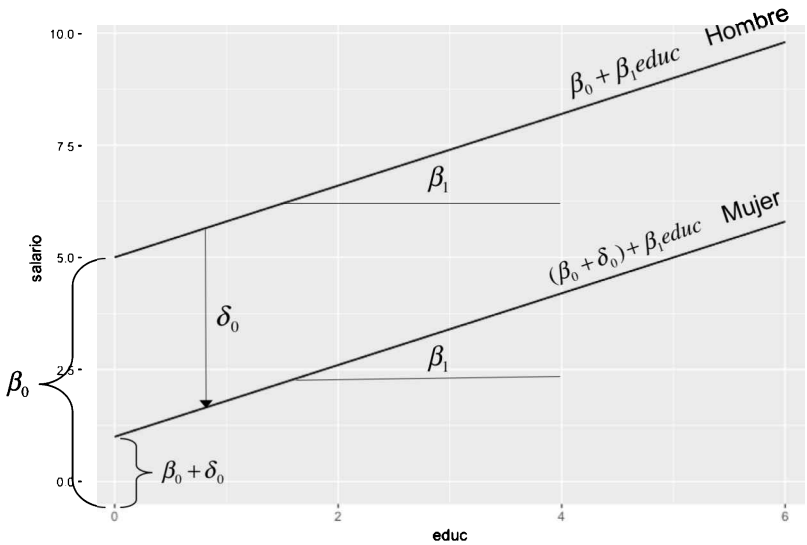
donde, como hemos comentado al presentar el caso:

$$mujer = \begin{cases} 1 & \text{si la persona es una mujer} \\ 0 & \text{si la persona es un hombre} \end{cases}$$

El atributo *sexo* tiene dos *categorías*: mujer y hombre. Como puede verse, la categoría mujer ha sido la incluida en el modelo; mientras que la categoría hombre, omitida, es la categoría de referencia. El modelo (8.71) se muestra en la figura 8.11, tomando $\delta_0 < 0$. La interpretación de δ_0 es la siguiente: δ_0 es la diferencia en el salario por hora entre mujeres y hombres, dado el mismo nivel de educación y el mismo término de perturbación ε. Así, el coeficiente δ_0 determina si existe una discriminación contra las mujeres o no. Si $\delta_0 < 0$, entonces, para el mismo nivel de otros factores (educación, en este caso), las mujeres ganan menos que los hombres en promedio. Suponiendo que la esperanza de la perturbación es cero, si se toman esperanzas en ambas categorías, se obtiene:

Estadísticos e-Books & Papers

Figura 8.11.: Variable cualitativa con dos modalidades e influencia aditiva



$$\begin{aligned} \mu_{salario|mujer} &= E\left(salario|mujer = 1, educ\right) = \beta_0 + \delta_0 + \beta_1 educ \\ \mu_{salario|hombre} &= E\left(salario|mujer = 0, educ\right) = \beta_0 + \beta_1 educ \end{aligned} \tag{8.72}$$

Como puede verse en (8.72), el término independiente para los hombres es β_0, y $\beta_0 + \delta_0$ para las mujeres. Gráficamente, como puede verse en la figura 8.11, hay un desplazamiento del término independiente, pero las líneas para hombres y mujeres son paralelas.

Estimamos el modelo mediante `lm{stats}` y los visualizamos en el cuadro 8.17 con `stargazer{stargazer}`.

```
fit1<-lm(log(salario)~mujer+educ,data=datos)
stargazer(fit1,type="text",report = "vct*")
```

Para responder a la pregunta planteada más arriba, tenemos que contrastar $H_0 : \delta_0 = 0$ contra de$H_1 : \delta_0 < 0$. Dado que el estadístico t es igual a –14,27 se rechaza la hipótesis nula para $\alpha = 0,01$. Es decir, hay evidencias de una discriminación en España contra la mujer en el año 2002.

8.7.2. Una variable cualitativa con más de dos modalidades

En el supuesto anterior hemos considerado un atributo (sexo) que tiene dos modalidades o categorías (mujer y hombre). Ahora vamos a considerar atributos

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 8.17.: Estimación del modelo $\ln(salario) = \beta_0 + \delta_0 mujer + \beta_1 educ + \varepsilon$

```
=================================================
                        Dependent variable:
                    ---------------------------
                           log(salario)
-------------------------------------------------
mujer                          -0.307
                         t = -14.268***

educ                            0.055
                          t = 21.768***

Constant                        1.731
                          t = 67.745***

-------------------------------------------------
Observations                    2,000
R2                              0.243
Adjusted R2                     0.242
Residual Std. Error      0.444 (df = 1997)
F Statistic          320.541*** (df = 2; 1997)
=================================================
Note:                 *p<0.1; **p<0.05; ***p<0.01
```

con más de dos categorías. En concreto, vamos a examinar un atributo con 3 categorías.

Para medir el impacto del tamaño de la empresa sobre el salario, podemos utilizar variables dicotómicas. Supongamos que las empresas se clasifican en tres grupos según su tamaño: pequeñas (hasta 49 trabajadores), medianas (de 50 a 199 trabajadores) y grandes (más de 199 trabajadores). Con esta información podemos construir 3 variables ficticias:

$$pequeña = \begin{cases} 1 & \text{hasta 49 trabajadores} \\ 0 & \text{en otros casos} \end{cases}$$
$$mediana = \begin{cases} 1 & \text{de 50 a 199 trabajadores} \\ 0 & \text{en otros casos} \end{cases}$$
$$grande = \begin{cases} 1 & \text{más de 199 trabajadores} \\ 0 & \text{en otros casos} \end{cases}$$

Si queremos explicar el salario por hora introduciendo en el modelo el tamaño de la empresa, es necesario omitir una de las categorías. En el siguiente modelo la categoría omitida son las empresas pequeñas:

$$salario = \beta_0 + \theta_1 mediana + \theta_2 grande + \beta_1 educ + \varepsilon \tag{8.73}$$

Vamos a ver ahora cuál es la media teórica en el modelo anterior, condicionada a que el individuo correspondiente esté en una empresa pequeña. Cuando

Estadísticos e-Books & Papers

se dice que una empresa es pequeña, es lo mismo que decir, tal como se han definido las variables ficticias, que no es ni mediana ni grande, es decir, que $grande_i = 0$ y que $mediana_i = 0$. Por lo tanto,

$$E\left(salario_i \mid mediana_i = 0, grande_i = 0\right) = \beta_0 + \beta_1 educ_i \tag{8.74}$$

lo que implica que la ordenada en el origen sea, para el caso de un individuo en una empresa pequeña y sin estudios, igual a β_0. Para los casos de empresas medianas o grandes, la media teórica del modelo, condicionada respectivamente a cada uno de estos casos es:

$$\begin{aligned} E\left(salario_i \mid mediana_i = 1, grande_i = 0\right) &= \beta_0 + \theta_1 + \beta_1 educ_i \\ E\left(salario_i \mid mediana_i = 0, grande_i = 1\right) &= \beta_0 + \theta_2 + \beta_1 educ_i \end{aligned} \tag{8.75}$$

La representación de este modelo da lugar a tres rectas paralelas con la misma pendiente β_1, pero con tres ordenadas diferentes: pequeña (β_0), mediana ($\beta_0 + \theta_1$) y grande ($\beta_0 + \theta_2$).

Obsérvese que al prescindir en la especificación del modelo de una de las tres variables ficticias (es decir, de una de las tres modalidades), se ha procedido de forma análoga a cuando se consideraban dos modalidades, ya que entonces no se incluyó una variable ficticia que tomara el valor 1 cuando el asalariado es hombre. No obstante, de forma alternativa se pueden incluir en la especificación del modelo las tres modalidades si en contrapartida se excluye el término independiente. Por lo tanto, una especificación alternativa a (8.73) con una parametrización diferente, pero completamente equivalente, es la siguiente:

$$salario = \theta_0 pequeña + \theta_1 mediana + \theta_2 grande + \beta_1 educ + \varepsilon \tag{8.76}$$

Como fácilmente puede comprobarse, tomando esperanzas condicionadas a cada una de las modalidades, se obtienen también tres rectas paralelas con las siguientes ordenadas para cada una de las modalidades: pequeña (θ_0), mediana (θ_1) y grande (θ_2).

Si estimamos el modelo (8.73) con la siguiente sintaxis, el cuadro 8.18 nos ofrece los resultados en su primera columna, la siguiente la utilizaremos para el contraste, que explicaremos inmediatamente.

```
fit2<-lm(log(salario)~mediana+grande+educ,data=datos)
fit3<-lm(log(salario)~educ,data=datos)
stargazer(fit2,fit3,type="text",report = "vct*")
anova(fit2)
anova(fit3)
```

Para responder a la pregunta inicial no haremos un contraste individual de θ_1 o θ_2. En vez de ello, contrastaremos conjuntamente si el tamaño de las empresas tiene una influencia significativa sobre el salario. Es decir, debemos

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 8.18.: Estimación del modelo $\ln(salario) = \beta_0 + \theta_1 mediana + \theta_2 grande + \beta_1 educ + \varepsilon$

```
=====================================================================
                                Dependent variable:
                    ---------------------------------------------------
                                     log(salario)
                              (1)                        (2)
---------------------------------------------------------------------
mediana                     0.162
                         t = 6.665***

grande                      0.281
                         t = 11.338***

educ                        0.048                      0.053
                         t = 18.591***              t = 19.931***

Constant                    1.566                      1.657
                         t = 57.718***              t = 63.105***

---------------------------------------------------------------------
Observations                2,000                      2,000
R2                          0.218                      0.166
Adjusted R2                 0.217                      0.165
Residual Std. Error     0.451 (df = 1996)          0.466 (df = 1998)
F Statistic         185.158*** (df = 3; 1996) 397.243*** (df = 1; 1998)
=====================================================================
Note:                                        *p<0.1; **p<0.05; ***p<0.01
```

contrastar si las medianas y grandes empresas tomadas conjuntamente tienen una influencia significativa en la determinación del salario. En este caso, las hipótesis nula y alternativa, tomando a (8.73) como el modelo no restringido, serán las siguientes:

$$H_0 : \theta_1 = \theta_2 = 0$$
$$H_1 : H_0 \text{ no es cierta}$$

El modelo restringido en este caso es el siguiente:

$$\ln(salario) = \beta_0 + \beta_1 educ + \varepsilon \tag{8.77}$$

que es precisamente el estimado en la segunda columna del cuadro 8.18 con la sintaxis que también mostrábamos. Por lo tanto, si retomamos el contraste para un subconjunto de parámetros que mostrábamos en la subsección 8.4.3, la expresión (8.48) genera el siguiente estadístico F:

$$F = \frac{(SSR_R - SSR_G)/r}{SSR_G/(n-1-k)} \tag{8.78}$$

y como necesitamos la suma de cuadrados de los residuos de los modelos restringidos y general, por esa razón los hemos solicitado en la sintaxis anterior mediante la opción `anova{stats}`. El cuadro 8.19 nos da esta información, con

Estadísticos e-Books & Papers

Cuadro 8.19.: Anova de los modelos original y restringido

```
> anova(fit2)
Analysis of Variance Table

Response: log(salario)
            Df Sum Sq Mean Sq  F value  Pr(>F)
mediana      1   0.98   0.982   4.8255 0.02816 *
grande       1  41.72  41.721 205.0171 < 2e-16 ***
educ         1  70.34  70.337 345.6325 < 2e-16 ***
Residuals 1996 406.19   0.204
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(fit3)
Analysis of Variance Table

Response: log(salario)
            Df Sum Sq Mean Sq F value    Pr(>F)
educ         1  86.11  86.113  397.24 < 2.2e-16 ***
Residuals 1998 433.12   0.217
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

la que ya podemos calcular el estadístico:

$$F = \frac{(SSR_R - SSR_G)/r}{SSR_G/(n-1-k)} = \frac{(433{,}12 - 406{,}19)/2}{406{,}19/(2000-1-3)} = 66{,}16$$

y como el valor crítico para $\alpha = 0{,}01$ puede obtenerse mediante R como sigue, podemos concluir que el tamaño de la empresa tiene una influencia significativa en la determinación de los salarios para los niveles usuales de significación en la medida en que el estadístico supera el valor crítico de $F^{0,01}_{2;1996} = 4{,}62$.

```
F<-((433.12-406.19)/2)/(406.19/(2000-1-3))
F

## [1] 66.16642

qf(0.99,2,2000-1-3)

## [1] 4.615812
```

8.7.3. La trampa de las variables ficticias

¿Qué ocurrirá si en (8.76) un investigador incluye además un término independiente? Anticipemos que entonces el investigador habrá caído en la trampa de las variables ficticias. Para ver en qué consiste esta trampa vamos a caer, aunque sea momentáneamente, en la misma, especificando el siguiente modelo:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

$$salario = \beta_0 + \theta_0 pequeña + \theta_1 mediana + \theta_2 grande + \beta_1 educ + \varepsilon \qquad (8.79)$$

Ahora, consideremos que tenemos una muestra de seis observaciones: las observaciones 1 y 2 corresponden a empresas pequeñas, la 3 y la 4 a medianas, y la 5 y 6 a grandes. En este caso, la matriz $\mathbf{X}$ de regresores tendría la siguiente configuración:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 & educ_1 \\ 1 & 1 & 0 & 0 & educ_2 \\ 1 & 0 & 1 & 0 & educ_3 \\ 1 & 0 & 1 & 0 & educ_4 \\ 1 & 0 & 0 & 1 & educ_5 \\ 1 & 0 & 0 & 1 & educ_6 \end{bmatrix} \qquad (8.80)$$

Como puede verse en la matriz $\mathbf{X}$, la columna 1 de esta matriz es igual a la suma de las columnas 2, 3 y 4. Por lo tanto, existe multicolinealidad perfecta, debido a la llamada *trampa de las variables ficticias*. Generalizando, si un atributo tiene g categorías, en el modelo únicamente tenemos que incluir $g-1$ variables ficticias junto con el término independiente. El término independiente para la categoría de referencia es el término independiente general del modelo, y el coeficiente de la variable ficticia de un grupo particular representa la diferencia estimada entre los términos independientes entre esa categoría y la categoría de referencia.

Así pues, la trampa de las variables ficticias es un error en que puede caer un investigador no experimentado por no haber especificado adecuadamente el modelo econométrico. Pero a la postre, el investigador, al ver que no puede estimar un modelo como el (8.79), reflexionando un poco, o consultando un manual de econometría, saldrá del atolladero especificando el modelo (8.73) o (8.76).

8.7.4. Interacción entre una variable cualitativa y una variable cuantitativa

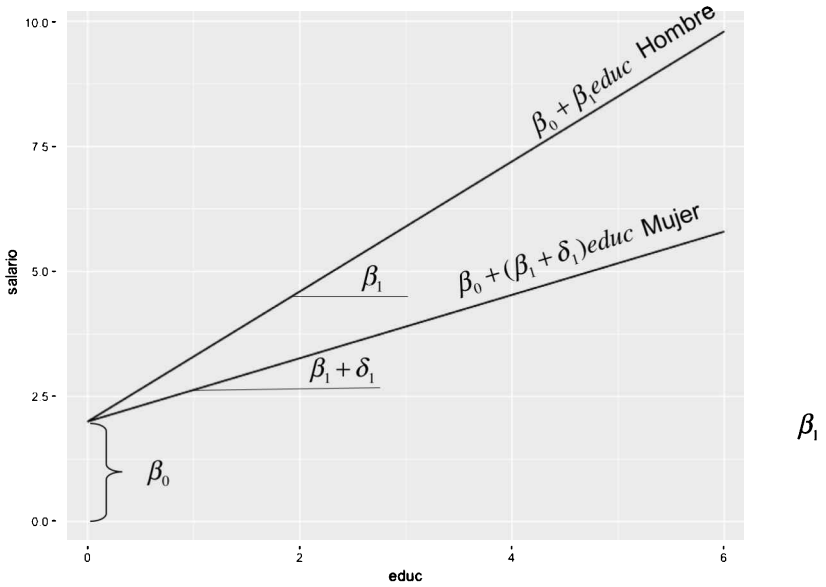
Hasta ahora, en los ejemplos sobre determinación del salario se ha utilizado una variable ficticia para desplazar el término independiente. Ahora bien, también se pueden utilizar las variables ficticias para interactuar con una variable explicativa continua. Por ejemplo, en el siguiente modelo la variable ficticia *mujer* interactúa con la variable continua *educ*:

$$salario = \beta_0 + \beta_1 educ + \delta_1 mujer \times educ + \varepsilon \qquad (8.81)$$

En el modelo (8.81), como puede verse en la figura 8.12, el término independiente es el mismo para hombres y para mujeres, pero la pendiente es mayor en hombres que en mujeres si asumimos que δ_1 es negativa.

Estadísticos e-Books & Papers

Figura 8.12.: Variable cualitativa con dos modalidades e influencia multiplicativa



En el modelo de (8.81), los rendimientos de un año adicional en educación dependen del género del asalariado. De hecho:

$$\frac{\partial salario}{\partial educ} = \begin{cases} \beta_1 + \delta_1 & \text{para mujeres} \\ \beta_1 & \text{para hombres} \end{cases} \tag{8.82}$$

Estimamos el modelo (8.81) donde la interacción la denotamos mediante el operador [:] y visualizamos la estimación con **stargazer{stargazer}** en el cuadro 8.20.

```
fit4<-lm(log(salario)~educ+mujer:educ,data=datos)
stargazer(fit4,type="text",report = "vct*")
```

En este caso necesitamos contrastar $H_0 : \delta_1 = 0$ contra $H_1 : \delta_1 < 0$. Dado que el estadístico t es –12,81, se rechaza la hipótesis nula en favor de la hipótesis alternativa para cualquier nivel de significación. Es decir, existe evidencia empírica de que el rendimiento de un año adicional de educación es mayor para hombres que para mujeres.

8.7.5. Contraste de cambio estructural

Hasta ahora hemos contrastado las hipótesis de que un parámetro, o un subconjunto de parámetros del modelo, son diferentes para dos grupos (mujeres

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 8.20.: Estimación del modelo $\ln(salario) = \beta_0 + \beta_1 educ + \delta_1 mujer \times educ + \varepsilon$

```
=================================================
                        Dependent variable:
                    ----------------------------
                           log(salario)
-------------------------------------------------
educ                           0.063
                            t = 23.693***

educ:mujer                    -0.027
                           t = -12.808***

Constant                       1.640
                            t = 64.849***

-------------------------------------------------
Observations                   2,000
R2                             0.229
Adjusted R2                    0.228
Residual Std. Error      0.448 (df = 1997)
F Statistic          296.851*** (df = 2; 1997)
=================================================
Note:                 *p<0.1; **p<0.05; ***p<0.01
```

y hombres, por ejemplo). Pero a veces queremos contrastar la hipótesis nula de si dos grupos tienen la misma función de regresión poblacional, frente a la alternativa de que no la tienen. En este procedimiento, contrastar si hay diferencias entre grupos consiste en realizar un contraste de significación conjunto de la variable ficticia que diferencia entre los dos grupos y de sus interacciones con todos los otros regresores. Por lo tanto, estimamos el modelo con (modelo no restringido) y sin (modelo restringido) la variable ficticia y todas sus interacciones.

De la estimación de ambas ecuaciones se obtiene el estadístico F, a través de la SSR, como hemos visto con anterioridad. En el siguiente modelo, para la determinación del salario, tanto el término independiente como la pendiente son diferentes para hombres y mujeres:

$$salario = \beta_0 + \delta_0 mujer + \beta_1 educ + \delta_1 mujer \times educ + \varepsilon \tag{8.83}$$

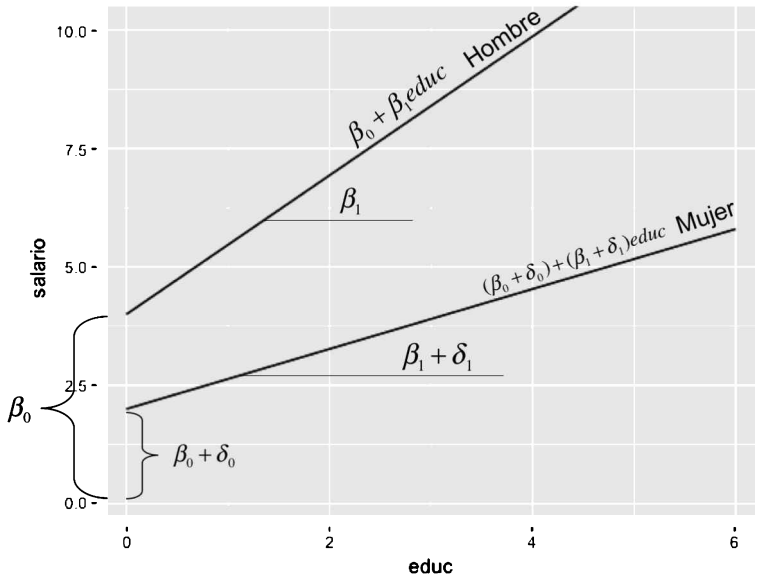
En la figura 8.13 ha sido representada la función de regresión poblacional de este modelo. Como puede verse, si mujer=1, se obtiene que:

$$salario = (\beta_0 + \delta_0) + (\beta_1 + \delta_1)\, educ + \varepsilon \tag{8.84}$$

Entonces, para las mujeres, el término independiente es $\beta_0 + \delta_0$ y la pendiente $\beta_1 + \delta_1$. Para mujer=0, obtenemos la ecuación (8.70). En este caso, para los hombres, el término independiente es β_0 y la pendienteβ_1. Por lo tanto, δ_0 mide la diferencia entre los términos independientes para mujeres y hombres y δ_1 mide a su vez la diferencia en el rendimiento de la educación entre mujeres

Estadísticos e-Books & Papers

Figura 8.13.: Pendiente y término independiente diferentes



y hombres. La figura 8.13 muestra un término independiente y una pendiente menores para mujeres que para hombres. Esto significa que las mujeres ganan menos que los hombres en todos los niveles de la educación, y que la brecha aumenta a medida que *educ* se hace más grande, es decir, un año adicional de educación tiene un rendimiento inferior para mujeres que para hombres.

La estimación de (8.83) es equivalente a la estimación de dos ecuaciones de salarios, uno para hombres y otro para las mujeres, por separado. La única diferencia es que (8.83) impone la misma varianza a los dos grupos, mientras que las regresiones por separado no lo hacen. Esta especificación del modelo es ideal, como veremos más adelante, para contrastar la igualdad de pendientes, la igualdad de términos independientes, o la igualdad tanto de términos independientes como de pendientes en los dos grupos.

Si los parámetros δ_0 y δ_1 son iguales a 0 en el modelo de (8.83), implica que la ecuación para la determinación de los salarios es la misma para hombres y mujeres. Entonces para responder a la cuestión planteada tomamos (8.83), pero expresando el salario en logaritmos, como el modelo no restringido. Las hipótesis nula y alternativa serán las siguientes:

$$H_0 : \delta_0 = \delta_1 = 0$$
$$H_1 : \ H_0 \text{ no es cierta}$$

Realizamos la estimación de los dos modelos con la siguiente sintaxis, que es la habitual, y añadimos el cuadro del análisis de la varianza para contar con las

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 8.21.: Estimación del modelo general y restringido

	Dependent variable:	
	log(salario)	
	(1)	(2)
educ	0.054***	0.053***
	(0.003)	(0.003)
mujer	-0.332***	
	(0.055)	
educ:mujer	0.003	
	(0.005)	
Constant	1.739***	1.657***
	(0.030)	(0.026)
Observations	2,000	2,000
R2	0.243	0.166
Adjusted R2	0.242	0.165
Residual Std. Error	0.444 (df = 1996)	0.466 (df = 1998)
F Statistic	213.694*** (df = 3; 1996)	397.243*** (df = 1; 1998)
Note:		*p<0.1; **p<0.05; ***p<0.01

SSR de los modelos general y restringido. El cuadro 8.21 ofrece la estimación de los modelos, y el 8.22, los resultados de análisis de la varianza.

```
fit5<-lm(log(salario)~educ+mujer+mujer:educ,data=datos)
fit6<-lm(log(salario)~educ,data=datos)
stargazer(fit5,fit6,type="text")
anova(fit5)
anova(fit6)
```

Con esta información podemos calcular el estadístico de contraste:

$$F = \frac{(SSR_R - SSR_G)/r}{SSR_G/(n-1-k)} = \frac{(433{,}12 - 393{,}00)/2}{393{,}00/(2000-1-3)} = 101{,}88$$

y como el valor crítico para $\alpha = 0{,}01$ puede obtenerse mediante `R` como sigue, podemos concluir que las ecuaciones de salarios para hombres y mujeres son distintas para los niveles usuales de significación en la medida en que el estadístico supera el valor crítico de $F^{0,01}_{2;1996} = 4{,}62$.

```
F<-((433.12-406.19)/2)/(406.19/(2000-1-3))
F

## [1] 66.16642
```


Estadísticos e-Books & Papers

Cuadro 8.22.: ANOVA del modelo general y restringido

```
> anova(fit5)
Analysis of Variance Table

Response: log(salario)
             Df Sum Sq Mean Sq  F value Pr(>F)
educ          1  86.11  86.113 437.3511 <2e-16 ***
mujer         1  40.07  40.066 203.4875 <2e-16 ***
educ:mujer    1   0.05   0.048   0.2435 0.6217
Residuals  1996 393.00   0.197
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(fit6)
Analysis of Variance Table

Response: log(salario)
            Df Sum Sq Mean Sq F value    Pr(>F)
educ         1  86.11  86.113  397.24 < 2.2e-16 ***
Residuals 1998 433.12   0.217
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
qf(0.99,2,2000-1-3)

## [1] 4.615812
```

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

9. Análisis discriminante

9.1. Introducción

El análisis discriminante se utiliza para explicar la pertenencia de distintos individuos a grupos —o poblaciones— alternativos a partir de los valores de un conjunto de variables que describen a los individuos a los que se pretende clasificar. Cada individuo puede pertenecer a un solo grupo. La pertenencia a uno u otro grupo se introduce en el análisis mediante una variable categórica que toma tantos valores como grupos existentes. En el análisis discriminante, esta variable juega el papel de variable dependiente.

A las variables que se utilizan para realizar la clasificación de los individuos las denominaremos variables *clasificadoras*. También se utilizan las denominaciones de variables criterio o variables predictoras, o la denominación genérica de variables explicativas. En el análisis discriminante, la información de las variables clasificadoras se sintetiza en unas funciones, denominadas funciones discriminantes, que son las que finalmente se utilizan en el proceso de clasificación.

El análisis discriminante se aplica para fines explicativos y predictivos. En la utilización explicativa se trata de determinar la contribución de cada variable clasificadora a la clasificación correcta de cada uno de los individuos. En una aplicación predictiva, se trata de determinar el grupo al que es más probable que pertenezca un individuo para el que se conocen los valores que toman las variables clasificadoras.

El análisis discriminante está muy relacionado con el análisis multivariante de la varianza con un factor, aunque el papel que juegan los distintos tipos de variables está invertido en uno y otro método. Así, en el análisis de la varianza, la variable categórica (el factor) es la variable explicativa, mientras que, en el análisis discriminante, la variable categórica es precisamente la variable dependiente.

Para aclarar las posibilidades de aplicación del análisis discriminante presentaremos un primer caso, el caso 9.1 que, al estar conformado por muy pocos casos, nos ayudará a explicar el funcionamiento interno de la herramienta, el cálculo de las funciones discriminantes, de los estadísticos, etc. Posteriormente, en un segundo caso más complejo, mostraremos las posibilidades de aplicación en un contexto más realista.

Caso 9.1 Análisis de préstamos fallidos del Banco de Ademuz

Cuando un banco concede un préstamo personal a un cliente se enfrenta a la doble posibilidad de que sea reintegrado o de que no lo sea. En este último

Estadísticos e-Books & Papers

caso, el préstamo será finalmente clasificado como fallido. Así pues, se pueden considerar dos grupos de clientes: clientes cumplidores y clientes fallidos. Como es obvio, si el banco conociera de antemano que una persona va a resultar fallida, no le concedería el préstamo en ningún caso. Sin embargo, puede utilizar la información existente en el banco sobre préstamos concedidos en el pasado en la concesión de préstamos futuros de forma que se evite o, al menos, se reduzca la posibilidad de conceder préstamos que después puedan resultar fallidos. Así, en los archivos del banco seguramente existirá información de las características de las personas a las que se ha concedido un préstamo, ya que el cliente, cuando realiza una petición de préstamo, debe facilitar datos acerca de cuestiones tales como ingresos, edad, sexo, situación familiar, antigüedad en su puesto de trabajo, régimen de tenencia de la vivienda, etc. Es muy posible que los clientes cumplidores tengan unas características distintas a las de los clientes fallidos. Utilizando estas características se trata de establecer unas funciones que clasifiquen lo más correctamente posible a los clientes a los que se les ha concedido un préstamo en cumplidores y fallidos (finalidad explicativa). Posteriormente, estas funciones se emplearán, en el caso de que haya realizado adecuadamente dicha clasificación, para determinar si se conceden o no los préstamos a futuros solicitantes (finalidad predictiva).

En el Banco de Ademuz se tiene información acerca de 16 clientes a los que se les concedió un préstamo de los llamados instantáneos por un importe de 12.000 euros cada uno. Una vez pasados tres años desde la concesión de los préstamos había 8 clientes, de ese grupo de 16, que fueron clasificados como fallidos, mientras que los otros 8 clientes son cumplidores, ya que reintegraron el préstamo. Para cada uno de los clientes se dispone de información sobre su patrimonio neto y deudas pendientes correspondientes al momento de la solicitud. En el cuadro 9.1 se ha reflejado esta información, así como la indicación de si resultaron o no fallidos.

Por otra parte, en la mesa del director del banco hay dos nuevas solicitudes de un préstamo instantáneo. El primer solicitante dispone de un patrimonio neto de 10,1 (decenas de miles de euros), con unas deudas pendientes de 6,8 (decenas de miles de euros). Para el segundo solicitante, los valores de estas variables son 9,7 y 2,2 respectivamente.

¿Cómo se realizaría en este caso la aplicación del análisis discriminante? Con la información sobre las variables de patrimonio neto y deudas pendientes se trata de construir una función discriminante que clasifique con los menores errores posibles a los clientes en dos grupos: fallidos y no fallidos. Si se obtienen buenos resultados en esta clasificación, en un paso posterior se utilizará la función discriminante construida para determinar si se concede el préstamo a los dos nuevos solicitantes. De esta forma, si a un nuevo solicitante se le clasifica a priori como fallido, no se le concederá el préstamo solicitado.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 9.1.: Datos sobre características de préstamos concedidos en el Banco de Ademuz (datos en decenas de miles de euros)

Fallidos (=1)			Fallidos (=2)		
Cliente	Patrim. neto	Deuda pendiente	Cliente	Patrim. neto	Deuda pendiente
1	1,3	4,1	9	5,2	1,0
2	3,7	6,9	10	9,8	4,2
3	5,0	3,0	11	9,0	4,8
4	5,9	6,5	12	12,0	2,0
5	7,1	5,4	13	6,3	5,2
6	4,0	2,7	14	8,7	1,1
7	7,9	7,6	15	11,1	4,1
8	5,1	3,8	16	9,9	1,6
Total	40,0	40,0	Total	72,0	24,0
Media	5,0	5,0	Media	9,0	3,0

9.2. Clasificación con dos grupos

En esta sección se va a ofrecer una visión descriptiva, y preferentemente intuitiva, de la aplicación del análisis discriminante a la clasificación de individuos en el caso de que se puedan clasificar solamente en dos grupos.

En una primera subsección se estudiarán los problemas de clasificación cuando existen dos grupos y una sola variable clasificadora. En la segunda subsección se introducirá una segunda variable clasificadora, ofreciéndose una solución gráfica. En la tercera subsección se expondrá la función discriminante de Fisher en el caso general de p variables clasificadoras.

9.2.1. Clasificación con dos grupos y una variable clasificadora

En esta subsección se va a considerar el supuesto más sencillo en el que existen dos poblaciones o grupos, a los que se denomina I y II, y una sola variable clasificadora, a la que se denomina X. El problema que se plantea es el de clasificar a cada individuo en el grupo correcto atendiendo al valor de la variable clasificadora.

En la figura 9.1 se han representado unas hipóteticas funciones de frecuencias de la variable X correspondientes a dos grupos. Tanto la configuración de distribución de frecuencias como la varianza son las mismas en los dos grupos; es decir, los dos grupos coinciden en todo excepto en su media. Como puede verse, las distribuciones de frecuencias, que se han representado de forma estilizada, están entrelazadas en el sentido de que se solapan. Precisamente, al existir solapamiento, se cometen o pueden cometerse errores de clasificación. De no existir tal solapamiento, el problema de clasificar a cada individuo en

Estadísticos e-Books & Papers

uno de los dos grupos sería trivial.

Dados los supuestos establecidos y denominando $\overline{X}_I$ y $\overline{X}_{II}$ a las medias de los grupos I y II respectivamente, el punto de intersección de las dos funciones corresponde al valor medio de $\overline{X}_I$ y $\overline{X}_{II}$. Es decir, este punto medio, al que denominaremos C, es igual a:

$$C = \frac{\overline{X}_I + \overline{X}_{II}}{2} \tag{9.1}$$

A la vista de la figura 9.1, parece razonable tomar el siguiente criterio para clasificar a un individuo i:

- Si $X_i < C$, se clasifica al individuo i en el grupo I.
- Si $X_i > C$, se clasifica al individuo i en el grupo II.

Por la función que cumple, designaremos a C como el *punto de corte discriminante* o simplemente como el *punto de corte*, en el sentido de que es el punto que se toma como referencia para clasificar a un individuo en uno u otro grupo.

Aplicando este criterio se cometen errores de clasificación, como puede comprobarse al examinar la figura 9.1. Así, el área tramada existente a la derecha de C recoge individuos pertenecientes al grupo I pero en los que $X_i > C$, es decir, son individuos del grupo I incorrectamente clasificados en el grupo II. Recíprocamente, el área rayada existente a la izquierda de C recoge individuos pertenecientes al grupo II, pero en los que $X_i < C$, es decir, son individuos del grupo II incorrectamente clasificados en el grupo I.

Vamos a aplicar las ideas anteriores al caso 9.1 de los préstamos del Banco de Ademuz. Como variable clasificadora se va a utilizar el patrimonio neto de los clientes, al que se denominará X_1. El grupo de clientes fallidos será el I, mientras que el II corresponderá a los no fallidos.

Las medias muestrales de los dos grupos son las siguientes:

$$\overline{X}_{1,I} = 5 \quad \overline{X}_{1,II} = 9$$

Aplicando (9.1) a este caso concreto se obtiene el siguiente valor para el punto de corte:

$$C_1 = \frac{\overline{X}_{1,I} + \overline{X}_{1,II}}{2} = \frac{5+9}{2} = 7$$

Por lo tanto, este punto de corte C_1 se utilizará para clasificar a los clientes a los que se les ha concedido préstamos en el Banco de Ademuz. Si el patrimonio neto es menor que 7 (decenas de miles de euros), se clasifica al cliente como fallido (I), mientras que se clasifica como no fallido (II) si el patrimonio es mayor a esa cifra. Con este criterio, ¿a cuántos clientes se clasifica incorrectamente? El examen de la columna de patrimonio neto en el cuadro 9.1 permite dar una respuesta inmediata a esta cuestión. En el cuadro 9.2 se ha reflejado el

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

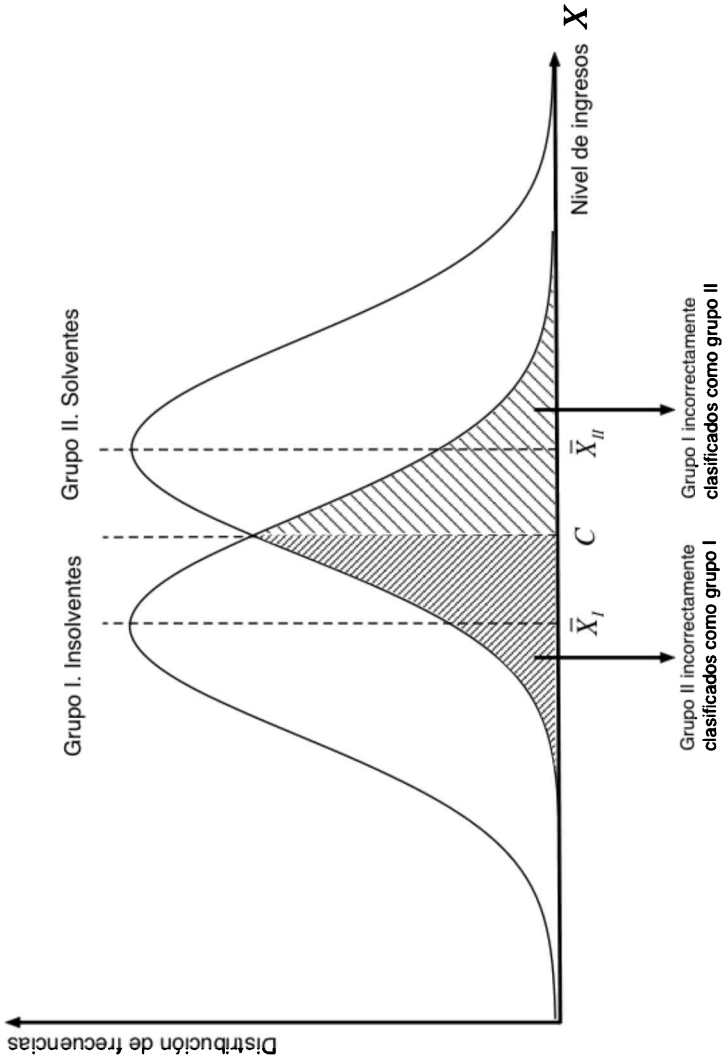


Figura 9.1.: Funciones de distribución hipotéticas de 2 grupos

Estadísticos e-Books & Papers

Cuadro 9.2.: Porcentaje de clasificaciones correctas e incorrectas utilizando la variable *patrimonio neto*

Situación real	Clasificado como		Total
	Fallido	No fallido	
Fallidos	6	2	8
No fallidos	2	6	8
Fallidos	75 %	25 %	100 %
No fallidos	25 %	75 %	100 %

Cuadro 9.3.: Porcentaje de clasificaciones correctas e incorrectas utilizando la variable *deudas pendientes*

Situación real	Clasificado como		Total
	Fallido	No fallido	
Fallidos	5	3	8
No fallidos	4	4	8
Fallidos	62,5 %	37,5 %	100 %
No fallidos	50 %	50 %	100 %

porcentaje de clasificaciones correctas e incorrectas en cada grupo. De un total de 16 clientes se han clasificado correctamente a 10, lo que equivale a un 75 % del total. En concreto, se han clasificado incorrectamente como no fallidos a los clientes 5 y 7. Por el contrario, se han clasificado erróneamente como fallidos a los clientes 9 y 13.

Vamos a utilizar ahora como variable clasificadora a la variable deudas pendientes, a la que designaremos por X_2, para ver si se obtienen o no mejores resultados que con X_1. Los datos sobre las deudas pendientes también aparecen en el cuadro 9.1. Las medias muestrales de las deudas pendientes de los dos grupos son:

$$\overline{X}_{2,I} = 5 \quad \overline{X}_{2,II} = 3$$

con lo que, ahora, el punto de corte es el siguiente:

$$C_2 = \frac{\overline{X}_{2,I} + \overline{X}_{2,II}}{2} = \frac{5+3}{2} = 4$$

Si las deudas pendientes son mayores que 4 (decenas de miles de euros), se clasifica al cliente como fallido (I), mientras que se clasifica como no fallido (II) si las deudas pendientes son menores que esa cifra.

En el cuadro 9.3 se ha reflejado el porcentaje de clasificaciones correctas e incorrectas para los fallidos y para los no fallidos.

Los resultados obtenidos con esta segunda variable clasificadora son peores,

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

ya que de los 16 casos solo se clasifican correctamente el 56,25 %. En concreto, se han clasificado incorrectamente como no fallidos a los clientes 3, 6 y 9. Por el contrario, se han clasificado erróneamente como fallidos a los clientes 10, 11, 13 y 15.

En lo que antecede se han utilizado dos variables clasificadoras, pero de forma separada. ¿Se puede mejorar el porcentaje de clientes clasificados correctamente si se utilizan las dos variables clasificadoras de forma conjunta? Esta cuestión será abordada en el siguiente punto, pero, en principio, cabe esperar que la clasificación mejore, ya que los casos que se clasifican incorrectamente son distintos para ambas variables, con la excepción hecha del caso número 13.

9.2.2. Clasificación con dos grupos y dos variables clasificadoras

Con objeto de obtener una representación gráfica, se va a considerar el caso de dos variables clasificadoras. En el siguiente apartado se realizará una exposición formalizada y utilizando p variables clasificadoras se obtendrá la función discriminante de Fisher (Fisher, 1936). En la figura 9.2 se han representado las elipses de concentración de los datos correspondientes a dos distribuciones de frecuencias bivariantes, en las que las variables X_1 y X_2 están correlacionadas positivamente.

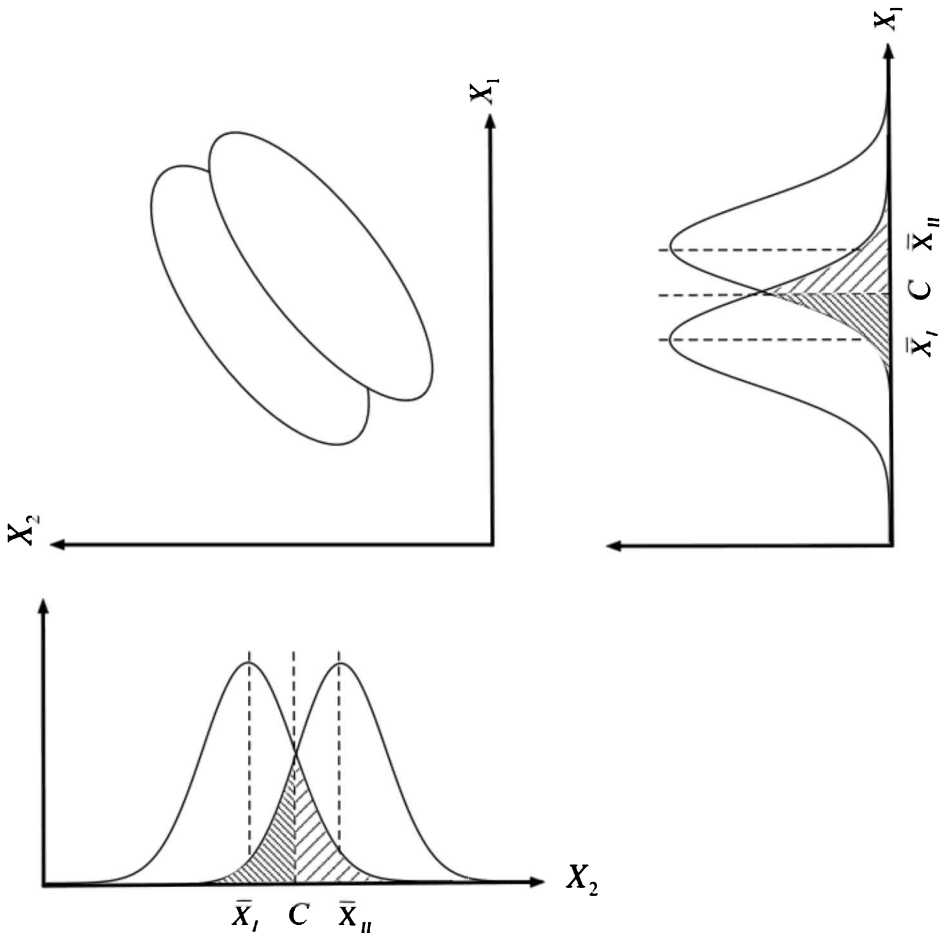
Las dos elipses de la figura 9.2 tienen el mismo tamaño y difieren en su centro. Debajo del eje X_1 se ha representado la proyección de las distribuciones de frecuencias bivariantes sobre este eje. Esta proyección nos ofrece las distribuciones univariantes marginales de la variable X_1. Como puede verse, las distribuciones de los dos grupos están solapadas, tal como ocurría en la figura 9.1.

Sobre el eje X_2 se han proyectado igualmente las distribuciones de frecuencias bivariantes, obteniéndose las correspondientes distribuciones de frecuencias marginales de la variable X_2. También en este caso aparecen muy solapadas las distribuciones marginales. Como ya hemos visto, cuanto mayor sea el grado de solapamiento, mayor será el porcentaje de individuos clasificados erróneamente.

A la vista de la figura 9.3 parece que se puede obtener una mejor función discriminante (es decir, que separe mejor) utilizando las dos variables conjuntamente. Proyectando las dos distribuciones de frecuencias sobre un eje oblicuo se pueden obtener distribuciones de frecuencias que estén menos solapadas que las distribuciones marginales. Variando la inclinación de este eje se obtienen distribuciones con un distinto grado de solapamiento. El eje con el que se logre un menor solapamiento será el óptimo. A este eje se le denomina *eje discriminante* y a las proyecciones de los valores de las variables X_1 y X_2 sobre dicho eje se las denomina *puntuaciones discriminantes*. La variable obtenida en la proyección, que será designada por D, es la función discriminante. En la figura 9.3 se ha representado, o se ha pretendido representar, esta proyección óptima. Por lo tanto, las distribuciones de frecuencias que aparecen sobre el eje discriminante son las correspondientes a la función discriminante.

Estadísticos e-Books & Papers

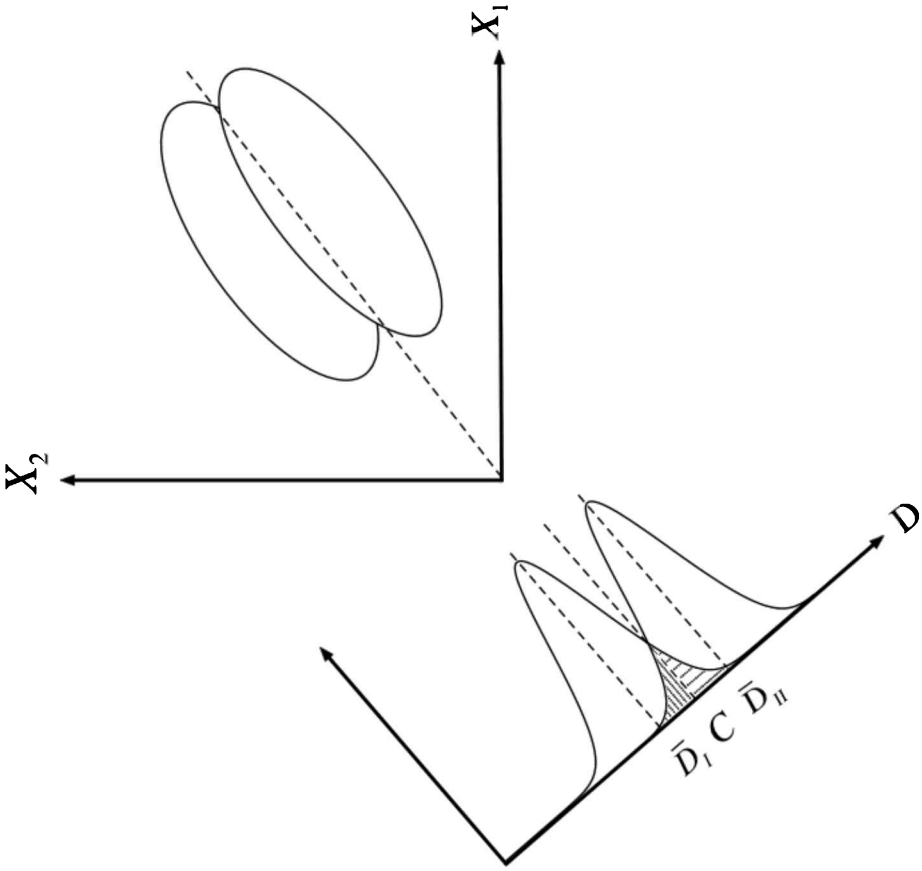
Figura 9.2.: Elipses de concentración de funciones de distribución de frecuencias y su proyección sobre los ejes X_1 y X_2



 © Ediciones Paraninfo

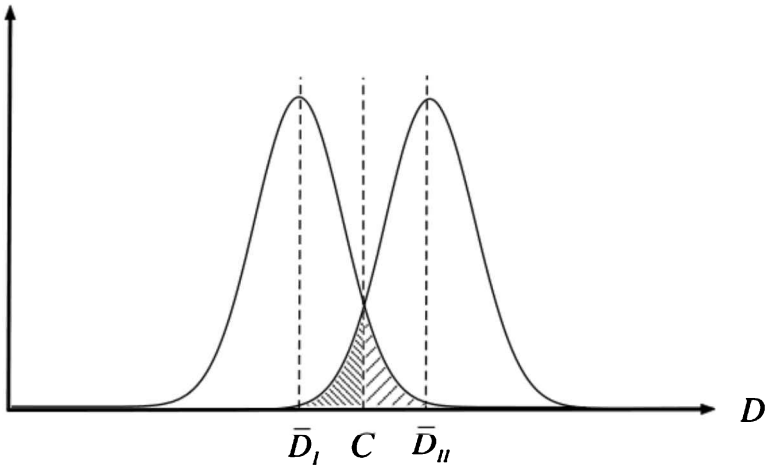
Estadísticos e-Books & Papers

Figura 9.3.: Elipses de concentración de funciones de distribución de frecuencias y su proyección sobre el eje discriminante



Estadísticos e-Books & Papers

Figura 9.4.: Funciones de distribución de frecuencias de las puntuaciones sobre el eje discriminante



En la figura 9.4 se han representado también las distribuciones de frecuencias de la función discriminante, pero tomando como eje de abscisas al eje que en la figura 9.3 aparecía como el eje discriminante. Para ello, se ha hecho simplemente una rotación. Comparando las distribuciones de frecuencias de la figura 9.4 con las distribuciones de frecuencias de la variable X en la figura 9.1 se observa que el solapamiento de la función discriminante es inferior al que presentan las distribuciones de frecuencias de la variable X. Esto implica que con las distribuciones de frecuencias de la función discriminante se consigue que el porcentaje de individuos clasificados correctamente sea mayor.

A. Función discriminante de Fisher

El problema que hemos ilustrado gráficamente para 2 variables clasificadoras, el estadístico Fisher lo resolvió analíticamente en 1936 para el caso general de p variables. La *función discriminante de Fisher* D se obtiene como función lineal de K variables explicativas X, es decir:

$$D = u_1X_1 + u_2X_2 + \cdots + u_KX_K \tag{9.2}$$

El problema planteado es la obtención de los coeficientes de ponderación u_j. Si consideramos que existen n observaciones, podemos expresar la función discriminante para las n observaciones:

$$D_i = u_1X_{1i} + u_2X_{2i} + \cdots + u_KX_{Ki} \quad i = 1, 2, \cdots, n \tag{9.3}$$

Así, D_i es la puntuación discriminante correspondiente a la observación i-

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

ésima. Expresando las variables explicativas en desviaciones respecto a la media, D_i también lo estará. La anterior relación se puede expresar en forma matricial para el conjunto de las observaciones:

$$\begin{bmatrix} D_1 \\ D_2 \\ \vdots \\ D_n \end{bmatrix} = \begin{bmatrix} X_{11} & X_{21} & \cdots & X_{K1} \\ X_{12} & X_{22} & \cdots & X_{K2} \\ \vdots & \vdots & \vdots & \vdots \\ X_{1n} & X_{2n} & \cdots & X_{Kn} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_K \end{bmatrix} \tag{9.4}$$

o en notación matricial compacta:

$$\mathbf{d} = \mathbf{X}\mathbf{u} \tag{9.5}$$

La variabilidad de la función discriminante (es decir, la suma de los cuadrados de las variables discriminantes en desviaciones respecto a su media) se puede expresar de la siguiente forma:

$$\mathbf{d}'\mathbf{d} = \mathbf{u}'\mathbf{X}'\mathbf{X}\mathbf{u} \tag{9.6}$$

El segundo miembro de (9.6) es una forma cuadrática de la matriz $\mathbf{X}'\mathbf{X}$. Esta matriz, al estar expresadas las variables en desviaciones respecto a la media, es la *matriz de suma de cuadrados y productos cruzados* (SCPC) total de las variables X. En el análisis multivariante de la varianza ya vimos que se calculaban también matrices de este tipo. La diferencia radica en que, en el análisis discriminante, estas matrices se obtienen para las variables explicativas, mientras que en el análisis multivariante de la varianza se obtienen para las variables dependientes. Más adelante volveremos sobre este punto. En cualquier caso, esta matriz se puede descomponer en la matriz de la SCPC *entre-grupos* y la SCPC *residual o intragrupos*. Utilizando para estas matrices la misma terminología que en el capítulo 7, la descomposición de $\mathbf{X}'\mathbf{X}$ puede expresarse así:

$$\mathbf{X}'\mathbf{X} = \mathbf{T} = \mathbf{F} + \mathbf{W} \tag{9.7}$$

donde $\mathbf{T}$, $\mathbf{F}$ y $\mathbf{W}$ son las matrices de SCPC total, entre-grupos e intragrupos respectivamente. Sustituyendo (9.6) en (9.7), se obtiene:

$$\mathbf{d}'\mathbf{d} = \mathbf{u}'\mathbf{T}\mathbf{u} = \mathbf{u}'\mathbf{F}\mathbf{u} + \mathbf{u}'\mathbf{W}\mathbf{u} \tag{9.8}$$

Obsérvese que en la expresión anterior $\mathbf{T}$, $\mathbf{F}$ y $\mathbf{W}$ se pueden calcular con los datos muestrales mientras que los coeficientes u_i están por determinar. Para su estimación, Fisher utilizó el siguiente criterio, que es totalmente racional. ¿Cuándo una función discriminante estará separando mejor los grupos? Cuando los grupos resultantes sean lo más homogéneos posible dentro de ellos y lo más diferentes posible unos de otros, es decir:

Estadísticos e-Books & Papers

$$\text{máx}\ \frac{\text{Variabilidad entre grupos}}{\text{Variabilidad intragrupos}} \tag{9.9}$$

Con este criterio se trata de determinar el eje discriminante de forma que las distribuciones proyectadas sobre el mismo estén lo más separadas posible entre sí (mayor variabilidad entre grupos) y, al mismo tiempo, que cada una de las distribuciones esté lo menos dispersa (menor variabilidad dentro de los grupos). Analíticamente, el criterio de Fisher se puede expresar de la siguiente forma:

$$\text{máx}\ \lambda = \frac{\mathbf{u'Fu}}{\mathbf{u'Wu}} \tag{9.10}$$

Como puede verse, se trata de que el primer término (*entre-grupos*) de (9.9) sea lo mayor posible en detrimento del segundo término (*intragrupos*). La forma de realizar esta maximización se verá con detalle en la sección 9.4 pero en estos momentos preferimos que el desarrollo de optimización necesario no oculte la intuición de lo que se está presentado.

La función discriminante de Fisher (9.2) suele ir acompañada del calificativo de *lineal*, debido a que se obtiene como combinación lineal de las variables originales.

Los coeficientes $u_1, u_2, ..., u_K$ (normalizados) que se obtienen en el proceso de maximización de (9.10) pueden contemplarse como un conjunto de cosenos que definen la situación del eje discriminante. Para esta interpretación, la normalización a la que nos referimos es que la suma de sus cuadrados sea la unidad; más adelante veremos otro tipo de normalización de los coeficientes de la función discriminante.

Las *puntuaciones discriminantes* son pues los valores que se obtienen al dar valores a $X_1, X_2, ..., X_K$ en la ecuación (9.2), y se corresponden con los valores obtenidos al proyectar cada punto del espacio K-dimensional de las variables originales sobre el eje discriminante.

Los *centros de gravedad o centroides* (es decir, el vector de medias) son los estadísticos básicos que resumen la información sobre los grupos. Las denominaciones que utilizaremos para designar a los centroides de los grupos I y II son las siguientes:

$$\overline{\mathbf{x}}_{\mathbf{I}} = \begin{bmatrix} \overline{X}_{1,I} \\ \overline{X}_{2,I} \\ \vdots \\ \overline{X}_{K,I} \end{bmatrix} \qquad \overline{\mathbf{x}}_{\mathbf{II}} = \begin{bmatrix} \overline{X}_{1,II} \\ \overline{X}_{2,II} \\ \vdots \\ \overline{X}_{K,II} \end{bmatrix} \tag{9.11}$$

Para calcular el *punto de corte discriminante* que nos va a permitir separar de la mejor manera posible los dos grupos, simplemente calculamos los valores que toma la función discriminante de Fisher en los dos centroides y luego promediamos esos valores. Por lo tanto sustituyendo los centroides (9.11) en la función de Fisher (9.2) se obtiene:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

$$\overline{D}_I = u_1\overline{X}_{1,I} + u_2\overline{X}_{2,I} + \cdots + u_p\overline{X}_{K,I} \tag{9.12}$$

$$\overline{D}_{II} = u_1\overline{X}_{1,II} + u_2\overline{X}_{2,II} + \cdots + u_p\overline{X}_{K,II} \tag{9.13}$$

y el *punto de corte discriminante* C *se calcula promediando* $\overline{D}_I$ *y* $\overline{D}_{II}$, es decir:

$$C = \frac{\overline{D}_I + \overline{D}_{II}}{2} \tag{9.14}$$

y esto nos da el criterio para clasificar al individuo i:

- Si la puntuación del individuo i en la función discriminante D_i —sustitución del valor de sus dos variables en (9.2)— es menor al punto de corte, se le clasifica en el grupo I: $D_i < C \longrightarrow$ grupo I.
- Si la puntuación del individuo i en la función discriminante D_i es mayor al punto de corte, se le clasifica en el grupo II: $D_i > C \longrightarrow$ grupo II.

En general, cuando se aplica el análisis discriminante se resta el valor de C a la función para hacer más fácil la aplicación del criterio todavía. Si $D_i - C < 0 \longrightarrow$ grupo I y $D_i - C > 0 \longrightarrow$ grupo II. De esta forma, la función discriminante viene dada por:

$$D - C = u_1X_1 + u_2X_2 + \cdots + u_KX_K - C \tag{9.15}$$

que delimita en el plano (X_1, X_2) a los grupos I y II. Es la recta que representábamos en trazos discontinuos en la figura 9.3.

Aplicación al caso 9.1. Apliquemos todo lo expuesto al caso 9.1 con el fin de ilustrar los cálculos. En primer lugar, la figura 9.5 muestra los datos de partida que ofrecíamos en el cuadro 9.1. En los mismos se observa que la separación de ambos no debería ser un problema y no deberían cometerse muchos errores de clasificación al trazar una recta que los separara. En segundo lugar es necesario que obtengamos la función discriminante de Fisher. En cuanto que es un proceso de optimización complicado que abordaremos matemáticamente en la sección 9.4, aplicamos directamente la función `lda{MASS}` para la estimación, que es uno de los paquetes de `R` que permite la aplicación del análisis discrimimante. La sintaxis no necesita ninguna aclaración.

```
library(ggplot2)
library(MASS)
datos<-Datos_9_1_Caso
fit<-lda(data=datos,fallido~patrneto+deudapen)
```

La salida (cuadro 9.4) nos ofrece los coeficientes de la función discriminante de Fisher u_1 (*patrneto*) y u_2 (*deudapen*). Por lo tanto, esta función es:

Estadísticos e-Books & Papers

Figura 9.5.: Representación gráfica de los datos de partida

Cuadro 9.4.: Coeficientes de la función discriminante de Fisher

```
Coefficients of linear discriminants:
                    LD1
patrneto  0.4224919
deudapen -0.3802226
```

$$D = 0{,}422X_1(patrneto) - 0{,}380X_2(deudapen) \tag{9.16}$$

Con esta función ya podemos estimar la función en los centroides de los grupos de acuerdo con (9.12) y (9.13) y calcular el punto de corte con (9.14). Todos estos cálculos se ilustran en el cuadro 9.5.

Vemos que la separación que provoca esta función en los planos es casi perfecta. Solo comete un error. El caso 13, que pertenece al grupo 2, aparece clasificado en el I debido a que:

$$D_{13} < C \rightarrow 0{,}422(6{,}3) - 0{,}380(5{,}2) = 0{,}6845 > 1{,}4366$$

o visto bajo la otra perspectiva:

$$D_{13} - C < 0 \rightarrow 0{,}422(6{,}3) - 0{,}380(5{,}2) - 1{,}4366 = -0{,}7520 < 0$$

Si representamos la recta $D - C = 0$ sobre los datos que contenía la figura (9.5) como hacemos en la figura 9.6 se ve claramente la separación y el error cometido, que es un porcentaje pequeño sobre el total. A efectos, simplemente,

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 9.5.: Aplicación de la función discriminante de Fisher

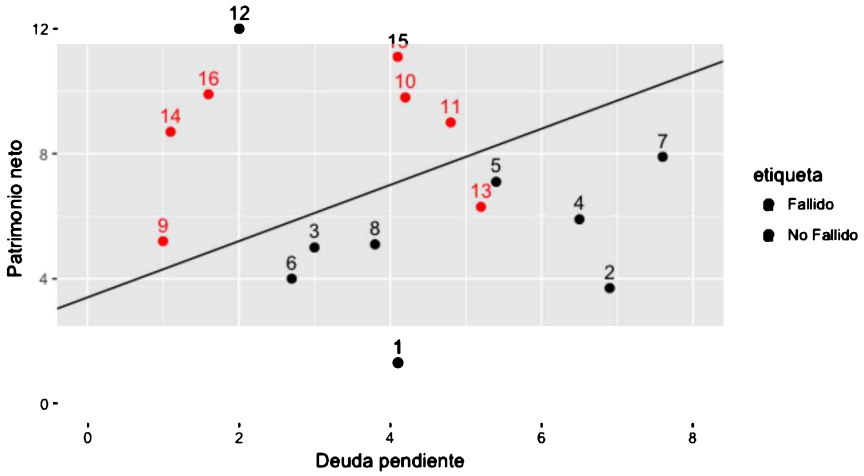
Cliente	Fallido	Patrimonio neto	Deuda pendiente	$D = 0{,}422X_1 - 0{,}380X_2$	$D - C$	Clasificación propuesta
		X_1	X_2			
1	1	1,3	4,1	-1,0097	-2,4462	1
2	1	3,7	6,9	-1,0603	-2,4969	1
3	1	5,0	3,0	0,9718	-0,4648	1
4	1	5,9	6,5	0,0213	-1,4153	1
5	1	7,1	5,4	0,9465	-0,4901	1
6	1	4,0	2,7	0,6634	-0,7732	1
7	1	7,9	7,6	0,4480	-0,9886	1
8	1	5,1	3,8	0,7099	-0,7267	1
9	2	5,2	1,0	1,8167	0,3802	2
10	2	9,8	4,2	2,5435	1,1069	2
11	2	9,0	4,8	1,9774	0,5408	2
12	2	12,0	2,0	4,3095	2,8729	2
13	2	6,3	5,2	0,6845	-0,7520	**1**
14	2	8,7	1,1	3,2574	1,8209	2
15	2	11,1	4,1	3,1307	1,6942	2
16	2	9,9	1,6	3,5743	2,1378	2

	Grupo	$\overline{X}_1$	$\overline{X}_1$	D_I D_{II}	C
Media	1	5	5	0,21134	1,43655
	2	9	3	2,66176	

© Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 9.6.: Función de Fisher sobre la representación gráfica de los datos de partida



de que se vea claramente que $D-C=0$ es la recta representada, la reescribimos del siguiente modo:

$$D-C=0{,}422X_1(patrneto)-0{,}380X_2(deudapen)-1{,}437=0$$

$$X_1=\frac{1{,}437}{0{,}422}+\frac{0{,}380}{0{,}422}X_2=3{,}400+0{,}900X_2$$

B. Calidad de la clasificación

Que el porcentaje de error es pequeño o, alternativamente, que el porcentaje de clasificaciones correctas es muy elevado, se observa fácilmente en el cuadro 9.6 que se conoce como *matriz de confusión*. Vemos que el 100 % de los integrantes del grupo I está correctamente clasificado y el 87,5 % de los del II. En solo un caso (12,5 %) la clasificación era incorrecta. En síntesis el 93,8 % (15/16) de los casos ha sido correctamente clasificado. Este cuadro no lo ofrece directamente la función `lda{MASS}` pero es fácil de obtener a partir de los datos que la misma proporciona. Como hemos visto en la sintaxis anterior, el resultado de estimar el análisis discriminante lo hemos guardado en el objeto `fit`. Pues bien `lda{MASS}` tiene una segunda función llamada `predict{MASS}` que permite clasificar nuevos casos de acuerdo con la función de Fisher obtenida. Si no se le indican los valores de las variables independientes, entonces clasifica los casos que se han utilizado para generar dicha función, es decir, la clasificación pronosticada que mostrábamos en la última columna del cuadro 9.5. Solo es

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 9.6.: Aciertos y errores en la clasificación

Grupo real	Grupo pronosticado 1	2	Total
1	8 100.0%	0 0.0%	8 50.0%
2	1 12.5%	7 87.5%	8 50.0%
Total	9	7	16

Cuadro 9.7.: Pertenencia al grupo predicha

```
$class
 [1] 1 1 1 1 1 1 1 1 2 2 2 2 1 2 2 2
Levels: 1 2
```

necesario aplicársela al objeto que contiene los resultados de la estimación, es decir, el que llamamos `fit`.

```
fit.p<-predict(fit)
```

De todos los componentes de la predicción, el que contiene la pertenencia al grupo predicha se llama `class`. Como vemos en el cuadro 9.7 coincide con la última columna del cuadro 9.5.

Pues bien, para obtener el cuadro 9.6 solo era necesario cruzar la pertenencia real que contiene el fichero datos en su primera columna (`datos[,1]`) con la pertenencia predicha en `class`. Para facilitar las cosas hemos añadido la pertenencia predicha al fichero original de datos.

```
# Obtenemos la pertenencia predicha y se anexa al fichero de datos
fit.p<-predict(fit)$class
datos<-data.frame(datos,fit.p)
# Obtenemos la tabla cruzada
CrossTable(datos[,1],datos$fit.p,digits=2,format="SPSS",
prop.c=FALSE, prop.chisq =FALSE,prop.t = FALSE,
dnn=c("Grupo real","Grupo pronosticado"))
```

En resumen y como puede verse ha mejorado sustancialmente el porcentaje de casos clasificados correctamente respecto a la utilización de las variables

Estadísticos e-Books & Papers

explicativas por separado.

C. Contraste de hipótesis

La obtención de la función discriminante la realizó Fisher aplicando un enfoque puramente descriptivo, como ha sido el seguido en las dos subsecciones anteriores. Sin embargo, si con el análisis discriminante se desea ir más lejos de la simple clasificación se requiere la formulación previa de hipótesis estadísticas. Formulando estas hipótesis se pueden abordar algunos temas de carácter inferencial y otros relativos al modelo poblacional. Los temas de tipo inferencial se refieren a diversos contrastes de significación sobre el modelo, así como contrastes utilizados en el proceso de selección de variables cuando el número de estas es muy grande y no se conocen a priori las variables que son relevantes en el análisis. Por otra parte, el cálculo de probabilidad de pertenencia a un grupo requiere que previamente se haya postulado algún modelo probabilístico de la población.

Hemos visto que la utilidad del análisis discriminante reside, principalmente, en la capacidad que tenga para explicar la separación entre los grupos. Esto exige, evidentemente, analizar si la función discriminante genera grupos predichos en los cuales las variables independientes toman valores medios significativamente distintos. En el caso de que la respuesta fuese negativa carecería de interés continuar con el análisis (qué peso tienen las variables en la explicación de la pertenencia a cada grupo), ya que significaría que las variables introducidas como variables clasificadoras no tienen una capacidad discriminante significativa. La hipótesis nula que debería poder rechazarse es la siguiente:

$$H_0 : \boldsymbol{\mu_1} = \boldsymbol{\mu_2} \tag{9.17}$$

es decir, que los vectores de medias de las variables independientes deben ser significativamente distintos.

Estamos ante un problema que es idéntico al que abordábamos en el MANOVA, siendo el factor el grupo de pertenencia y, por tanto, la solución es la misma en el estadístico planteado (Λ de Wilks) y su interpretación. Recordemos de la expresión (9.10) que la Λ de Wilks estaba en la base de generación de la función discriminante. Dado que la función `lda{MASS}` no realiza este contraste, lo hemos de realizar nosotros repitiendo exactamente lo visto en el capítulo 7. El cuadro 9.8 muestra como la hipótesis nula puede rechazarse para p<0,01, lo que implica que los dos grupos generados en la partición por la función discriminante son significativamente distintos respecto a las dos variables consideradas conjuntamente, lo que nos permite concluir que esta función es estadísticamente significativa (Sharma, 1996).

```
fit.manova<-manova(data=datos,cbind(deudapen,patrneto)~fallido)
summary((fit.manova),test="Wilks")
```

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 9.8.: Significatividad de la función discriminante

```
          Df   Wilks approx F num Df den Df   Pr(>F)
fallido    1 0.36825   11.151      2     13 0.001513 **
Residuals 14
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Cuadro 9.9.: Autovalores del MANOVA

```
$Eigenvalues
            [,1] [,2]
fallido 1.715578    0
```

D. Ajuste y tamaño del efecto

No es raro, como hemos señalado en repetidas ocasiones, sobre todo si se tienen muestras grandes, que la diferencia entre los dos grupos sea significativa sin que esta diferencia sea, a efectos prácticos, muy marcada. Esto nos ha llevado a introducir reiteradamente el concepto de tamaño del efecto. En el análisis discriminante el ajuste del modelo se obtiene como en el MANOVA a través de la η^2 que definíamos como la parte de la varianza total explicada por el factor de manera equivalente a la R^2 de una regresión y que obteníamos a partir de los autovalores (λ) de este modo (con una función discriminante solo tenemos un autovalor):

$$\eta^2 = \frac{\lambda}{1+\lambda}$$

El autovalor aparece en la salida del MANOVA antes efectuado (cuadro 9.9) por lo que el tamaño del efecto sería:

$$\eta^2 = \frac{1{,}715578}{1+1{,}715578} = 0{,}632$$

Por su parte, el tamaño del efecto se define como la raíz cuadrada de la η^2 y se conoce como *correlación canónica* (CR), en nuestro ejemplo:

$$CR = \eta = \sqrt{\frac{1{,}715578}{1+1{,}715578}} = 0{,}795$$

que es un tamaño del efecto grande de acuerdo con el criterio de Cohen (1988;1992).

E. Importancia relativa de las variables en la explicación de los grupos

Es conveniente conocer cuáles son las variables que tienen mayor poder discriminante en orden a clasificar a un individuo en uno de los dos grupos. Igual que

Estadísticos e-Books & Papers

Cuadro 9.10.: Matriz SCPC residual

```
$Residuals
         deudapen patrneto
deudapen    45.62    14.01
patrneto    14.01    66.70
```

en una regresión, los coeficientes no estandarizados de la función discriminante que hemos obtenido en el cuadro 9.4 no nos permiten la comparación directa del efecto relativo de una y otra variable. Sin embargo, los **coeficientes estandarizados** pueden calcularse directamente a partir de los no estandarizados mediante la expresión:

$$\hat{b}_j^* = \hat{b}_j \sqrt{\hat{s}_j^2} \tag{9.18}$$

donde $\hat{b}_j^*$ es el coeficiente estandarizado, $\hat{b}_j$, el no estandarizado del cuadro 9.4 y $\hat{s}_j^2$, la varianza de cada una de las variables obtenida de la matriz de *matriz de suma de cuadrados y productos cruzados* (SCPC) residual. La raíz cuadrada es porque, obviamente, lo que utilizamos es la desviación típica, pero queríamos dejar claro que lo que ofrecen las salidas es la varianza. Como pasaba con el MANOVA, en la estimación obtenemos la SCPC y en la diagonal no están las varianzas, pero, como veíamos en el capítulo 7, basta dividir por los grados de libertad para tener las varianzas. Mostramos primero la SCPC residual que ya se ha obtenido en el MANOVA (cuadro 9.10) y dividiendo la diagonal por los 14 grados de libertad que el cuadro 9.8 muestra que tienen los residuos, obtendríamos las varianzas.

```
summary((fit.manova),test="Wilks")$SS
```

Basta aplicar la expresión (9.18) para obtener los coeficientes estandarizados:

$$\sqrt{\hat{s}_1^2}(patrneto) = \sqrt{\frac{66{,}70}{14}} = \sqrt{4{,}764} = 2{,}1827$$

$$\sqrt{\hat{s}_2^2}(deudapen) = \sqrt{\frac{45{,}62}{14}} = \sqrt{3{,}2585} = 1{,}8052$$

$$\hat{b}_1^*(patrneto) = \hat{b}_1 \sqrt{\hat{s}_1^2} = 0{,}422 \times 2{,}1827 = 0{,}9222$$

$$\hat{b}_2^*(patrneto) = \hat{b}_2 \sqrt{\hat{s}_2^2} = -0{,}380 \times 1{,}8052 = -0{,}6864$$

Es fácil automatizar en R estos cálculos porque toda la información está contenida en los objetos estimados. En el resultado del cuadro 9.11 se puede

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 9.11.: Coeficientes estandarizados de la función discriminante

```
> std.b.patrneto
[1] 0.9221834
> std.b.deudapen
[1] -0.6863594
```

comprobar que son idénticos a los obtenidos manualmente.

```
std.b.deudapen=(sqrt(summary(fit.manova)$SS$Residuals[1,1]/14))
*fit$scaling[2,1]
std.b.patrneto=(sqrt(summary(fit.manova)$SS$Residuals[2,2]/14))
*fit$scaling[1,1]
```

Dado que estos coeficientes se han obtenido tipificando —o estandarizando— cada una de las variables clasificadoras para que tengan media 0 y desviación típica unitaria, se evitan los problemas de escala que pudieran existir entre las variables y, consecuentemente, la magnitud de los coeficientes estandarizados son un indicador de la importancia que tiene cada variable en el cálculo de la función discriminante. Así, la variable *patrneto* tiene una influencia que es casi un 50 % superior a la ejercida por la variable *deudapen*.

Sin embargo, es necesaria cierta precaución en el uso de los coeficientes estandarizados (Sharma, 1996) cuando las variables están muy correlacionadas entre sí, puesto que, dependiendo del grado de multicolinealidad, el efecto relativo puede estar muy distorsionado, como en una regresión. Dado que las puntuaciones discriminantes son una combinación lineal de las variables originales puede ser interesante saber qué significa exactamente la función discriminante viendo la contribución relativa que las variables tienen a su formación de manera análoga a las cargas factoriales de un análisis factorial. Es otra forma de ver la contribución relativa pero obviando el problema de la multicolinealidad. La carga, como en un factorial, sería simplemente la correlación entre la puntuación factorial y la variable y estará acotada entre +1 y −1. A estas cargas se las denomina **coeficientes de estructura** y la matriz que las recoge **matriz de estructura**. Pueden obtenerse del siguiente modo:

$$l_i = \sum_{j=1}^{p} r_{ij} b_j^* \tag{9.19}$$

donde l_i es la carga de la variable i, r_{ij} es la correlación entre la variable i y j en la matriz de correlaciones obtenida de la SCPC residual y b_j^* es el coeficiente estandarizado de la función discriminante que acabamos de obtener. Vemos que el único paso previo necesario es obtener la matriz de correlaciones a partir de la SCPC residual, lo que ya ilustramos en el MANOVA. Si lo repetimos ahora, solo hay que dividir la matriz SCPC residual por los grados de libertad de los

Estadísticos e-Books & Papers

Cuadro 9.12.: Matriz de correlaciones residual

```
> SCPC.residual.correlaciones
          deudapen  patrneto
deudapen 1.0000000 0.2539789
patrneto 0.2539789 1.0000000
```

Cuadro 9.13.: Matriz de estructura

```
> l.deudapen
[1] -0.4521443

> l.patrneto
[1] 0.7478626
```

residuos (14) y tenemos la matriz de covarianzas y esta la transformamos en una matriz de correlaciones mediante la función `cov2cor{stats}`. La matriz de correlaciones aparece en el cuadro 9.12 y con ella ya podemos calcular las cargas tomando los coeficientes estandarizados del cuadro 9.11.

```
SCPC.residual<-summary((fit.manova),test="Wilks")$SS$Residuals
SCPC.residual.varianzas<-SCPC.residual/14
SCPC.residual.correlaciones<-cov2cor(SCPC.residual.varianzas)
```

$$l_1(deudapen) = 1{,}0000 \times (-0{,}6864) + 0{,}2540 \times 0{,}9221 = -0{,}4521$$

$$l_2(patrneto) = 0{,}2540 \times (-0{,}6864) + 1{,}0000 \times 0{,}0{,}9221 = 0{,}7479$$

De nuevo es fácil automatizar los cálculos con `R` pudiéndose comprobar en el cuadro 9.13 que los resultados coinciden. Ahora podemos confirmar el resultado anterior de la mayor relevancia del patrimonio neto para explicar la separación en grupos frente a la deuda pendiente.

```
l.deudapen=SCPC.residual.correlaciones[1,1]*std.b.deudapen
+SCPC.residual.correlaciones[1,2]*std.b.patrneto
l.patrneto=SCPC.residual.correlaciones[1,2]*std.b.deudapen
+SCPC.residual.correlaciones[2,2]*std.b.patrneto
```

F. Supuestos del análisis discriminante

Las hipótesis estadísticas que se adoptan, análogas a las postuladas en el análisis multivariante de la varianza, se refieren tanto a la población como al proceso de obtención de la muestra. Las hipótesis sobre la población son las siguientes:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 9.14.: Test M de Barlett-Box

```
Box's M-test for Homogeneity of Covariance Matrices

data:  datos[2:3]
Chi-Sq (approx.) = 0.8035, df = 3, p-value = 0.8486
```

- La matriz de covarianzas de todos los grupos es la misma (hipótesis de homocedasticidad).
- Cada uno de los grupos sigue una distribución normal multivariante (hipótesis de normalidad).
- Se ha extraído una muestra aleatoria multivariante independiente en cada uno de los G grupos.

Los procedimientos son los mismos que explicamos en el tema 7, por lo que no reiteraremos el desarrollo de los estadísticos y nos limitaremos a aplicarlos al caso 9.1. Por lo tanto, la hipótesis nula para contrastar la homocedasticidad sería:

$$H_0 : \mathbf{\Sigma_1} = \mathbf{\Sigma_1} = \cdots = \mathbf{\Sigma_G} \tag{9.20}$$

Para determinar si la matriz de covarianzas es la misma para los distintos grupos se puede utilizar el contraste de Barlett-Box (Box, 1949), que utiliza el estadístico M definido en el tema 7. Su aplicación mediante la función `boxM{biotools}` se solicitaría de este modo. El resultado (cuadro) muestra que no puede rechazarse la hipótesis nula de igualdad de covarianzas, por lo que podemos asumir la interpretación de los resultados anteriores.

```
library(biotools)
boxM(datos[2:3],datos[,1])
```

El análisis de normalidad multivariante está disponible mediante la función `mshaphiro.test {mvnormtest}`. Esta función se aplica grupo a grupo, por lo que primero es necesario segregar la base de datos en los grupos, dos en nuestro ejemplo. Otro detalle es que el test necesita que las variables estén en filas, por lo que es necesario transponer la base de datos. Estos detalles se sintetizan en esta sintaxis:

```
library(mvnormtest)
#desagregamos la base por grupos
grupo1<-datos[1:8,2:3]
grupo2<-datos[9:16,2:3]
```


Estadísticos e-Books & Papers

Cuadro 9.15.: Contraste de la normalidad multivariante

```
> mshapiro.test(grupo1)

     Shapiro-Wilk normality test

data:  Z
W = 0.92454, p-value = 0.4678

> mshapiro.test(grupo2)

     Shapiro-Wilk normality test

data:  Z
W = 0.90647, p-value = 0.3299
```

```
#El test necesita las variables en filas, transponemos
grupo1<-t(grupo1)
grupo2<-t(grupo2)

#Ejecutamos el test
mshapiro.test(grupo1)
mshapiro.test(grupo2)
```

El cuadro 9.15 confirma que en ningún caso puede rechazarse la hipótesis de normalidad multivariante, reforzando de nuevo la viabilidad de interpretar los resultados como lo hicimos con anterioridad.

G. Predicción de nuevos casos

La función discriminante obtenida y mostrada en la ecuación (9.15) nos permite estimar si nuevos solicitantes tienen más probabilidad de ser de los que devolverán el préstamo o de los que no los harán. Recordemos que en la mesa del director del banco hay dos nuevas solicitudes de un préstamo instantáneo. El primer solicitante (1*) dispone de un patrimonio neto de 10,1 (decenas de miles de euros), con unas deudas pendientes de 6,8 (decenas de miles de euros). Para el segundo solicitante (2*), los valores de estas variables son 9,7 y 2,2 respectivamente. Aplicando la función obtenida a estos casos:

$$D_{1^*} - C < 0 \rightarrow 0{,}422(10{,}1) - 0{,}380(6{,}8) - 1{,}4366 = 0{,}2416 > 0$$

$$D_{2^*} - C < 0 \rightarrow 0{,}422(9{,}7) - 0{,}380(2{,}2) - 1{,}4366 = 1{,}8208 > 0$$

Como la puntuación es positiva en ambos casos, se clasifican los dos solicitantes en el grupo de los no fallidos, si bien hay que hacer notar que el segundo

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 9.16.: Predicción de nuevos casos

```
> predict(fit,newdata=nuevos.casos)$class
[1] 2 2
```

solicitante tiene una puntuación discriminante mucho más elevada.

La función `predict{MASS}` puede hacer automáticamente por nosotros este cálculo. Basta decirle que coja la información de la estimación realizada con los datos anteriores (`fit`) y la aplique a una nueva base de datos que tenga los casos que queremos predecir con sus valores de las variables independientes. La salida (cuadro 9.16) nos señala la `$class`, es decir, el grupo en el que quedan clasificados que son el segundo (no fallidos), tal como habíamos calculado. Más adelante veremos que también nos puede ofrecer la probabilidad de pertenencia a cada grupo pero, antes de eso, hemos de profundizar teóricamente en ese concepto.

```
d<-c(10.1,9.7)
p<-c(6.8,2.2)
nuevos.casos<-data.frame(d,p)
colnames(nuevos.casos)<-c("patrneto","deudapen")
predict(fit,newdata=nuevos.casos)$class
```

H. Cálculo de probabilidades de pertenencia a una población

Las funciones discriminantes del tipo (9.2) —o (9.15)— clasifican a los diferentes individuos en uno u otro grupo, pero no dan más información acerca de los individuos investigados.

Como hemos visto en el apartado anterior, las puntuaciones discriminantes matizan la clasificación dejando algo más claro si la clasificación en un grupo es sólida o no (por ejemplo, si $D - C$ estuviera muy cerca de cero, siendo positivo, clasificaríamos en el grupo II pero con cierta precaución). Con estas puntuaciones se puede clasificar a cada individuo, pero es interesante disponer además de información sobre la probabilidad de su pertenencia a cada grupo, ya que ello permitiría realizar análisis más matizados e incluir otras informaciones, tales como la información *a priori* o los costes que implica una clasificación errónea. Para realizar este tipo de cálculos se suelen asumir las hipótesis de normalidad y homocedasticidad, pero considerando que se conocen los parámetros poblacionales. Esta forma de proceder ocasiona ciertos problemas de los que nos ocuparemos posteriormente.

El cálculo de probabilidades se va a realizar en el contexto de la teoría de la decisión, que permite tener en cuenta tanto la probabilidad de pertenencia a un grupo como los costes de una clasificación errónea. La clasificación de los individuos se va a realizar utilizando el teorema de Bayes. La aplicación

Estadísticos e-Books & Papers

del teorema de Bayes permite el cálculo de las probabilidades *a posteriori* a partir de estas probabilidades *a priori* y de la información muestral contenida en las puntuaciones discriminantes. Considerando el caso general de G grupos, el teorema de Bayes establece que la probabilidad a posteriori de pertenencia a un grupo g con una puntuación discriminante D ($\Pr(g/D)$) es la siguiente:

$$\Pr(g/D) = \frac{\pi_g \times \Pr(D/G)}{\sum_{i=1}^{G} \pi_i \times \Pr(D/i)} \tag{9.21}$$

donde el elemento importante son las probabilidades a priori π_g. En todos los desarrollos anteriores hemos asumido que la probabilidad a priori de pertenecer al grupo de fallidos era la misma que la de pertenecer al de no fallidos: $\pi_1 = \pi_2 = 0{,}5$. Con ese supuesto hemos encontrado el punto de corte C que nos permitía la clasificación. Una alternativa habría sido asignar como probabilidad a priori la proporción de casos en la muestra que pertenecen al grupo I y al grupo II. En nuestro ejemplo son los mismos (8 y 8) por lo que nada habría cambiado. En la estimación mediante la función `lda{MASS}` si se quiere que se usen las proporciones derivadas de los tamaños muestrales iniciales en cada grupo, basta indicarlo así:

```
fit<-lda(data=datos,fallido~patrneto+deudapen,
prior=proportions)
```

Pues bien, sin las consideraciones posteriores que vamos a realizar inmediatamente sobre distintas probabilidades a priori o incorporación de restricciones de coste, `lda{MASS}` nos dará las probabilidades a posteriori con la función `predict{MASS}` y el modificador `$posterior`. Las probabilidades a posteriori aparecen en el cuadro 9.16. Este tipo de clasificación se conoce como **cálculo de probabilidades sin información a priori**. Nótese como el caso 13 tiene una probabilidad muy superior de pertenencia al grupo I pese a que en la realidad devolvió el préstamo.

```
fit<-lda(data=datos,fallido~patrneto+deudapen)
predict(fit)$posterior
```

Cálculo de probabilidades con información a priori. En ocasiones se dispone de información de la probabilidad a priori sobre pertenencia de un individuo a cada uno de los grupos. Así, en el caso de concesión de préstamos se puede tener la información de que los préstamos fallidos suponen un 10 % del total de préstamos concedidos a lo largo de los cinco años. Esta información puede tenerse en cuenta en el cálculo de las probabilidades a posteriori. El punto de corte discriminante se ve modificado de la siguiente forma:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 9.17.: Probabilidades a posteriori partiendo de probabilidades a priori iguales para los dos grupos

```
> predict(fit)$posterior
                1           2
1  0.9975131890 0.002486811
2  0.9978027741 0.002197226
3  0.7574696815 0.242530318
4  0.9697653007 0.030234699
5  0.7686769420 0.231323058
6  0.8692820481 0.130717952
7  0.9185178455 0.081482154
8  0.8557821467 0.144217853
9  0.2825993848 0.717400615
10 0.0622431333 0.937756867
11 0.2099549005 0.790045099
12 0.0008755318 0.999124468
13 0.8632722914 0.136727709
14 0.0114086116 0.988591388
15 0.0154972169 0.984502783
16 0.0052807262 0.994719274
```

$$C_p = \frac{\overline{D}_I + \overline{D}_{II}}{2} - \ln \frac{\pi_{II}}{\pi_I} \tag{9.22}$$

La *ratio* entre las proporciones debe establecerse de manera que el punto de corte se desplace hacia el grupo con menor probabilidad a priori. Al desplazar el punto de corte de esta forma, se tenderá a clasificar una proporción menor de individuos en el grupo con menor probabilidad a priori. Nótese que cuando las probabilidades a priori son iguales, la expresión anterior se convierte en el punto de corte C que mostrábamos en (9.14). La forma de solicitarlo es muy sencilla y el cuadro 9.18 muestra que los clientes 3, 5, 6 y 8, que antes estaban clasificados como fallidos, se clasifican ahora como cumplidores. Lo mismo ocurre con el cliente 13, que anteriormente estaba clasificado erróneamente como fallido siendo cumplidor.

.

```
fit<-lda(data=datos,fallido~patrneto+deudapen)
predict(fit,prior=c(0.1,0.9))$posterior
```

Cálculo de probabilidades con información a priori y consideración de costes. Hasta ahora no se ha considerado el coste que una clasificación errónea puede tener. En muchas aplicaciones el coste de clasificación errónea puede diferir para cada uno de los grupos. Veamos el ejemplo de concesión de préstamos. Cuando se está tratando de clasificar a los clientes en fallidos y no fallidos,

Estadísticos e-Books & Papers

Cuadro 9.18.: Probabilidades a posteriori partiendo de probabilidades a priori de 10 % para el grupo I y 90 % para el grupo II

```
                    1          2
1  9.780553e-01 0.02194472
2  9.805666e-01 0.01943344
3  2.576215e-01 0.74237850
4  7.808864e-01 0.21911362
5  2.696559e-01 0.73034409
6  4.249223e-01 0.57507772
7  5.560516e-01 0.44394841
8  3.973465e-01 0.60265351
9  4.193364e-02 0.95806636
10 7.320952e-03 0.99267905
11 2.868095e-02 0.97131905
12 9.735708e-05 0.99990264
13 4.122950e-01 0.58770496
14 1.280610e-03 0.99871939
15 1.745964e-03 0.99825404
16 5.895145e-04 0.99941049
```

las dos posibilidades de clasificación errónea son las siguientes: clasificar como fallido a un cliente cumplidor y clasificar como cumplidor a un fallido. En la primera de las posibilidades, el coste para el banco es dejar de percibir los intereses del préstamo y la posible pérdida de un cliente que en realidad es cumplidor. En cambio, en la segunda posibilidad, el coste para el banco es la pérdida de la cantidad prestada, ya que el cliente clasificado como cumplidor es realmente un cliente fallido. En principio, y bajo el criterio de una prudente administración financiera, parece que el segundo tipo de coste es superior al primero.

Cuando se introducen costes de clasificación no puede hablarse ya de cálculo de probabilidades a posteriori. No obstante se puede obtener un criterio para clasificar minimizando el **coste total de clasificación errónea**. Este coste total viene dado por la siguiente expresión:

$$\pi_I \times \Pr(II/I) \times \text{Coste}\,(II/I) + \pi_{II} \times \Pr(I/II) \times \text{Coste}\,(I/II) \tag{9.23}$$

Como puede verse en (9.23), cada probabilidad va multiplicada por el coste en que se incurre. Cuando se minimiza (9.23) bajo las hipótesis de normalidad y homocedasticidad, el punto de corte discriminante $C_{p,c}$ que se obtiene es el siguiente:

$$C_{p,c} = \frac{\overline{D}_I + \overline{D}_{II}}{2} - \ln \frac{\pi_{II} \times \text{Coste}\,(I/II)}{\pi_I \times \text{Coste}\,(II/I)} \tag{9.24}$$

Ahora vamos a calcular el punto de corte de nuestro ejemplo teniendo en

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

cuenta la información a priori e incorporando también los costes de la clasificación errónea. Con respecto a este último punto se adopta el supuesto de que el coste de clasificar como cumplidor a un cliente fallido es 20 veces superior al coste de clasificar como fallido a un cliente cumplidor. Es decir, se establece que la ratio Coste(II/I)/Coste(I/II) es igual a 20. Sustituyendo este valor en (9.24) se obtiene que:

$$C_{p,c} = \frac{0{,}211 + 2{,}662}{2} - \ln \frac{0{,}9}{0{,}1 \times 20} = 2{,}235 \tag{9.25}$$

Aunque la función `predict{MASS}` no tiene implementada la función de costes la reclasificación sería elemental, porque bastaría ir al cuadro 9.5 y cambiar el cálculo de la columna $D - C$ con el valor de $C_{p,c}$. Si el lector realiza el ejercicio, podrá comprobar que no altera la clasificación de ningún cliente respecto a la utilización del punto de corte inicial C. Es decir, la incorporación de los costes de clasificación errónea ha compensado, más o menos, la menor probabilidad a priori de ser un cliente fallido.

En todos los desarrollos anteriores se ha supuesto que las probabilidades son conocidas. En la práctica, sin embargo, se utilizan estadísticos muestrales en su lugar. El empleo de estadísticos muestrales tiene como consecuencia que se subestime la probabilidad de clasificación errónea, cometiéndose, por lo tanto, sesgos sistemáticos en la clasificación. Para disminuir estos sesgos se han propuesto, entre otros, dos procedimientos alternativos que pasamos a examinar.

Un procedimiento consiste en dividir la muestra total en dos submuestras, utilizando la primera muestra para estimar la función discriminante, mientras que la segunda se utiliza para su validación. Así, la potencia discriminatoria de la función vendrá determinada por el porcentaje de individuos clasificados correctamente en esta segunda submuestra.

El segundo procedimiento consiste en excluir un individuo del grupo I, calcular la función discriminante, y clasificar después al individuo que se ha excluido. Haciendo lo mismo con el resto de los individuos del grupo I, se estima la Pr(II/I) con el porcentaje de individiuos que han sido clasificados en el grupo II. Procediendo de la misma forma con los individuos del grupo II, se estima la Pr(I/II). A este segundo procedimiento se le conoce con la denominación *jacknife*.

9.3. Análisis discriminante con más de dos grupos

En esta sección se va a considerar la existencia de G grupos, generalizando los resultados obtenidos en la sección 9.2. En el caso de dos grupos se obtenía un solo eje discriminante. En el caso general, el número máximo de ejes discriminantes que se pueden obtener viene dado por:

$$\text{mín}(G - 1, K)$$

Estadísticos e-Books & Papers

De acuerdo con la expresión anterior, se pueden obtener hasta $G-1$ ejes discriminantes si el número de variables explicativas (K) es mayor o igual que $G-1$. Dado que en las aplicaciones prácticas, en general, suele existir un número abundante de variables explicativas, consideraremos a partir de ahora que no existen restricciones por esa parte, y que el número potencial de ejes es $G-1$.

Al análisis discriminante con más de dos grupos se le denomina ***análisis discriminante múltiple***. En esta sección se abordarán las dos cuestiones siguientes. En primer lugar, se examinará la obtención de las funciones discriminantes. En segundo lugar, se tratará la cuestión específica de la significatividad estadística de cada función discriminante.

Los contrastes y procedimientos aplicados en la sección 9.2 son fácilmente generalizables al caso de G grupos, no requiriendo aclaraciones especiales.

Antes de exponer las dos cuestiones que hemos indicado, se presenta un caso de concesión de préstamos en el que se consideran tres posibilidades, en lugar de las dos del Banco de Ademuz.

Caso 9.2. Concesión de préstamos en el Banco de Buñol

El director del Banco de Buñol se muestra preocupado por el aumento de clientes morosos y fallidos. Con objeto de paliar o reducir este problema encarga al investigador Joaquín Pastoriza la realización de un estudio que permita identificar con la mayor precisión posible aquellas solicitudes de préstamos que, probablemente, puedan llegar a convertirse en préstamos morosos o fallidos en el caso de que se concedieran.

Después de analizar la documentación existente en el Banco de Buñol, el investigador Joaquín Pastoriza solamente puede conseguir información relativamente completa acerca de 25 clientes a los que se ha concedido préstamos en los tres últimos años. Esta información es la que aparece en el cuadro 9.19. Los códigos que aparecen en la cabecera tienen el siguiente significado:

- *Categ*: grado de cumplimiento del cliente en el reintegro del préstamo. Toma los siguientes valores:
  - 1. Cliente cumplidor.
  - 2. Cliente moroso. Tiene pagos pendientes.
  - 3. Cliente fallido. Préstamo irrecuperable.
- *Ingresos*: ingresos anuales netos en decenas de euros.
- *Patrneto*: patrimonio neto en decenas de miles de euros.
- *Proviv*: variable dicotómica que toma los valores:
  - 1. Si es propietario de la vivienda que habita.
  - 0. En cualquier otra situación.
- *Casado*: variable dicotómica que toma los valores:
  - 1. Si está casado.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 9.19.: Datos sobre los préstamos concedidos en el Banco de Buñol

Cliente	Categ	Ingresos	Patrneto	Proviv	Casado	Salfij
1	1	5450	56	1	1	0
2	1	3100	34	1	0	1
3	1	2100	8	0	1	1
4	1	6200	45	1	0	1
5	1	975	10	0	1	1
6	1	1250	22	1	1	1
7	1	4900	15	0	1	1
8	1	8900	38	1	1	1
9	1	3350	54	0	1	1
10	1	5200	80	1	1	0
11	1	2850	11	1	1	1
12	1	6500	22	1	1	1
13	1	7600	36	1	1	1
14	2	4350	39	1	1	0
15	2	1350	8	0	1	1
16	2	2100	19	1	0	1
17	2	1450	25	1	0	0
18	2	6400	4	0	1	1
19	2	3800	19	1	1	0
20	2	2450	24	0	1	0
21	2	3300	7	0	1	1
22	2	850	12	0	1	0
23	2	1200	5	1	1	1
24	2	1850	6	1	0	0
25	2	2650	25	0	0	0

  - 0. Cualquier otra situación.

- *Salfij*: variable dicotómica que toma los valores:
  - 1. Si es asalariado con contrato fijo.
  - 0. Cualquier otra situación.

El problema que se plantea es construir funciones discriminantes que permitan clasificar, con los menores errores posibles, a los clientes en los diferentes grupos. Si se obtienen buenos resultados, estas funciones discriminantes se podrán utilizar para analizar si se concede un préstamo o no a un futuro peticionario.

Estadísticos e-Books & Papers

9.3.1. Obtención de las funciones discriminantes

Bajo el supuesto de que el número de variables explicativas (K) es mayor o igual que $G-1$, se pueden obtener potencialmente $G-1$ funciones discriminantes.

Cada una de las funciones discriminantes D_i se obtiene como función lineal de las K variables explicativas X, es decir:

$$D_i = u_{i1}X_1 + u_{i2}X_2 + \cdots + u_{iK}X_K \tag{9.26}$$

Los $G-1$ ejes discriminantes vienen definidos, respectivamente, por los vectores $\mathbf{u_1}, \mathbf{u_2}, ..., \mathbf{u_{G-1}}$, que toman las siguientes expresiones:

$$\mathbf{u_1} = \begin{bmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{1K} \end{bmatrix} \quad \mathbf{u_2} = \begin{bmatrix} u_{21} \\ u_{22} \\ \vdots \\ u_{2K} \end{bmatrix} \cdots \mathbf{u_{G-1}} = \begin{bmatrix} u_{G-1,1} \\ u_{G-1,2} \\ \vdots \\ u_{G_1,K} \end{bmatrix}$$

Para la obtención del primer eje discriminante se maximiza la ratio variabilidad *entre-grupos*/variabilidad *intragrupos* dada en (9.9), es decir:

$$\text{máx } \lambda_1 = \frac{\mathbf{u_1'Fu_1}}{\mathbf{u_1'Wu_1}} \tag{9.27}$$

La solución a este problema se obtiene derivando la ratio anterior e igualando a $\mathbf{0}$, es decir:

$$\frac{\partial \lambda_1}{\partial \mathbf{u_1}} = \mathbf{0} \tag{9.28}$$

Derivando en (9.27) se obtiene que:

$$\frac{\partial \lambda_1}{\partial \mathbf{u_1}} = \frac{2\mathbf{Fu_1}\left(\mathbf{u_1'Wu_1}\right) - 2\mathbf{Wu_1}\left(\mathbf{u_1'Fu_1}\right)}{\left(\mathbf{u_1'Wu_1}\right)^2} = \mathbf{0} \tag{9.29}$$

es decir,

$$2\mathbf{Fu_1}\left(\mathbf{u_1'Wu_1}\right) - 2\mathbf{Wu_1}\left(\mathbf{u_1'Fu_1}\right) = \mathbf{0} \tag{9.30}$$

Operando en la expresión anterior y teniendo en cuenta (9.27), se puede establecer que:

$$\frac{2\mathbf{Fu_1}}{2\mathbf{Wu_1}} = \frac{\mathbf{u_1'Fu_1}}{\mathbf{u_1'Wu_1}} = \lambda_1 \tag{9.31}$$

Igualando el primero y el último miembro de la expresión anterior, se obtiene:

$$\mathbf{Fu_1} = \mathbf{Wu_1}\lambda_1 \tag{9.32}$$

Premultiplicando ambos miembros de (9.32) por $\mathbf{W^{-1}}$, bajo el supuesto de que $\mathbf{W}$ es no singular, se llega a la expresión:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

$$\mathbf{W^{-1}Fu_1} = \lambda_1 \mathbf{u_1} \tag{9.33}$$

La obtención del vector $\mathbf{u_1}$ es pues un problema de cálculo de un vector característico (autovector) asociado a la matriz no simétrica $\mathbf{W^{-1}F}$. De las raíces características (autovalores) que se obtienen al resolver la ecuación (9.33) se retiene la mayor, ya que λ_1 es precisamente la ratio que pretendemos maximizar y $\mathbf{u_1}$ es el autovector asociado a dicho autovalor.

Dado que λ_1 es la ratio (9.27), nos medirá, una vez calculado, el poder discriminante del primer eje discriminante. El resto de los ejes discriminantes son otros vectores característicos de la matriz $\mathbf{W^{-1}F}$, ordenados según el orden decreciente de las raíces características. Así, el segundo eje tendrá menos poder discriminante que el primero, pero más que cualquiera de los restantes.

Puesto que la matriz $\mathbf{W^{-1}F}$ no es simétrica, esto implicará que, en general, los ejes discriminantes no serán ortogonales, es decir, no serán perpendiculares entre sí.

9.3.2. Contrastes de significación

En el análisis discriminante múltiple se plantean contrastes específicos para determinar si cada uno de los valores λ_i es estadísticamente significativo, es decir, para determinar si cada uno de los valores λ_i contribuye o no a la discriminación entre los diferentes grupos.

Este tipo de contrastes se realiza a partir del estadístico V de Barlett. El estadístico V es una función de la Λ de Wilks y se aproxima a una ji-cuadrado; tiene interés en el análisis discriminante por su descomponibilidad. En el capítulo 7 se utilizó otra función de la Λ de Wilks para realizar el contraste de igualdad de medias. Concretamente se aplicó la aproximación de Rao por la ventaja de tener distribución exacta F en algunos casos. La expresión del estadistíco V de Barlett es la siguiente:

$$V = -\left\{ n - 1 - \frac{K+G}{2} \right\} \ln \Lambda \tag{9.34}$$

estadístico que se distribuye como una $\chi^2_{K(G-1)}$. Este estadístico se utiliza en el análisis multivariante para contrastar las siguientes hipótesis:

$$H_0 : \boldsymbol{\mu_1} = \boldsymbol{\mu_2} = \cdots = \boldsymbol{\mu_G} \tag{9.35}$$

En el contexto del análisis multivariante de la varianza con un factor se contrastaba esta hipótesis para determinar si el factor (la variable categórica con G grupos) explicaba la variabilidad del vector de variables dependientes de forma significativa. En el contexto del análisis discriminante múltiple la hipótesis a contrastar sigue siendo la misma, pero los papeles se han invertido. Ahora se realiza el contraste para tratar de dar respuesta a la siguiente pregunta: ¿las K variables clasificadoras contribuyen de forma significativa a discriminar entre los G grupos? Si no se rechaza la hipótesis nula (9.35), no se debería continuar

Estadísticos e-Books & Papers

el análisis, puesto que las variables clasificadoras utilizadas en la investigación no tienen ningún poder discriminante significativo.

Examinemos ahora el poder discriminante de cada uno de los ejes que se construyen en el análisis discriminante. El estadístico V se puede descomponer a partir de la descomposición de la Λ de Wilks en productos en los que aparecen las raíces características λ_j. De acuerdo con su definición, el recíproco de Λ se puede expresar así:

$$\frac{1}{\Lambda} = \frac{|\mathbf{T}|}{|\mathbf{W}|} = |\mathbf{W}|^{-1}|\mathbf{T}| = \left|\mathbf{W^{-1}T}\right| = \left|\mathbf{W^{-1}}(\mathbf{W}+\mathbf{F})\right| = \left|\mathbf{I}+\mathbf{W^{-1}}+\mathbf{F}\right| \tag{9.36}$$

En la deducción anterior se ha tenido en cuenta que el determinante del producto de dos matrices es igual al producto de sus determinantes. Por otra parte, teniendo en cuenta que el determinante de una matriz es igual al producto de sus raíces características, se obtiene que:

$$\frac{1}{\Lambda} = (1+\lambda_1)(1+\lambda_2)\cdots(1+\lambda_{G-1}) \tag{9.37}$$

Sustituyendo (9.37) en (9.34), se obtiene la expresión alternativa del estadístico V de Barlett:

$$V = \left\{n-1-\frac{K+G}{2}\right\}\sum_{g=1}^{G-1}\ln(1+\lambda_g) \tag{9.38}$$

Si se rechaza la hipótesis nula (9.35), significa que al menos uno de los ejes discriminantes es estadísticamente significativo. Esto implica a su vez que el primer eje es estadísticamente significativo, debido a que es precisamente el que tiene mayor poder discriminante. Pero ¿qué pasa con la significatividad del resto de los ejes?

En el caso de que se acepte la hipótesis de que el primer eje discriminante es significativo, se pasaría a contrastar la significatividad conjunta del resto de los ejes discriminantes, utilizando el siguiente estadístico:

$$V = \left\{n-1-\frac{K+G}{2}\right\}\sum_{g=2}^{G-1}\ln(1+\lambda_g) \tag{9.39}$$

Como puede verse en la V anterior no figura la raíz característica λ_1. De forma general se puede establecer el siguiente esquema de contrastación secuencial mediante el estadístico V:

$$V_j = \left\{n-1-\frac{K+G}{2}\right\}\sum_{g=j+1}^{G-1}\ln(1+\lambda_g) \;\; j=0,1,2,\cdots,G-2 \tag{9.40}$$

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 9.20.: Medias de las variables explicativas en los tres grupos

```
Group means:
  ingresos patrneto    proviv    casado   salfijo
1 4873.077 35.53846 0.7692308 0.8461538 0.8461538
2 3128.571 17.71429 0.4285714 0.7142857 0.7142857
3 1970.000 11.00000 0.4000000 0.6000000 0.2000000
```

donde el estadístico V_j se distribuirá como una $\chi^2_{(K-j)(G-j-1)}$.

Así, en este proceso secuencial se van eliminando del estadístico V las raíces características que van resultando significativas, deteniendo el proceso cuando se acepte la hipótesis nula de no significatividad de los ejes discriminantes que queden por contrastar.

En (9.40) cuando se da a j el valor 0, se obtiene V_0, que se corresponde con la expresión (9.34). Cuando se da a j el valor 1, se está contrastando la significatividad de todos los ejes excluido el primero.

Como una medida descriptiva complementaria de este contraste se suele calcular el porcentaje acumulativo de la varianza después de la incorporación de cada nueva función discriminante.

Apliquemos todo lo expuesto al caso del Banco de Buñol. La estimación de un análisis discriminante con más de dos grupos tiene exactamente la misma sintaxis en la función `lda{MASS}` que cuando hay solo un grupo. Lo primero que nos ofrece la salida (cuadro 9.20) son la medias de las variables explicativas en cada uno de los tres grupos de la variable dependiente.

```
fit<-lda(data=datos,
categ~ingresos+patrneto+proviv+casado+salfijo)
```

Las medias de las 5 variables introducidas en el análisis son mayores en la categoría de cumplidores que en las otras dos categorías. Así, los clientes cumplidores, en relación con los otros dos grupos, tienen unos mayores ingresos y un mayor patrimonio, son propietarios de la vivienda en que habitan, están casados y son asalariados con contrato fijo.

Las dos funciones discriminantes que se necesitan ahora para separar los tres grupos aparecen en el cuadro 9.21. Aplicando esas funciones discriminantes con el mismo esquema que vimos para el caso de dos grupos, estos se han clasificado. Para saber la calidad de la clasificación, es decir, cuántos de los casos originales se habrían clasificado de manera correcta, procedemos como hicimos para el caso de dos grupos. Guardaremos la pertenencia estimada tras estimar el discriminante y lo cruzaremos con la variable *categ* que tiene la clasificación original:

```
# Obtenemos la pertenencia predicha
#y la anexamos al fichero de datos
```


Estadísticos e-Books & Papers

Cuadro 9.21.: Funciones discriminantes

```
Coefficients of linear discriminants:
                    LD1           LD2
ingresos -0.0001758512  5.423942e-05
patrneto -0.0688731887  1.184634e-02
proviv   -0.5905758045  9.217797e-01
casado   -1.0053538699  1.315451e-01
salfijo  -2.8156685539 -1.500319e+00
```

Cuadro 9.22.: Matriz de confusión

Grupo real	Grupo pronosticado 1	2	3	Total
1	12 92.31%	1 7.69%	0 0.00%	13 52.00%
2	0 0.00%	6 85.71%	1 14.29%	7 28.00%
3	0 0.00%	1 20.00%	4 80.00%	5 20.00%
Total	12	8	5	25

```
fit.p<-predict(fit)$class
datos<-data.frame(datos,fit.p)
# Obtenemos la tabla cruzada
CrossTable(datos[,1],datos$fit.p,digits=2,format="SPSS",
prop.c=FALSE, prop.chisq =FALSE,prop.t = FALSE,
dnn=c("Grupo real","Grupo pronosticado"))
```

El resultado muestra que el 88 % de los casos (22/25) ha sido correctamente clasificado. El mayor acierto se produce con el grupo 1 (92,3 %) y el menor en el grupo 3 (80 %).

Pero antes de comenzar a interpretar los resultados hemos de estar seguros de la significatividad de las funciones discriminantes puesto que, de no ser capaces de haber generado grupos estimados donde las variables explicativas tengan valores medios significativamente distintos, cualquier interpretación que hagamos de la influencia de cada una de esas variables sería absurda. El test de Wilks que realizamos mediante un MANOVA como en el tema 7 —dado que `lda{MASS}` no realiza esta prueba— confirma la significatividad de las

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 9.23.: Significatividad conjunta de las funciones discriminantes

```
          Df   Wilks approx F num Df den Df    Pr(>F)
categ      1 0.22082   13.408      5     19 1.113e-05 ***
Residuals 23
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cuadro 9.24.: Autovalores y proporción de la varianza explicada

```
Proportion of trace:
   LD1    LD2
0.9842 0.0158

$svd
[1] 6.2300478 0.7885377
```

diferencias entre estos vectores de medias (cuadro 9.23).

De la estimación del análisis discriminante obtenemos los autovalores y el porcentaje de la varianza que explica cada función discriminante (*proportion of trace*) (cuadro 9.24). Vemos que la primera función discriminante es responsable de la mayor parte de la separación entre los grupos (98,42 %) y la segunda tiene un papel secundario (1,58 %).

Nos resta por analizar la contribución relativa de cada variable a las funciones discriminantes. Para ello recordemos, del caso de dos grupos, que contamos con los coeficientes estandarizados de dichas funciones (dado que los que aparecen en el cuadro 9.21 eran los no estandarizados) y la matriz de estructura cuyos coeficientes interpretábamos de manera equivalente a las cargas de un análisis factorial y que corregían el problema que podíamos tener en una situación de elevada correlación con las cargas estandarizadas. Recordemos que `lda{MASS}` no ofrecía esta información pero que era bastante sencillo calcularla en `R` a partir de las salidas del análisis discriminante. Dado que en este caso la sintaxis es extensa y el lector siempre puede acudir a ella en los ficheros que acompañan a este manual, ofrecemos solo la misma para la primera variable de la primera de las funciones discriminantes en el caso de la matriz de estructura y para la primera función discriminante en el caso de los coeficientes estandarizados:

```
#=========================================
# Calculo de los coeficientes estandarizados
#=========================================
#grados libertad residuos = 23
```


Estadísticos e-Books & Papers

```
std.b.ingresosLD1=(sqrt(summary(fit.manova)
$SS$Residuals[1,1]/23))
*fit$scaling[1,1]
std.b.patrnetoLD1=(sqrt(summary(fit.manova)
$SS$Residuals[2,2]/23))
*fit$scaling[2,1]
std.b.provivLD1=(sqrt(summary(fit.manova)
$SS$Residuals[3,3]/23))
*fit$scaling[3,1]
std.b.casadoLD1=(sqrt(summary(fit.manova)
$SS$Residuals[4,4]/23))
*fit$scaling[4,1]
std.b.salfijoLD1=(sqrt(summary(fit.manova)
$SS$Residuals[5,5]/23))
*fit$scaling[5,1]


#==========================================
# Calculo de la matriz de estructura
#==========================================
SCPC.residual<-summary((fit.manova),test="Wilks")$SS$Residuals
SCPC.residual.varianzas<-SCPC.residual/23
SCPC.residual.correlaciones<-cov2cor(SCPC.residual.varianzas)

l.ingresosLD1=
SCPC.residual.correlaciones[1,1]*std.b.ingresosLD1+
SCPC.residual.correlaciones[1,2]*std.b.patrnetoLD1+
SCPC.residual.correlaciones[1,3]*std.b.provivLD1+
SCPC.residual.correlaciones[1,4]*std.b.casadoLD+
SCPC.residual.correlaciones[1,5]*std.b.salfijoLD1
```

El cuadro 9.25 nos ofrece los coeficientes estandarizados y el 9.26 la matriz de estructura. Si leemos el cuadro 9.25 podemos concluir que las variables que más influencia tienen en construir la primera función discriminante —que recordemos era la que tenía mayor capacidad de separación— son fundamentalmente el patrimonio neto y contar con un salario fijo.

Aunque el sentido de la influencia parece bastante evidente, conviene ser capaz de evaluar a qué grupos separa fundamentalmente esa función. Para ello vamos a mostrar tres gráficos complementarios. La figura 9.7 representa los individuos sobre los dos ejes discriminantes. Ya se intuye que la separación va a ser bastante nítida, aunque individuos como el 23 son proclives a estar mal clasificados. La figura 9.8 muestra como el grupo 3 (fallido) no se solapa con los demás a diferencia de un caso puntual que se debe corresponder con el 23. El solapamiento entre el 1 (cumplidor) y el 2 (moroso) también es menor. La figura 9.9, conocida como *mapa territorial*, que es el gráfico más informativo,

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 9.25.: Coeficientes estandarizados de las funciones discriminantes

	LD1	LD2
ingresos	-0.3223302	0.09941932
patrneto	-1.0777635	0.18537776
proviv	-0.2844800	0.44402065
casado	-0.4357721	0.05701841
salfijo	-1.1878251	-0.63292824

Cuadro 9.26.: Matriz de estructura

	LD1	LD2
ingresos	-0.3475734	0.31673748
patrneto	-0.3561321	0.70080901
proviv	-0.1832637	0.63862920
casado	-0.1223667	0.04069247
salfijo	-0.2990385	-0.88999276

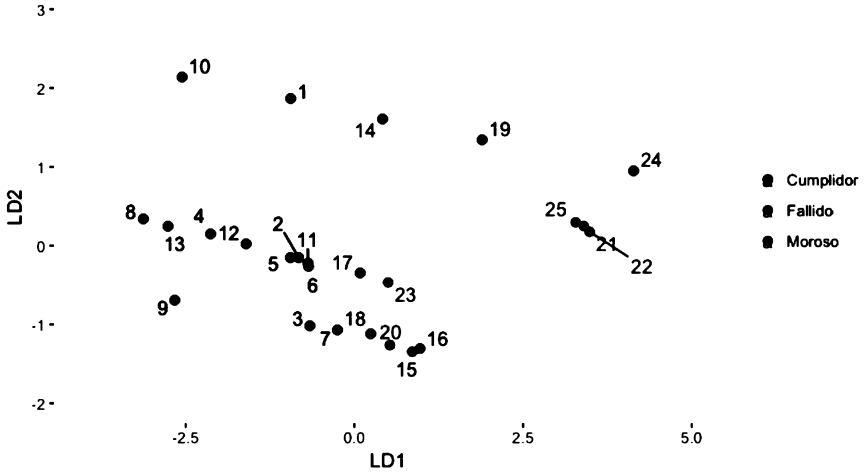
nos ofrece los centroides de los tres grupos para apreciar su separación (sería la traducción gráfica del MANOVA efectuado que confirmaba la diferencia de los vectores de medias) y nos muestra, marcados en gris claro, los tres casos mal clasificados. Los del 1 y el 2 están en la frontera pero el del grupo 3, que se corresponde con el caso 23, es una clasificación incorrecta muy clara que muestra un fallo en las funciones discriminantes debido, probablemente, al carácter anómalo de ese caso particular.

La figura 9.7 se ha obtenido añadiendo al fichero de datos las puntuaciones discriminantes obtenidas mediante la función `predict{MASS}` como se muestra a continuación:

```
grupo<-c(rep("1",13),rep("2",7),rep("3",5))
etiqueta<-c(rep("Cumplidor",13),
rep("Moroso",7),rep("Fallido",5))
individuo<-c(rep(1:25))

datos.grafico<-data.frame(datos,grupo,etiqueta,individuo,
predict(fit)$x[,1],predict(fit)$x[,2])

library(plyr)
datos.grafico<-rename(datos.grafico,
c("predict.fit..x...1."="LD1", "predict.fit..x...2."="LD2"))

ggplot(datos.grafico,
aes(x=LD1, y=LD2,colour=etiqueta,label=individuo)) +
  geom_point(size=2)+
```


Estadísticos e-Books & Papers

Figura 9.7.: Representación de las puntuaciones discriminantes de los individuos sobre las funciones discriminantes




```
scale_color_manual(values=c("black","red","blue"))+
geom_text_repel()+
expand_limits(x=c(-4,5), y=c(-2, 3))+
labs (x="LD1", y="LD2")+
guides(colour="legend",label=FALSE)+
theme(legend.title=element_blank())
```

Sin embargo, los otros dos gráficos son más sencillos de obtener. El primero es una opción del mismo paquete que realiza el análisis discriminante **plot{MASS}**, mientras que el segundo requiere de la instalación del paquete **{klaR}** pues es su función **partimat{klaR}**:

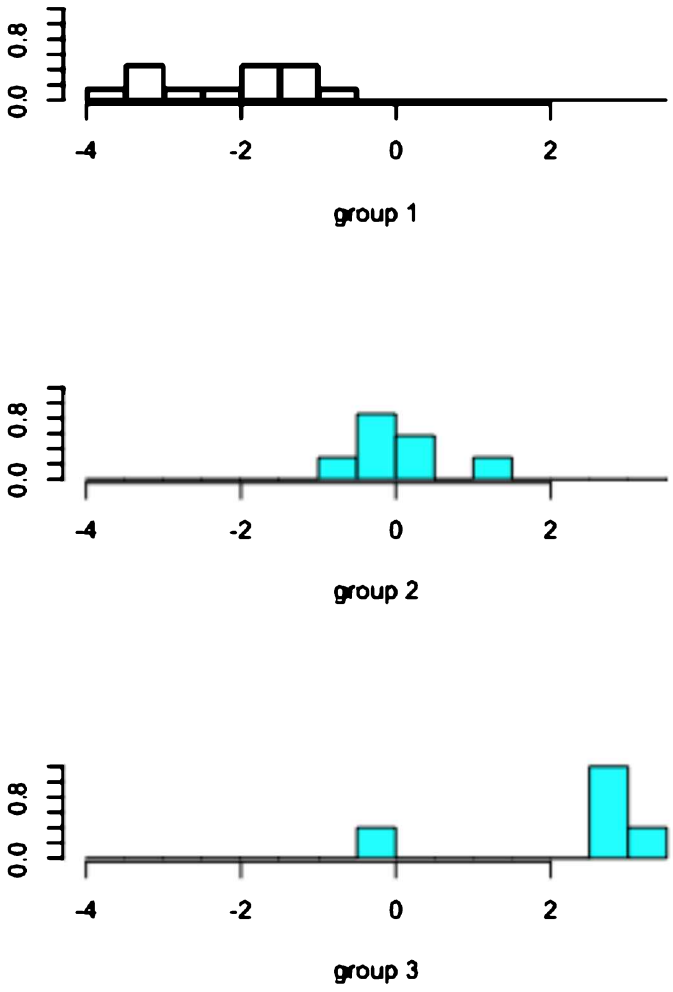
```
plot(fit,dimen=1)

colores<-c("gray","white","gray")
partimat(data=datos.grafico,grupo~LD2+LD1,prec=1000,
col.mean="darkred", image.colors =colores,
display.points=TRUE)
```

Solo resta, para cerrar el análisis, comprobar que se cumplen los supuestos en que se basa el análisis discriminante. Ya hemos señalado que con ejemplos de pocos casos necesarios para una ilustración didáctica de las herramientas, estos supuestos son poco plausibles, pero nada en la contrastación cambiaría para casos más complejos. No repetiremos la sintaxis porque es análoga al caso de dos grupos. El cuadro 9.27 muestra los resultados del test M de Box para

 © Ediciones Paraninfo

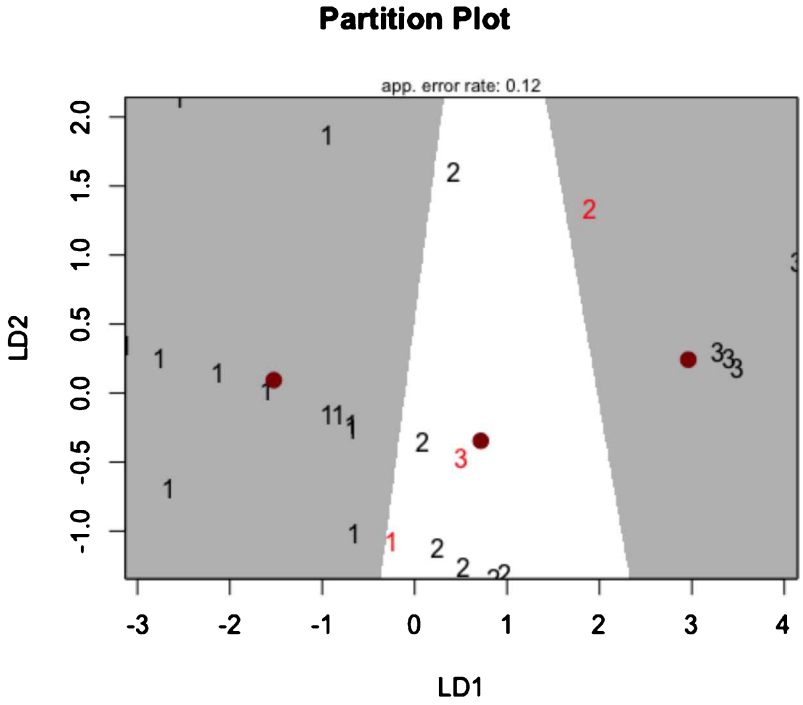
Estadísticos e-Books & Papers

Figura 9.8.: Proyecciones de los casos sobre los ejes discriminantes



Estadísticos e-Books & Papers

Figura 9.9.: Mapa territorial



© Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 9.27.: Comprobación de las hipótesis del modelo

```
Box's M-test for Homogeneity of Covariance Matrices

data:  datos[2:6]
Chi-Sq (approx.) = 105.04, df = 30, p-value = 2.925e-10

> mshapiro.test(grupo1)

      Shapiro-Wilk normality test

data:  Z
W = 0.78982, p-value = 0.005173

> mshapiro.test(grupo2)

      Shapiro-Wilk normality test

data:  Z
W = 0.45297, p-value = 4.136e-06
```

el análisis de la homocedasticidad y los test de normalidad multivariante de Shapiro. Ninguna de las hipótesis se sostiene. No se ofrece el test de normalidad para el grupo 3 debido a que no es posible invertir la matriz en el proceso de cálculo al haber filas linealmente dependientes, lo que no es extraño con pocos casos y con tres variables que son variables ficticias 1/0 donde es fácil provocar combinaciones lineales.

Aunque los gráficos son muy ilustrativos, si se desea tener información sobre las probabilidades a posteriori de pertenencia a los grupos, que es lo que el director de la oficina contaría para tomar las decisiones de conceder o no el préstamo, basta, como vimos en la sección anterior, aplicar la función `predict{MASS}`. Aplicada a los casos originales —se podría aplicar como vimos a los casos futuros sin clasificar— los resultados serían los que aparecen en el cuadro 9.28:

```
predict(fit)$posterior
```

Los resultados de la investigación satisfacen plenamente al director del Banco de Buñol, ya que se obtiene un porcentaje elevado de clientes clasificados correctamente (88 %). Además, el tipo de errores de clasificación que se cometen son los que tienen menor importancia para el banco. En efecto, al banco le preocupa sobre todo que un cliente moroso o fallido pueda ser considerado como cumplidor, ya que el coste de una clasificación errónea de este tipo es muy elevada para la entidad. En este sentido, como puede verse en el cuadro 9.22, no hay ningún cliente moroso o fallido que haya sido clasificado como cumplidor. Como resultado de la investigación, el Banco de Buñol dispone de un instrumento valioso que utilizará en el análisis de las solicitudes de nuevos préstamos.

Estadísticos e-Books & Papers

Cuadro 9.28.: Probabilidades a posteriori

```
$posterior
              1            2            3
1  9.369266e-01 0.0628047142 2.686379e-04
2  8.247632e-01 0.1749391924 2.975905e-04
3  6.888359e-01 0.3107019496 4.621508e-04
4  9.900390e-01 0.0099599026 1.082559e-06
5  8.607999e-01 0.1390201655 1.799119e-04
6  7.634597e-01 0.2360176568 5.226373e-04
7  4.628503e-01 0.5352085003 1.941186e-03
8  9.990130e-01 0.0009869389 1.278715e-08
9  9.956169e-01 0.0043830455 8.782775e-08
10 9.983815e-01 0.0016183088 2.187285e-07
11 7.689106e-01 0.2305737018 5.157021e-04
12 9.666315e-01 0.0333575174 1.098035e-05
13 9.976666e-01 0.0023333642 6.512773e-08
14 3.674986e-01 0.5864233196 4.607809e-02
15 5.839850e-02 0.9079896520 3.361185e-02
16 4.591089e-02 0.9093953041 4.469381e-02
17 3.577726e-01 0.6347980084 7.429346e-03
18 2.169922e-01 0.7747304567 8.277336e-03
19 7.173790e-03 0.3494213143 6.434049e-01
20 1.208961e-01 0.8631547911 1.594912e-02
21 1.482388e-05 0.0339172237 9.660680e-01
22 1.021743e-05 0.0293055571 9.706842e-01
23 1.701121e-01 0.8074534277 2.243442e-02
24 5.012192e-07 0.0044092836 9.955902e-01
25 2.499298e-05 0.0428203536 9.571547e-01
```


© Ediciones Paraninfo

Estadísticos e-Books & Papers

10. Regresión logística

10.1. Introducción

La regresión logística es, de manera simplificada, una regresión múltiple cuando la variable dependiente es no métrica, bien sea una variable no métrica con dos niveles (regresión logística binomial), bien tenga más de dos niveles (regresión logística multinomial). El problema que pretendemos resolver, por tanto, es el mismo que abordábamos con el análisis discriminante. La diferencia está en la forma de abordarlo con un planteamiento, la regresión, mucho más flexible y menos sensible a ciertas propiedades de los datos como la normalidad multivariante. Problemas típicos serían la explicación o predicción de la quiebra de empresas, o la decisión de un consumidor de recomprar o no un producto o servicio, por ejemplo.

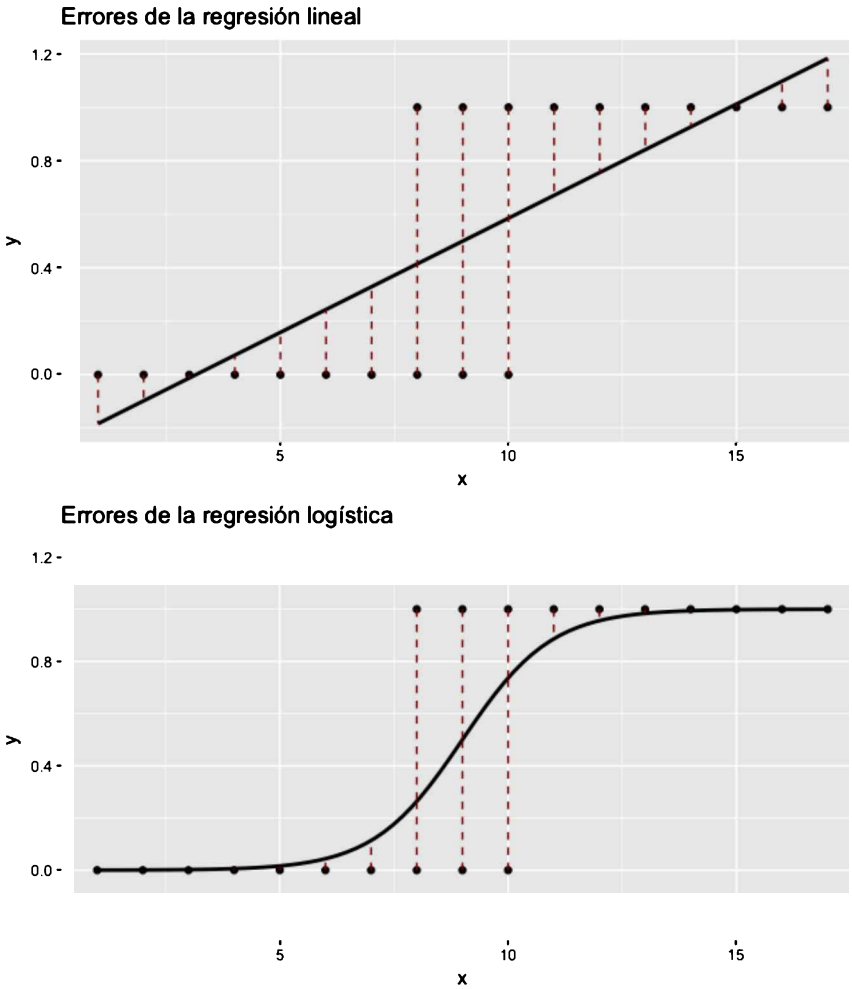
La otra pregunta que seguro que el lector se hará es qué razón hay para no abordar este problema mediante una regresión lineal múltiple como la que ya hemos presentado en este manual. La razón es que cuando la variable dependiente es no métrica, digamos dicotómica, es imposible que cumpla las condiciones que exigimos en una regresión múltiple: no puede seguir una distribución normal ni tener varianza constante. Si esto se ignora, los valores predichos —que deberían estar acotados entre 0 y 1— podrán ser negativos o superiores a 1 (Norusis, 2008). Veámoslo intuitivamente con un ejemplo sencillo donde tenemos una única variable explicativa X y una variable dependiente Y dicotómica. Como se puede comprobar en el panel (a) de la figura 10.1, si estimamos una regresión lineal y ajustamos una recta, vemos como la predicción de los dos primeros valores es negativa y la de los dos últimos superior a 1. Pero no solo eso, el error que cometemos en la estimación (los residuos serían la suma de líneas discontinuas desde el valor real al predicho sobre la recta) es muy superior en la estimación mediante esa recta de regresión que cuando ajustamos una función logística. Tampoco podría cumplirse la hipótesis de linealidad entre los valores observados (Berry, 1993), lo que es obvio solamente observando la figura 10.1. La solución ha de pasar por linealizar de alguna forma lo que es una relación no lineal. En eso se basa el modelo de regresión logística.

10.2. El modelo de regresión logística binomial

Como viene siendo habitual en este manual, comenzaremos describiendo el modelo, sus supuestos, los estadísticos para el contraste de hipótesis, los parámetros estimados para interpretar el modelo, indicadores de ajuste, etcétera,

Estadísticos e-Books & Papers

Figura 10.1.: Ajuste de una recta y una función logística a una variable dicotómica



 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 10.1.: Datos simulados para la ilustración 10.1

Y	X	Y	X
0	1	1	8
0	2	1	9
0	3	1	10
0	4	1	11
0	5	1	12
0	6	1	13
0	7	1	14
0	8	1	15
0	9	1	16
0	10	1	17

con una ilustración sencilla con pocos datos que nos permita obtener los resultados manualmente y generar la intuición de la herramienta y, posteriormente, lo aplicaremos a un caso más complejo.

Ilustración 10.1 Regresión logística binomial

Supongamos que queremos analizar la relación —bien para explicarla, bien para predecir— de una variable dependiente limitada dicotómica Y que toma los valores 0 y 1 en función de una variable métrica X. Los datos son los que aparecen en el cuadro 10.1 y la representación gráfica de los mismos es el panel (b) de la figura 10.1.

La relación entre la variable Y y X en un modelo de regresión lineal se plantearía del siguiente modo:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i \tag{10.1}$$

donde β_0 sería el punto en que la recta estimada intercepta al eje de ordenadas y β_1 la pendiente de esa recta. Si en lugar de una variable explicativa, como en nuestra ilustración, tuviéramos n de ellas, la expresión anterior sería:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni} + \varepsilon_i \tag{10.2}$$

A diferencia de la regresión lineal, la regresión logística no estima las expresiones (10.1) ni (10.2), es decir, no predice el valor de Y_i dados los valores de X_i, sino que directamente estima la probabilidad de que ocurra $Y_i(Y_i = 1)$ dados los valores de X_i. Por supuesto que incorpora las expresiones (10.1) y (10.2) para tener en cuenta la relación entre la variable dependiente y las independientes, pero las envuelve en la siguiente función para calcular la probabilidad de ocurrencia en lugar de predecir Y_i:

$$\Pr(Y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i})}} = \frac{1}{1 + e^{-Y}} \tag{10.3}$$

Estadísticos e-Books & Papers

y para el caso de n variables

$$\Pr(Y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni})}} = \frac{1}{1 + e^{-Y}} \tag{10.4}$$

donde $\Pr(Y)$ es la probabilidad de ocurrencia de Y, y e es la base del logaritmo natural. Es la forma de linealizar la relación que no era lineal e impedía la estimación de una regresión lineal como planteábamos en la introducción.

10.2.1. Estimación del modelo

En una regresión lineal, los coeficientes β del modelo se estiman —en el caso de mínimos cuadrados— minimizando como vimos en el capítulo 8 la suma de las distancias al cuadrado entre los valores observados y los que predeciría la función estimada. En el caso de la regresión logística, el modelo se estima mediante la minimización de la **función de máxima-verosimilitud,** que es un planteamiento análogo al evaluar cuánta información queda por explicar después de que el modelo se ha estimado:

$$LL = \sum_{i=1}^{N} [Y_i \ln(\Pr(Y_i)) + (1 - Y_i) \ln(1 - \Pr(Y_i))] \tag{10.5}$$

Veamos la lógica de la expresión (10.5). Imaginemos que, para un valor real de $Y = 0$, la función estimada ha pronosticado una probabilidad de ocurrencia cercana a 0 ($\Pr(Y) \simeq 0$) (lo que sería un acierto de la función), entonces (10.5) nos diría:

$$LL = \cancelto{0}{0 \times \ln(\simeq 0)} + (1 - 0) \times \cancelto{\simeq 0}{\ln(1 - (\simeq 0))} \simeq 0$$

puesto que la función logarítmica es 0 para $X = 1$. Dicho de otra manera, la función de máxima verosimilitud toma valores cercanos a 0 cuando la probabilidad predicha (cercana a 0 en caso de no ocurrencia, cercana a 1 en caso de ocurrencia) acierta en clasificar al caso como 0 o 1. Pero ¿qué pasa en caso de desacierto? Es decir, cuando, por ejemplo, se pronostica una probabilidad cercana a 1 ($\Pr(Y) \simeq 1$) para un caso cuyo valor real es $Y = 0$. Veamos qué ocurre con la expresión (10.5):

$$LL = \cancelto{0}{0 \times \ln(\simeq 1)} + (1 - 0) \times \cancelto{-\infty}{\ln(1 - (\simeq 1))} \simeq -\infty$$

puesto que $\ln(x) \to -\infty$ cuando $x \to 0$. Es decir, que en caso de desacierto la función de máxima verosimilitud crece haciéndose muy grande. La expresión (10.5) no es sino una suma de valores que miden aciertos (valores pequeños) y desaciertos (valores grandes). Mayor valor implicará menos capacidad de la estimación para replicar los valores reales. Dicho de otro modo, los parámetros β estimados serán aquellos que minimicen la función de máxima verosimilitud.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

10.2.2. Contraste de hipótesis para el modelo estimado

A. Contraste de significatividad global

Igual que en una regresión múltiple el primer paso es contrastar la hipótesis nula de que todos los coeficientes de regresión son nulos, dado que, de ser así, no tendría sentido continuar con la interpretación del modelo. El mismo paso es necesario dar en la estimación de un modelo de regresión logística. El planteamiento, sin embargo, es diferente al que realizábamos en aquel caso con el cálculo del estadístico F.

La estrategia es sencilla: calcular la máxima verosimilitud LL de un modelo en el que la función solo está formada por el intercepto β_0, es decir, un modelo en el que las variables explicativas no jugarían ningún papel o, dicho de otro modo, en el que todos los β serían nulos ($LL(0)$). A continuación estimamos la función de máxima verosimilitud para el modelo que estamos estimando ($LL(M)$). Si este segundo es significativamente más pequeño que el primero, podremos concluir que es más plausible (verosímil) por lo que alguna variable debe estar ejerciendo una influencia significativa en la predicción de la variable dependiente (su β sería distinto de cero).

El problema es que para ver si la diferencia es significativa es necesario que esta diferencia siga una distribución conocida y la función LL no la sigue. Sin embargo –2 veces esa función ($-2LL$) sí que se distribuye como una χ^2 con grados de libertad ($k_M - k_0$), siendo k_M el número de parámetros a estimar en el modelo, y k_0, el número de parámetros a estudiar en el modelo base (en el que solo se estima la constante):

$$\chi^2 = -2LL(M) - (-2LL(0)) = 2LL(0) - 2LL(M) \qquad gl = k_0 - k_M \quad (10.6)$$

A esta diferencia se la conoce como **razón de máxima verosimilitud**[1]. Es muy importante también, a efectos de interpretación de las salidas que nos ofrecerá R, saber que $-2LL$ suele etiquetarse bajo el término **deviance**.

Veamos este procedimiento para los datos de la ilustración. Utilizaremos para ello, de nuevo, la función `glm{stats}`, donde se señala que estamos estimando una regresión logística mediante el modificador `family=binomial`.

```
fit<-glm(y~x,family=binomial)
```

La función `glm{stats}` no nos proporciona el valor de la ji-cuadrado, pero su cálculo es inmediato:

$$\begin{gathered} -2LL(M) = 10{,}093 \quad gl = 18 \\ -2LL(0) = 27{,}726 \quad gl = 19 \\ \chi^2 = 27{,}726 - 10{,}093 = 17{,}6333 \quad gl = 1 \end{gathered}$$

[1] Si como hacen Field *et al.* (2012), el lector se pregunta por qué se denomina ratio o razón, solo tiene que recordar la propiedad de los logaritmos según la cual $\ln(a) - \ln(b) = \ln(a/b)$.

Estadísticos e-Books & Papers

Cuadro 10.2.: Resultados de -2LL

```
    Null deviance: 27.726  on 19  degrees of freedom
Residual deviance: 10.093  on 18  degrees of freedom
AIC: 14.093
```

Cuadro 10.3.: Resultados del contraste de significatividad global del modelo

```
> deviance.model
[1] 10.09252
> deviance.base
[1] 27.72589
> chi
[1] 17.63337
> chi.df
[1] 1
> sig.chi
[1] 2.678476e-05
```

Más operativo, sin embargo, es hacer el cálculo directamente en R. El cuadro 10.3 ofrece el resultado de la ji-cuadrado y su significatividad. Podemos comprobar como se rechaza la hipótesis nula de que las dos –2LL sean iguales (su diferencia sea nula) y, dado que la del modelo es más pequeña, este es significativamente mejor, por lo que algunos de los β han de ser significativamente distintos de cero. Solo nótese en la sintaxis que el número de grados de libertad del modelo están guardados en el objeto `fit` como `df.residual` ya que puede parecer un nombre poco intuitivo.

```
deviance.model<-fit$deviance
deviance.base<-fit$null.deviance
chi<-deviance.base-deviance.model
chi.df<-fit$df.null-fit$df.residual
sig.chi<-1-pchisq(chi,df=chi.df)
```

B. Contraste para los coeficientes individuales

Al igual, de nuevo, que en una regresión lineal, una vez descartada la hipótesis de que todos los coeficientes son nulos, necesitamos saber cuál es la contribución individual de los regresores a la explicación de la variable dependiente. En la regresión lineal construíamos un estadístico t como la ratio entre la estimación del parámetro no estandarizado y su error estándar, esto es[2], $t = \hat{B}/SE$, que

[2]Reconocemos cierta inconsistencia en la notación dado que hemos denominado β a los coeficientes sin distinguir si eran estandarizados o no y ahora estamos denotando como B a

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 10.4.: Contraste individual de los coeficientes de regresión

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -9.2448     4.9470  -1.869   0.0617 .
x             1.0272     0.5427   1.893   0.0584 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

sigue una distribución t de Student. El planteamiento en una regresión logística es análogo, solo que al estadístico se le denomina **test de Wald**.

$$W_j = \frac{\hat{B}_j}{SE_{B_j}} \tag{10.7}$$

Este test se distribuye como una normal y a partir de ella se establecen sus valores críticos. Hay que tener en cuenta algunas precauciones al analizar sus resultados. En primer lugar, cuando el coeficiente estimado B_j es grande, el error estándar tiene tendencia a crecer en exceso, lo que puede llevar a rechazar hipótesis nulas cuando no debería hacerse (error tipo II) al tener esa variable una contribución significativa real (Menard, 1995). Por eso la alternativa que se plantea —y que veremos posteriormente— consiste en una estimación del modelo con y sin la variable implicada cuando su coeficiente es muy grande, evaluando la significatividad mediante la razón de máxima verosimilitud, como hemos hecho con la significatividad global (Hauck y Donner, 1977).

El cuadro 10.4 nos ofrece la estimación del coeficiente no estandarizado B para nuestro modelo elemental y vemos que este sería significativo a un nivel de $p < 10\,\%$. ($B = 1{,}0272$, $W_x = 1{,}893$, $p < 0{,}1$).

El contraste de la razón de máxima verosimilitud no tiene sentido en este caso —o mejor dicho, ya lo habríamos aplicado—. Si quitamos el único regresor X, lo que nos queda es la constante, con lo que el contraste coincidiría con el ya efectuado para la significatividad global del modelo. En el caso 10.1 veremos un caso más completo y efectuaremos este contraste.

10.2.3. Interpretación de los coeficientes de regresión

En la regresión lineal los coeficientes no estandarizados B están afectados por las unidades de las variables a las que van asociados y por ello utilizamos los estandarizados β para ver la contribución relativa de cada variable independiente. El papel de los coeficientes estandarizados en la regresión logística la juegan los denominados ***odds ratio***. El concepto es muy común en el entorno anglosajón para tratar con las probabilidades, pero no tanto en nuestro entorno,

los no estandarizados. En general nos referiremos a β como el coeficiente de regresión general y matizaremos como B cuando sea relevante que sea el no estandarizado.

Estadísticos e-Books & Papers

por lo que requiere alguna elaboración, aunque el resultado final es fácilmente comprensible.

Se define el ***odd*** de un acontecimiento como la razón entre su probabilidad de ocurrencia y la de no ocurrencia:

$$odd = \frac{\Pr(Y=1)}{\Pr(Y=0)} \tag{10.8}$$

Como en nuestro caso la probabilidad de ocurrencia viene dada por (10.4), el odd puede escribirse como:

$$odd = \frac{\Pr(Y=1)}{\Pr(Y=0)} = \frac{\frac{1}{1+e^{-Y}}}{1-\frac{1}{1+e^{-Y}}} = \frac{\frac{1}{\cancel{1+e^{-Y}}}}{\frac{1+e^{-Y}-1}{\cancel{1+e^{-Y}}}} = \frac{1}{e^{-Y}} = e^{Y} \tag{10.9}$$

pero e^Y puede ponerse como:

$$e^{Y} = e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\cdots+\beta_n X_{ni}} = e^{\beta_0} e^{\beta_1 X_{1i}} \ldots e^{\beta_n X_{ni}} \tag{10.10}$$

Pues bien, al término e^{β_i} se le conoce como ***odd ratio*** y su interpretación es la siguiente: es el factor en que se incrementa el *odd* cuando la variable independiente i-ésima se incrementa en una unidad (y el resto permanecen constantes). Es fácil ver que siempre que el coeficiente no estandarizado B sea positivo, su *odd ratio* será superior a 1 (incrementa el *odd*) pero si B es negativo su *odd ratio* será inferior a 1 (disminuye el *odd* en la multiplicación). Mayores *odd ratio* pueden interpretarse como mayor influencia relativa de esa variable en la predicción de la ocurrencia del caso. Como vemos un uso equivalente a los coeficientes estandarizados de una regresión.

Es muy habitual leer en distintos trabajos interpretaciones incorrectas de los *odds ratio*. A veces se escribe que un *odd ratio* de, digamos, 1,2 supone que un incremento de una unidad en la variable a la que va asociado supone un incremento en un 20 % de la probabilidad de ocurrencia de la variable Y. No es así. Supone un incremento en un 20 % en el *odd*, es decir, en la probabilidad de ocurrencia frente a la de no ocurrencia. Traducir el *odd ratio* en incrementos de la probabilidad de ocurrencia es sencillo, pero la lectura no es así de directa. Calculemos primero los *odds ratio* para nuestra ilustración y veamos posteriormente esta traducción en términos de probabilidad.

La función `glm{stats}` no ofrece directamente los *odds ratio*. Es elemental el cálculo manual, por ejemplo para nuestra variable y teniendo en cuenta la información del cuadro 10.4, el *odd ratio* asociado a la variable X sería:

$$e^{B_x} = e^{1,0272} = 2,7932$$

Afortunadamente la función `toOR{LOGIT}` nos evita realizar manualmente este cálculo para modelos más complejos:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 10.5.: *Odd ratio* en el modelo estimado

```
                or  delta  zscore pvalue exp.loci. exp.upci.
(Intercept) 0.0001 0.0005 -1.8688 0.0617    0.0000    1.5701
x           2.7932 1.5158  1.8929 0.0584    0.9643    8.0913
```

```
toOR(fit)
```

Pues bien, volvamos sobre nuestra advertencia. Que el *odd ratio* de X sea 2,79 no quiere decir que cuando X pasa de 7 a 8 (por ejemplo) la probabilidad de ocurrencia se multiplique por 2,79, sino que el *odd* —la probabilidad de ocurrencia frente a la de no ocurrencia— lo hace. ¿En cuánto se incrementa entonces la probabilidad de ocurrencia? El cálculo es elemental:

Un sujeto donde X=7 tendrá una probabilidad de ocurrencia de:

$$\Pr(Y=1) = \frac{1}{1+e^{-Y}} = \frac{1}{1+e^{-(-9{,}2448+1{,}0272\times 7)}} = 0{,}1136$$

Si ese sujeto pasara a $X = 8$, entonces sus nuevos *odds* serían:

$$\text{Nuevos odds} = \text{Viejos odds} \times \text{odd ratio} \times \text{cambio en la variable} =$$

$$= 0{,}1136 \times 2{,}7932 \times 1 = 0{,}3173$$

Como:

$$Odd = \frac{\Pr(Y=1)}{1-\Pr(Y=1)} \rightarrow Odd\,[1-\Pr(Y=1)] = \Pr(Y=1)$$

$$Odd - Odd \times \Pr(Y=1) = \Pr(Y=1)$$

$$Odd = (1+Odd) \times \Pr(Y=1)$$

$$\Pr(Y=1) = \frac{Odd}{1+Odd}$$

El individuo pasará a tener una probabilidad de ocurrencia de:

$$\Pr(Y=1) = \frac{Odd}{1+Odd} = \frac{0{,}3173}{1+3173} = 0{,}2409$$

Es decir, al pasar de $X = 7$ a $X = 8$ el individuo ha visto incrementada su probabilidad de ocurrencia en un $(0{,}2409 - 0{,}1136)/0{,}1136 = 1,1206$, esto es, en un 112 % (no la ha visto multiplicada por 2,7932).

Estadísticos e-Books & Papers

10.2.4. Evaluando el ajuste del modelo

En la regresión múltiple, como vimos, el coeficiente de determinación R^2 es una medida muy intuitiva de lo bien que el modelo predice la variable dependiente, pues es la parte de la varianza total explicada por las variables independientes. Desgraciadamente no hay un equivalente tan intuitivo en la regresión logística. Veamos algunas propuestas realizadas para buscar un equivalente.

A. Pseudo R^2 de McFadden

McFadden (1979) propone calcular la R^2 del siguiente modo:

$$R^2_{MF} = \frac{-2LL(0) - (-2LL(M))}{-2LL(0)} \tag{10.11}$$

es decir, es la proporción de la reducción de la *deviance* que incorpora el modelo supone respecto al modelo ingenuo o base. Lógicamente puede variar entre 0 y 1. Si lo aplicamos a nuestro ejemplo:

$$R^2_{MF} = \frac{27{,}726 - 10{,}093}{27{,}726} = 0{,}6360$$

B. R^2 de Cox y Snell

La R^2 de Cox y Snell (1989) se define del siguiente modo:

$$R^2_{CS} = 1 - e^{\left[\frac{1}{N}(2LL(M) - 2LL(0)\right]} \tag{10.12}$$

donde N es el tamaño muestral. Una vez más el cálculo es inmediato:

$$R^2_{CS} = 1 - e^{\frac{1}{20}(10{,}093 - 27{,}726)} = 0{,}5859$$

Cabe apuntar que, aunque este estadístico se atribuye a estos autores, ya fue discutido con anterioridad por Cragg y Uhler (1970) y Maddala (1983). El problema que tiene esta medida de ajuste es que, por construcción, nunca puede alcanzar el 1. Por esta razón Nagelkerke (1991) propuso la modificación que se ofrece a continuación.

C. R^2 de Nagelkerke

Como se ha señalado Nagelkerke (1991) modifica la expresión de Cox y Snell (1989) para que pueda alcanzar el valor máximo de 1:

$$R^2_N = \frac{R^2_{CS}}{1 - e^{\left[\frac{2LL(0)}{N}\right]}} \tag{10.13}$$

donde R^2_{CS} es la R^2 de Cox y Snell (1989) y R^2_N es la R^2 corregida de Nagelkerke (1991). De manera inmediata:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

$$R_N^2 = \frac{0{,}5859}{1 - e^{\frac{-27{,}726}{20}}} = 0{,}7812$$

D. Matriz de confusión

En mi opinión una de las mejores herramientas para evaluar la precisión de las estimaciones es, simplemente, comparar los valores reales de la variable dependiente con los valores predichos. Si recordamos, la expresión (10.4) nos permite calcular la probabilidad de que cada caso pertenezca al grupo etiquetado como 1. Si adoptamos un criterio, por ejemplo, que cuando esta probabilidad es mayor a 0,5 se asigne al grupo 1 y, si es inferior a esta cifra, se asigne al 0 —es una de las muchas posibilidades, pero es la más intuitiva— podremos predecir el grupo al que, de acuerdo con las variables independientes, es más plausible que pertenezca. Dado que conocemos su pertenencia real, basta con generar una tabla cruzada de valores reales y predichos. Cuantos más elementos haya en la diagonal de esa tabla (predichos como 0 valores que son 0 y predichos como 1 valores que son 1), mejor será precisión del modelo estimado.

Veamos cómo realizar manualmente este cálculo a partir de los datos contenidos en el modelo estimado. El objeto `fit` guarda la probabilidad estimada por la expresión (10.4) como `fit$fitted.values`. En la sintaxis siguiente hemos guardado esa probabilidad en un objeto que llamamos `predict.fit` y lo hemos ido modificando para pasar de probabilidades a pertenencias predichas al grupo. Así `predict.fit[predict.fit>=.50]<-1` recodifica en sí mismo el objeto `predict.fit` y sustituye las probabilidades superiores a 0.50 por un 1 indicando la pertenencia al grupo y `predict.fit[predict.fit<.50]<-0` hace lo propio para las probabilidades inferiores y el grupo 0. Como el fichero datos tiene la variable y con el grupo de pertenencia original, basta con crear la tabla cruzada con `CrossTable{gmodels}` como hemos hecho en otros temas (cuadro 10.6).

```
#Matriz de confusi<U+00F3>n
predict.fit<-fit$fitted.values
predict.fit[predict.fit>=.50]<-1
predict.fit[predict.fit<.50]<-0
CrossTable(datos$y,predict.fit,prop.chisq=FALSE,prop.c=FALSE,
prop.r=FALSE)
```

Vemos que el porcentaje total de aciertos es de $(8 + 9)/20 = 85\,\%$. Sin embargo, aquí es necesario hacer una distinción que es muy importante para algunos estudios y es el hecho de que no todos los errores son iguales —o pueden no serlo—. En el cuadro 10.6 vemos que hemos cometido dos errores diferentes, hemos predicho que en dos casos ocurriría el suceso (1) cuando en la realidad no ocurrió (0) ($2/20 = 10\,\%$ de error) y en un caso hemos predicho que no ocurriría (0) cuando en realidad sí que lo hizo ($1/20 = 5\,\%$ de error). Normalmente se distinguen los siguientes conceptos en la precisión de un modelo:

Estadísticos e-Books & Papers

Cuadro 10.6.: Matriz de confusión

```
Total Observations in Table:  20


             | predict.fit
     datos$y |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
           0 |         8 |         2 |        10 |
             |     0.400 |     0.100 |           |
-------------|-----------|-----------|-----------|
           1 |         1 |         9 |        10 |
             |     0.050 |     0.450 |           |
-------------|-----------|-----------|-----------|
Column Total |         9 |        11 |        20 |
-------------|-----------|-----------|-----------|
```

1. **Sensibilidad**: % de positivos que son clasificados como positivos (9/10 = 90 %).
2. **Especificidad**: % de negativos que son clasificados como negativos (8/10 = 80 %).
3. **Falsos positivos**: % de negativos clasificados como positivos (2/10 = 5 %).
4. **Falsos negativos**: % de positivos clasificados como negativos (1/10 = 10 %).

Así, por ejemplo, un modelo que pretenda predecir, vista una analítica, si una persona tiene un tumor maligno, no puede valorar igual los falsos positivos —el paciente recibe un gran susto, pero está sano— que los falsos negativos —al paciente no se le da tratamiento pero sí que tiene ese tumor maligno—. El modelo puede calibrarse en función de estas consideraciones modificando el valor de la probabilidad de ocurrencia a partir del cual se predice la pertenencia al grupo 1.

Por razones didácticas hemos calculado de manera manual los distintos indicadores de ajuste, sin embargo, existen en `R` funciones preprogramadas para que no sea necesaria esta tarea. Así, por ejemplo, `PseudoR2 {BaylorEdPsych}` calcula los distintos R^2 que hemos ofrecido —y alguno más— mientras que `confusion_stat {LOGIT}` calcula la matriz de confusión detallando la precisión general, sensibilidad y especificidad. La sintaxis es muy sencilla y el lector puede comprobar en el cuadro 10.7 que los resultados coinciden con los calculados manualmente.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 10.7.: Indicadores de ajuste con funciones predefinidas

```
> PseudoR2(fit)
        McFadden     Adj.McFadden        Cox.Snell       Nagelkerke McKelvey.Zavoina
       0.6359892        0.4195850        0.5859085        0.7812113        0.8679852
          Effron            Count        Adj.Count              AIC    Corrected.AIC
       0.6492492        0.8500000        0.7000000       14.0925213       14.7984036
> confusion_stat(predict.fit, datos$y)
$matrix
     obs  0  1 Sum
pred
0         8  1   9
1         2  9  11
Sum      10 10  20

$statistics
   Accuracy Sensitivity Specificity
       0.85        0.80        0.90
```

Cuadro 10.8.: Descripción de los datos del caso 10.1

Variable	Descripción	Codificación
survived	¿Sobrevivió al naufragio?	1=Sí; 0=No
age	Edad del pasajero	Años
sex	Sexo del pasajero	1=Hombre; 0=Mujer
pclass	Clase del pasaje	1=Primera; 2=Segunda; 3=Tercera

```
PseudoR2(fit)
confusion_stat(predict.fit, datos$y)
```

Hemos presentado la regresión logística binomial basándonos en un caso muy sencillo que nos permitiera ilustrar con cálculos manuales los distintos elementos necesarios para interpretar el modelo pero, en contrapartida, determinadas cuestiones que exigen de un tamaño muestral más elevado y que tienen que ver, fundamentalmente, con aspectos como la capacidad discriminante del modelo (curva ROC, test de Homer-Lemeshow) no han podido ser abordados. Por esta razón, y otras como la incorporación de variables independientes categóricas, vamos a realizar un segundo caso que nos permita cerrar adecuadamente la presentación de la regresión logística binomial.

Caso 10.1. ¿Quién sobrevivió al Titanic?

El fichero que acompaña al manual contiene los datos históricos del pasaje del Titanic, indicando si sobrevivieron o no al naufragio. Las variables que vamos a emplear en este caso son las recogidas en el cuadro 10.8.

El objetivo es analizar qué variables tuvieron más peso en explicar la supervivencia. Como siempre conviene comenzar con un análisis descriptivo univariante, viendo la relación entre la variable dependiente (*survived*) y las independientes por separado. Cuando la independiente sea no métrica (clase del pasaje, sexo) utilizaremos tablas cruzadas y cuando sea métrica (edad) compa-

Estadísticos e-Books & Papers

Cuadro 10.9.: Relación univariante entre las variables independientes y la supervivencia de los pasajeros del Titanic

```
             | datos$survived
datos$pclass |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
           1 |       103 |       181 |       284 |
             |     0.363 |     0.637 |     0.272 |
-------------|-----------|-----------|-----------|
           2 |       146 |       115 |       261 |
             |     0.559 |     0.441 |     0.250 |
-------------|-----------|-----------|-----------|
           3 |       370 |       131 |       501 |
             |     0.739 |     0.261 |     0.479 |
-------------|-----------|-----------|-----------|
Column Total |       619 |       427 |      1046 |
-------------|-----------|-----------|-----------|

             | datos$survived
   datos$sex |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
           0 |        96 |       292 |       388 |
             |     0.247 |     0.753 |     0.371 |
-------------|-----------|-----------|-----------|
           1 |       523 |       135 |       658 |
             |     0.795 |     0.205 |     0.629 |
-------------|-----------|-----------|-----------|
Column Total |       619 |       427 |      1046 |
-------------|-----------|-----------|-----------|
        Welch Two Sample t-test

data:  datos$age by datos$survived
t = 1.7707, df = 868.26, p-value = 0.07696
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.176415  3.430696
sample estimates:
mean in group 0 mean in group 1
       30.54537        28.91823
```

raremos la media entre los grupos de la dependiente. El cuadro 10.9 muestra los resultados que parecen indicarnos una fuerte influencia de la clase del pasaje (sobrevivieron el 63,7 % de los de primera, pero solo el 26,1 % de los de tercera), del sexo (sobrevivieron el 75,3 % de las mujeres pero solo el 20,5 % de los hombres) y una influencia menor de la edad, puesto que aunque es algo más joven en promedio el grupo de supervivientes (28,9 años) esta diferencia no es significativa ($t = 1{,}77, p > 0{,}05$).

```
CrossTable(datos$pclass,datos$survived,chisq=TRUE,
prop.chisq = FALSE, prop.c = FALSE, prop.t=FALSE)
CrossTable(datos$sex,datos$survived,chisq=TRUE,
prop.chisq = FALSE, prop.c = FALSE, prop.t=FALSE)
t.test(datos$age~datos$survived)
```

Comenzamos estimando la regresión logística binomial. Hay que tener en cuenta una cuestión operativa. A todos los efectos la variable clase del pasaje (*pclass*) es una variable métrica si no hacemos nada cuando, realmente, es

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 10.10.: Significatividad global: ratio de máxima verosimilitud

```
    Null deviance: 1414.62  on 1045  degrees of freedom
Residual deviance:  982.45  on 1041  degrees of freedom
AIC: 992.45

Number of Fisher Scoring iterations: 4

> deviance.model<-fit$deviance
> deviance.base<-fit$null.deviance
> chi<-deviance.base-deviance.model
> chi.df<-fit$df.null-fit$df.residual
> sig.chi<-1-pchisq(chi,df=chi.df)
>
> deviance.model
[1] 982.4531
> deviance.base
[1] 1414.62
> chi
[1] 432.1673
> chi.df
[1] 4
> sig.chi
[1] 0
```

claramente una variable categórica ordinal. Para que el programa la trate como una variable categórica es necesario señalarle que lo es, lo que hemos hecho en la sintaxis generando una nueva variable que llamamos *pclass.f* definida como "factor" que es el término para que R sepa que es no métrica.

```
# Transformamos en categórica la clase del pasaje datos
$pclass.f <- factor(datos$pclass)

# Estimamos la regresión logística
fit<-glm(data=datos,survived~pclass.f+age+sex,family=binomial)
```

La primera cuestión que hay que analizar es la significatividad global del modelo mediante la ratio de máxima verosimilitud. El cuadro 10.10 nos permite comprobar que, efectivamente, la diferencia entre la *deviance* del modelo nulo y el modelo estimado es significativa y menor en el modelo ($\chi^2(4) = 432{,}16$; $p < 0{,}01$).

El paso siguiente es determinar qué variables individuales ejercen una influencia significativa (test de Wald), el sentido de esta (signo de los coeficientes no estandarizados) y la importancia relativa de cada una (*odds ratio*). El cuadro 10.11 ofrece los resultados. Recordemos que los *odds ratio* no los ofrece directamente la función `glm{stats}`, sino que los hemos obtenido mediante la función `toOR{LOGIT}`. En primer lugar vemos que todas las variables son significativas, el sexo ($B = -2{,}49, W = -15{,}04, p < 0{,}01$) y la edad tienen signo negativo ($B = -0{,}03, W = -5{,}43, p < 0{,}01$), lo que implica para la edad que ser más joven aumentaba la probabilidad de supervivencia pero hemos de tener precaución al interpretar el signo del sexo. Recordemos que la mujer estaba codificada como 0 y el hombre como 1, luego ese signo negativo implica que

Estadísticos e-Books & Papers

Cuadro 10.11.: Significatividad de las variables individuales y *odds ratio*.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.522074   0.326702  10.781  < 2e-16 ***
pclass.f2   -1.280570   0.225538  -5.678 1.36e-08 ***
pclass.f3   -2.289661   0.225802 -10.140  < 2e-16 ***
age         -0.034393   0.006331  -5.433 5.56e-08 ***
sex         -2.497845   0.166037 -15.044  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                 or   delta   zscore pvalue exp.loci. exp.upci.
(Intercept) 33.8546 11.0604  10.7807      0   17.8455   64.2254
pclass.f2    0.2779  0.0627  -5.6778      0    0.1786    0.4323
pclass.f3    0.1013  0.0229 -10.1401      0    0.0651    0.1577
age          0.9662  0.0061  -5.4325      0    0.9543    0.9783
sex          0.0823  0.0137 -15.0439      0    0.0594    0.1139
```

más valor de la variable (ser hombre) reduce la probabilidad de supervivencia, lo que ya intuíamos con los datos del cuadro 10.9.

Un esquema de razonamiento similar hay que aplicar con la variable clase del pasaje. Al ser una variable categórica, el programa la ha descompuesto en tres ficticias (una para cada clase del pasaje) y es necesario fijarse en cuál ha dejado fuera para interpretar los coeficientes en función de ella. Así vemos como se ha dejado fuera a la clase primera (en la salida están *pclass.2* y *pclass.3*). Por lo tanto, los signos negativos de esta dos variables nos indican que estar en 2.ª o en 3.ª disminuye la probabilidad de supervivencia respecto a los que viajan en primera. Tanto para *pclass.2* ($B = -1{,}28, W = -5{,}67, p < 0{,}01$) como para *pclass.3* ($B = -2{,}28, W = -10{,}14, p < 0{,}01$) estos coeficientes son significativos.

Al analizar la significatividad de los parámetros hablamos de las limitaciones del test de Wald cuando el tamaño de los coeficientes es grande por la inflación del error estándar que se produce. Planteábamos como alternativa la estimación del -2LL del modelo con esa variable y sin ella. Si lo aplicamos, por ejemplo, a la variable edad, que es la que está más en el límite de la significatividad, el modelo lo calcularíamos de esta forma (que es la misma función quitando la edad):

```
fit2<-glm(data=datos,survived~pclass.f+sex,family=binomial)
```

El resultado nos da un –2LL de 1013,8 basado en 1042 grados de libertad. Si nos fijamos en el –2LL del modelo con la variable edad (cuadro 10.10) es de 982,45 con 1041 grados de libertad, la diferencia es significativa $\chi^2(1) = 31,35; p < 0{,}01$, lo que implica que eliminar la variable empeora significativamente el ajuste del modelo y esta debe considerarse como significativa.

Pero los coeficientes estandarizados están afectados por las unidades y no nos permiten interpretar los efectos relativos de las variables. Para eso nos hemos de fijar en los *odds ratio*. Aunque obviamente sus estadísticos de significatividad coinciden con los de los no estandarizados —solo están ahí porque los

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

ha generado otra función de `R`— es la columna etiquetada como `or` en la que nos hemos de fijar. Al ser inferiores a cero —reducen la probabilidad de supervivencia— cuanto más pequeños sean, más importante es el efecto sobre esta probabilidad (más la reducen). Como intuíamos, el efecto de la edad es muy pequeño (su *odd ratio* es cercano a 1, que implicaría un efecto nulo). El sexo, sin embargo, es la variable que más influencia tiene, con el *odd ratio* más pequeño. Ir en tercera tiene un efecto reductor de la probabilidad de supervivencia prácticamente equivalente al del sexo, también muy importante. Vemos que ir en segunda, disminuyendo la probabilidad, no lo hace con tanta intensidad como ir en tercera.

En un caso con más de una variable explicativa es en el que tiene más sentido realizar el ejercicio de cálculo de las probabilidades de supervivencia y la relación entre *odds ratios* y probabilidad. Veámoslo con los personajes de la película Titanic. Jack Dawson (Leonardo DiCaprio) era un joven de 22 años que viajaba en tercera, mientras que Rose DeWitt Bukater (Kate Winslet) tenía 17 años y viajaba en primera. Del cuadro 10.11 tenemos que la función que calcula la probabilidad de sobrevivir vendrá dada por:

$$Pr(Y=1) = \frac{1}{1+e^{-(3{,}522-1{,}281\times pclass\cdot 2-2{,}289\times pclass\cdot 3-0{,}0343\times age-2{,}498\times sex)}}$$

Si aplicamos esta función a los dos personajes:

$$Pr(Y=1;Jack) = \frac{1}{1+e^{-(3{,}522-1{,}281\times 0-2{,}289\times 1-0{,}0343\times 22-2{,}498\times 1)}} = 0{,}1169$$

$$Pr(Y=1;Rose) = \frac{1}{1+e^{-(3{,}522-1{,}281\times 0-2{,}289\times 0-0{,}0343\times 17-2{,}498\times 0)}} = 0{,}9497$$

luego vemos que el desenlace encaja con los resultados, la probabilidad de supervivencia de Jack era muy reducida y la de Rose prácticamente era 1.

¿Qué hubiera pasado si Jack en lugar de viajar en tercera lo hubiera hecho en primera? Si nos fijamos en el *odd ratio* de la variable *pclass.3*, este toma el valor –2,2897. ¿Cómo podemos traducir este *odd ratio* en términos de cambio en la probabilidad de sobrevivir si viajara en primera? Como vimos anteriormente, los *odds* de Jack son:

$$odds = \frac{\Pr(Y=1)}{\Pr(Y=0)} = \frac{0{,}1169}{1-0{,}1169} = 0{,}1323$$

El cambio en los *odds* vendría dado por:

$$\text{Nuevos } odds = \text{Viejos } odds \times odd \text{ ratio} \times \text{cambio en la variable} =$$

Estadísticos e-Books & Papers

Cuadro 10.12.: Indicadores de ajuste y matriz de confusión

```
> PseudoR2(fit)
        McFadden     Adj.McFadden        Cox.Snell       Nagelkerke McKelvey.Zavoina
       0.3055006        0.2970177        0.3384448        0.4565043        0.4446390
          Effron            Count        Adj.Count              AIC    Corrected.AIC
       0.3793344        0.7848948        0.4730679      992.4531050      992.5107973
> confusion_stat(predict.fit, datos$survived)
$matrix
     obs    0    1  Sum
pred
0         520  126  646
1          99  301  400
Sum       619  427 1046

$statistics
   Accuracy Sensitivity Specificity
  0.7848948   0.8400646   0.7049180
```

$$= 0{,}1323 \times (2{,}2897) \times 2 = 0{,}6062$$

y, por lo tanto, su nueva probabilidad sería:

$$\Pr(Y = 1) = \frac{Odd}{1 + Odd} = \frac{0{,}6062}{1 + 0{,}6062} = 0,3774$$

es decir, se hubiera visto incrementada significativamente.

El siguiente paso es analizar el ajuste del modelo —o la capacidad explicativa del mismo quizás sea una expresión más adecuada— a partir de la matriz de confusión y los distintos R^2 que vimos con anterioridad. La sintaxis es idéntica y no la repetiremos, pero el cuadro 10.12 nos ofrece los resultados. Los R^2 se mueven alrededor de un 30 % y el porcentaje de predicciones correctas llega al 78,5 %. El modelo es más preciso en la predicción de los supervivientes (Sensibilidad, 84 %) que en los que no lo hicieron (Especificidad, 70,5 %).

Al tener ahora un caso con un tamaño muestral adecuado, tiene sentido que planteemos un análisis que no habíamos abordado con la ilustración, el de la capacidad discriminante del modelo.

10.2.5. Capacidad discriminante del modelo

La capacidad discriminante del modelo es, en muchas ocasiones, más importante que el ajuste que puedan proporcionarnos unos estadísticos R^2 sobre los que, además, no hay acuerdo en cuanto a su interpretación ni niveles de referencia. La capacidad discriminante del modelo es la capacidad del mismo para distinguir entre los dos grupos de casos (supervivencia y no supervivencia) en función de la probabilidad predicha. En cierta forma, la matriz de confusión es un buen indicador de la capacidad discriminante del modelo, sobre todo si se completa con los porcentajes de sensibilidad y especificidad.

Sin embargo existen indicadores adicionales que pueden utilizarse, fundamen-

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

talmente el **estadístico c asociado a las curvas ROC** (*Receiver Operating Characteristic*). Las curvas ROC comparan, para diferentes puntos de corte de la probabilidad —en el ejemplo hemos predicho que sobrevivirían aquellos casos en los que la probabilidad obtenida fuera >0,50 pero podríamos haber elegido otro punto de corte— cuál es la tasa de clasificaciones correctas *true positives* (TPR) y la de falsos positivos *false positives* (FPR), definidas ambas como:

$$TPR = \frac{\text{Positivos correctamente predichos}}{\text{Positivos reales totales}}$$

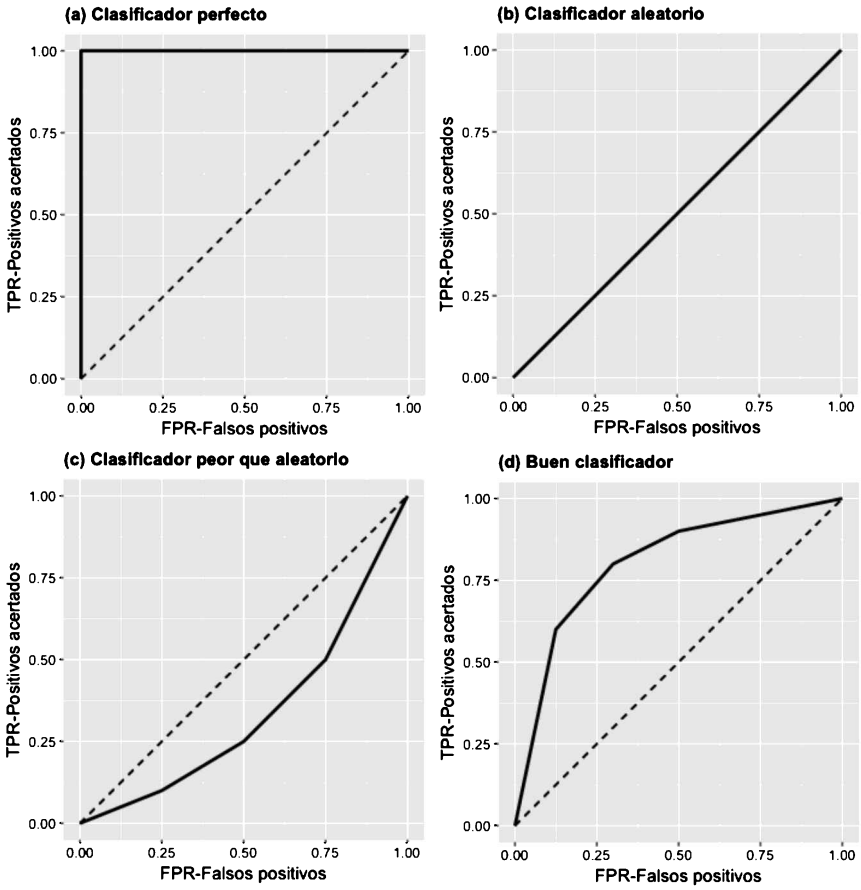
$$FPR = \frac{\text{Falsos positivos}}{\text{Negativos reales totales}}$$

La lógica se entiende mejor cuando se rastrea el origen de las ROC, que eran utilizadas en la Segunda Guerra Mundial para analizar la efectividad del radar, que había de ser calibrado de tal manera que se incrementara la tasa de acierto (TPR) sin incurrir en demasiados falsos positivos (FPR). Si ante cualquier señal en el radar (probabilidad 0) se activaba la alerta, la TPR sería 1, todos los positivos predichos serían positivos reales, pero también se incrementaría mucho la FPR, se detectarían muchos falsos positivos. Los cuatro gráficos de la figura 10.2 ilustran con casos extremos la interpretación de las ROC. El panel (a) sería un clasificador perfecto, la tasa de falsos positivos es 0 y la de clasificaciones correctas es 1. Luego cuanto mayor sea el área bajo la curva, mejor clasificador será el modelo. En el panel (b) tenemos una ilustración en la que el clasificador está haciendo predicciones aleatorias, como si se lanzara una moneda, ante un 50 % de clasificaciones correctas obtiene un 50 % de falsos positivos también. El panel (c) nos indica un clasificador cuyos resultados son peor que lanzar una moneda al aire, tiene siempre un porcentaje mayor de falsos positivos que de aciertos reales. El panel (d) sería un clasificador con un buen funcionamiento que se acerca bastante al panel (a) en cualquier caso para el porcentaje de clasificaciones correctas siempre es mejor que el de falsos positivos. Cuanto mayor sea el área bajo la curva, mejor clasificador será.

El estadístico c es el área bajo la curva ROC (Hanley y McNeil, 1982) y varía entre 0,5 y 1. Como hemos ilustrado anteriormente, un valor de 0,5 significa que el modelo no es mejor que lanzar una moneda al aire, mientras que el valor de 1 implica que el modelo siempre asigna mayores probabilidades de ocurrencia a los casos donde el suceso ocurre realmente que cuando no es así. En la mayoría de los programas a este valor c se le denomina AUC (*area under the curve*). En nuestro ejemplo lo vamos a obtener mediante la función `ROC{Epi}`, que es una función con una virtud: nos incorpora al gráfico el punto de corte óptimo, es decir, aquel que equilibra las FPR y TPR. Nótese en la figura 10.3 que este punto de corte sería 0,368 —cuando nosotros hemos utilizado el valor por defecto de 0,5— lo que daría una especificidad del 76,4 % (70,5 % en nuestro caso) mientras que la sensibilidad sería del 79,6 % (84 % en nuestro caso).

Estadísticos e-Books & Papers

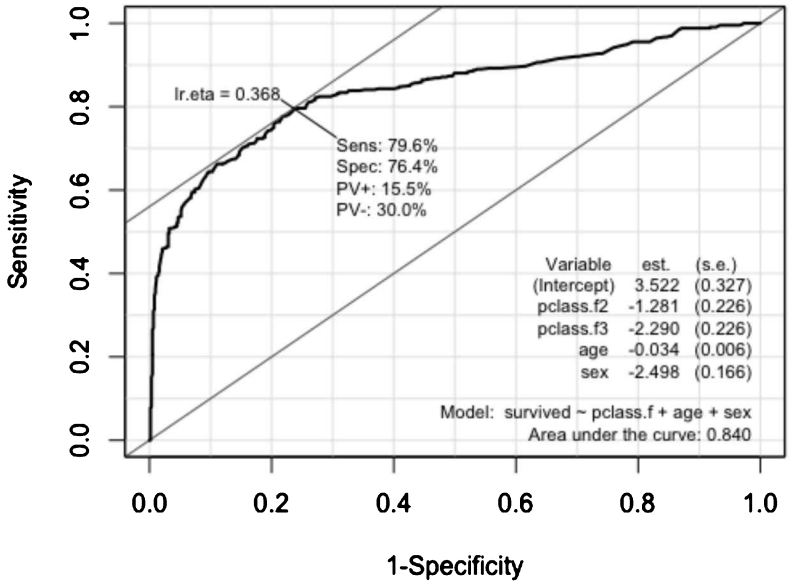
Figura 10.2.: Ejemplo de curvas ROC



© Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 10.3.: Curva ROC para el caso del Titanic




```
library(Epi) ROC(data=datos,form=survived~pclass.f+age+sex)
```

10.2.6. Calibración del modelo

El último aspecto que abordaremos con el caso 10.2 hace referencia a lo que Norusis (2008) denomina calibración del modelo y que no pudimos abordar en la ilustración porque, como veremos inmediatamente, el procedimiento diseñado por Hosmer y Lemeshow (1980) necesita de tamaños muestrales elevados.

La calibración del modelo evalúa cuán bien las probabilidades predichas y observadas coinciden a muy distintos niveles de esa probabilidad. El procedimiento propuesto por Hosmer y Lemeshow (1980) consiste en dividir los casos en, aproximadamente, 10 grupos iguales en función de la probabilidad estimada de ocurrencia del acontecimiento (supervivencia en nuestro caso) —deciles de riesgo— y se analiza cómo casan la ocurrencia y no ocurrencia del suceso mediante un test χ^2. Para facilitar la interpretación lo aplicamos al caso mediante la función `HL{vcdExtra}`.

Estadísticos e-Books & Papers

Cuadro 10.13.: Test de Hosmer-Lemeshow

```
$table
                 cut total obs         exp         chi
1  [0.0217,0.0885]   109  95 101.357004 -0.63143057
2    (0.0885,0.11]   102  83  91.647825 -0.90332872
3     (0.11,0.136]   109  96  95.524646  0.04863620
4    (0.136,0.215]    99  81  81.524923 -0.05813674
5    (0.215,0.303]   106  94  79.001875  1.68740086
6    (0.303,0.477]   103  61  62.333672 -0.16892254
7    (0.477,0.617]   110  59  48.896484  1.44488644
8    (0.617,0.762]    99  37  30.796918  1.11777373
9    (0.762,0.874]   104  10  19.315775 -2.11964391
10   (0.874,0.969]   105   3   8.600878 -1.90978607

glm(formula = survived ~ pclass.f + age + sex, family = binomial,
    data = datos)
 ChiSquare df      P_value
  40.07817  8 3.098073e-06
```

```
library(vcdExtra)
HL<-HLtest(fit,g=10)
HL
HL$table
```

El cuadro 10.13 nos ofrece, en primer lugar, las probabilidades a las que se han generado los 10 grupos aproximadamente iguales de casos (columna `cut`). En el primer grupo está el 10 % de la muestra, cuya probabilidad predicha es más pequeña y así sucesivamente. Supongamos que todas los casos del primer grupo tuvieran una probabilidad del 10 % —no tiene por qué ser así— entonces el porcentaje de casos en ese grupo que tuviera ocurrencia del caso (Y=1) debería ser del 10 %. Si fuera —digamos— del 90 % sería un indicador de mala calibración del modelo. El test de HL calcula los casos con Y=1 que esperaría en ese grupo como el promedio de las probabilidades estimadas y lo multiplica por el número de casos. A continuación hace lo mismo para los casos con Y=0 y aplica en esa decila el test χ^2 de Pearson:

$$\sum_{k=0}^{1}\sum_{l=1}^{g}\frac{(o_{kl}-e_{kl})^2}{e_{kl}} \tag{10.14}$$

donde o_{0l} es el número de Y=0 observados en el grupo l, o_{1l}, el de Y=1 observados en ese mismo grupo, y e_{0l} y e_{1l}, el número de casos esperados para Y=0 e Y=1. En su trabajo de 1980, los autores demuestran que su estadístico se distribuye aproximadamente como una χ^2 con $g-2$ grados de libertad, siendo g el número de grupos (10 en el ejemplo) en que se ha dividido la muestra. Si rechazamos la hipótesis nula de que los valores observados y esperados no difieren, estaríamos ante la evidencia de un modelo mal calibrado en la medida en que, estando en una decila en que esperamos un porcentaje determinado de ocurrencias, este es mucho mayor (o menor) de lo esperado.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

El cuadro 10.13 nos señala a cada uno de esos niveles de probabilidad cuántos casos hay (`total`), en cuántos no se ha producido el suceso realmente (`obs`) —aunque podría haber optado por mostrar Y=1— y en cuántos se ha pronosticado que el suceso no se daría (`exp`). El programa calcula un estadístico χ^2 para cada una de estas filas y el agregado del mismo se convierte en el estadístico χ^2 global para el test (grados de libertad son el número de grupos menos 2) como indicábamos en la expresión (10.14). En nuestro caso ($\chi^2(8) = 40{,}07817, p < 0{,}01$) la hipótesis nula puede rechazarse, lo que evidencia un ajuste deficiente. Sin embargo, este es un estadístico que depende del número de grupos que queramos crear, si escogemos pocos, será más difícil que detecte una mala calibración pero, si escogemos muchos, será difícil establecer si las diferencias entre frecuencias observadas y esperadas en cada fila son sistemáticas o debidas al azar.

10.3. Regresión logística multinomial

En el epígrafe anterior hemos visto modelos de elección discreta en los que existen dos alternativas en la variable dependiente. Sin embargo, en muchas ocasiones son más de dos las alternativas a las que se enfrenta el que tiene que tomar una decisión y el suceso que se quiere explicar o predecir. Por ejemplo, en unas elecciones legislativas, el elector puede optar entre varios partidos políticos o en la elección del medio de transporte para ir al trabajo, en una ciudad metropolitana, pueden considerarse, entre otras, el metro, el autobús o el automóvil propio.

Los modelos de elección discreta con más de dos alternativas se denominan **modelos multinomiales**. Según las alternativas estén o no ordenadas se denominan ordenados o no ordenados. En este epígrafe se analizan estos últimos, a los que denominaremos simplemente multinomiales.

Vamos a considerar que el número de alternativas son $J+1 \quad (0, 1, 2, \ldots, J)$, tomándose a la alternativa 0 como categoría de referencia. En los modelos binomiales solamente hay un vector de parámetros. Ahora, para construir el modelo logit multinomial se requieren J vectores de parámetros. (La regla que se aplica es análoga a la que se aplica con las variables ficticias explicativas: el número de vectores de parámetros es igual al número de categorías excluida la categoría de referencia).

Si definimos

$$Z_{ij} = \beta_{1j} + \beta_{2j} X_{2i} + \cdots \beta_{kj} X_{ki}$$

las probabilidades de cada alternativa en un modelo logit multinomial se ex-

Estadísticos e-Books & Papers

presan de la siguiente forma:

$$P_{ij} = \Pr(Y_i = j) = \frac{e^{-(\beta_{1j}+\beta_{2j}X_{2i}+\cdots\beta_{kj}X_{ki})}}{1+\sum_{g=1}^{J} e^{-(\beta_{1j}+\beta_{2j}X_{2i}+\cdots\beta_{kj}X_{ki})}}$$
$$P_{i0} = \Pr(Y_i = 0) = \frac{1}{1+\sum_{g=1}^{J} e^{-(\beta_{1j}+\beta_{2j}X_{2i}+\cdots\beta_{kj}X_{ki})}} \qquad j = 1, 2, \cdots, J \quad (10.15)$$

Cuando J es igual a 1, el modelo multinomial es igual al dicotómico. En el modelo anterior el logaritmo neperiano de los *odds ratio* entre la alternativa j y la alternativa de la categoría de referencia (0) viene dada por:

$$\ln\left[\frac{P_{ij}}{P_{i0}}\right] = \beta_{1j} + \beta_{2j}X_{2i} + \cdots \beta_{kj}X_{ki} \qquad (10.16)$$

También se puede calcular el logaritmo de los *odds ratio* entre cualquier par de alternativas. Así, por ejemplo, entre las alternativas j y g vendrá dado por:

$$\ln\left[\frac{P_{ij}}{P_{ig}}\right] = \beta_{1j} - \beta_{1g} + (\beta_{2j} - \beta_{2g})\, X_{2i} + \cdots + (\beta_{kj} - \beta_{kg})\, X_{ki} \qquad (10.17)$$

En los modelos multinomiales los parámetros deben interpretarse con mucho cuidado. En principio, se podría pensar que el signo del efecto marginal de una variable sobre la probabilidad de elegir una alternativa solo depende del correspondiente elemento del vector β_j. Sin embargo, derivando en (10.15), se puede establecer que el efecto marginal de la variable X_h es igual a:

$$\frac{\partial P_{ij}}{\partial X_h} = P_{ij}\left[\beta_{hj} - \overline{\beta}_h\right] \qquad (10.18)$$

donde $\overline{\beta}_h$ es la media de los parámetros β_{hj} para las J alternativas. Este efecto marginal depende de P_j que siempre es positiva y de la diferencia entre β_{hj} y $\overline{\beta}_h$. Por tanto, el signo dependerá del coeficiente y alternativa que se esté analizando.

Caso 10.2. Predicción de la clase social

Imaginemos que estamos interesados en saber en qué medida el desempeño de un individuo a lo largo de su paso por el sistema educativo es capaz de explicar su "éxito" final en la vida medido este como la clase social a la que pertenece tras unos años de acabar sus estudios. También nos interesa saber si el hecho de haber sido mejor estudiante en asignaturas relacionadas con las ciencias sociales o con las ciencias básicas y aplicadas influye en este resultado. La base

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 10.14.: Descripción datos caso 10.2

Variable	Descripción	Codificación
ses	Clase social	1=Baja; 2=Media; 3=Alta
female	Sexo	1=Mujer; 0=Hombre
science	Desempeño en Ciencias	Puntuación estandarizada
socst	Desempeño en Ciencias Sociales	Puntuación estandarizada

Cuadro 10.15.: Análisis bivariado de las variables dependiente e independientes

```
             | datos$ses
datos$female |       low |    middle |      high | Row Total |
-------------|-----------|-----------|-----------|-----------|
        male |        15 |        47 |        29 |        91 |
             |     0.165 |     0.516 |     0.319 |     0.455 |
-------------|-----------|-----------|-----------|-----------|
      female |        32 |        48 |        29 |       109 |
             |     0.294 |     0.440 |     0.266 |     0.545 |
-------------|-----------|-----------|-----------|-----------|
Column Total |        47 |        95 |        58 |       200 |
-------------|-----------|-----------|-----------|-----------|
```

```
ses  science    socst
1    low 47.70213 47.31915
2 middle 51.70526 52.03158
3   high 55.44828 57.13793
```

de datos contiene las variables que se recogen en el cuadro 10.14[3].

Como hemos recomendado para el caso anterior, debemos siempre realizar un análisis descriptivo bivariado que nos permita intuir la relación que va a existir entre la variable dependiente, la clase social, y las independientes. Como muestra el cuadro 10.15 parece que el porcentaje de hombres en las clases media y alta es mayor que el de mujeres y, también, que el desempeño en los estudios ha sido mayor cuanto mayor es la clase social.

```
CrossTable(datos$female,datos$ses,chisq=TRUE,
prop.chisq = FALSE,
prop.c = FALSE, prop.t=FALSE)
aggregate(cbind(science,socst)~ses, data=datos, mean, na.rm=TRUE)
```

Estamos en condiciones de estimar la regresión logística multinomial. La sintaxis es elemental. El único elemento que hemos de tener en cuenta es que

[3]Aunque está indicado en el fichero de sintaxis, a la base de datos se puede acceder en http://www.ats.ucla.edu/stat/data/hsb2.dta. Este caso es parte de la formación online de los Data & Statistical Services de la Universidad de Princeton, concretamente su autor es Oscar Torres-Reyna (otorres@princeton.edu).

Estadísticos e-Books & Papers

Cuadro 10.16.: Resultados de la estimación del modelo

```
multinom(formula = ses2 ~ science + socst + female, data = datos)

Coefficients:
       (Intercept)    science      socst femalefemale
middle   -1.912305 0.02356541 0.03892473   -0.8166207
high     -5.969577 0.04648473 0.08192982   -0.8494647

Std. Errors:
       (Intercept)    science      socst femalefemale
middle    1.127259 0.02097473 0.01951656    0.3909821
high      1.437546 0.02510004 0.02383389    0.4482127

Residual Deviance: 388.0697
AIC: 404.0697
```

hemos de decidir cuál ha de ser la categoría de referencia de la variable dependiente respecto a la que queremos interpretar los coeficientes de los otros dos niveles. Aunque en este caso la opción por defecto (el nivel codificado con el número más bajo) es el que queremos —la clase baja— indicamos en la sintaxis cómo hacerlo para el caso en que se deseara otro nivel de referencia. En esta base de datos las variables que son categóricas están incorporadas como texto, con lo cual la función que vamos a utilizar para la estimación `multinom{nnet}` sabe que son categóricas. De estar codificada como números habría que indicarle que son categóricas transformándolas en factores como hemos hecho en otras ocasiones mediante la opción `factor{base}`.

```
# Nivel de referencia dependiente: clase baja
datos$ses2 = relevel(datos$ses, ref = "low")
multi1 = multinom(ses2 ~ science + socst + female,data=datos)
summary(multi1)
```

El cuadro 10.16 nos ofrece los resultados de la estimación del modelo. Sin embargo enseguida observamos que nos faltan dos elementos importantes de información. Por un lado, no tenemos el indicador de ajuste global que era la significatividad del **ratio de verosimilitud**. Por otro lado, que, aunque están todos los elementos para calcular la **significatividad de cada parámetro** —estimación y error estándar—, esta no está calculada como tampoco están los *odds ratio* que en la regresión logística multinomial denominamos ***risk ratios***. Antes de comenzar la interpretación debemos tener toda esa información, por lo que, primero, vamos a proceder a obtenerla y luego la interpretaremos.

Para calcular el ratio de verosimilitud vamos a estimar el mismo modelo pero solo con la constante. Guardaremos las *deviance* (–2LL) y calcularemos la diferencia de ambas, que se distribuye como una χ^2 con tantos grados de libertad como la diferencia de grados de libertad entre ellas. Todo lo hemos empaquetado en la siguiente sintaxis y el cuadro 10.17 ofrece el resultado,

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 10.17.: Significatividad global del modelo

```
> print(cbind(chi2,df.chi2,Sig.chi2))
         chi2 df.chi2     Sig.chi2
[1,] 33.09537       6 1.005192e-05

> R2MF<-(multi0$deviance-multi1$deviance)/multi0$deviance
 [1] 0.07858052
```

Cuadro 10.18.: Coeficientes de regresión del modelo y su significatividad

```
> z <- summary(multi1)$coefficients/summary(multi1)$standard.errors
(Intercept)  science    socst femalefemale
middle   -1.696420 1.123514 1.994446    -2.088640
high     -4.152617 1.851978 3.437535    -1.895227

> p <- (1 - pnorm(abs(z), 0, 1)) * 2
(Intercept)    science        socst femalefemale
middle 8.980643e-02 0.26121914 0.0461033003   0.03674018
high   3.286941e-05 0.06402896 0.0005870359   0.05806237
```

donde vemos claramente que el modelo estimado mejora significativamente el modelo base, lo que permite concluir que no todos los coeficientes serán nulos. También podemos calcular fácilmente la R^2 de McFadden (1979) de acuerdo con la expresión (10.11) que presentamos con anterioridad.

```
# Estimamos el modelo solo con la constante
multi0 <- multinom(ses2 ~ 1,data=datos)
# Calculamos el estad<U+00ED>stico ji cuadrado como
# diferencia de sus -2LL (deviance)
chi2<-multi0$deviance-multi1$deviance
df.chi2<-multi1$edf-multi0$edf
Sig.chi2<-1-pchisq(chi2,df.chi2)
print(cbind(chi2,df.chi2,Sig.chi2))
# R2 de McFadden
R2MF<-(multi0$deviance-multi1$deviance)/multi0$deviance
```

Para calcular la significatividad de los parámetros basta dividir cada coeficiente por su error estándar y compararlo con el valor crítico de la normal. Una alternativa es el cálculo manual, cuyo resultado se ofrece en el cuadro 10.18.

```
z <- summary(multi1)$coefficients/summary(multi1)
$standard.errors z p <- (1 - pnorm(abs(z), 0, 1)) * 2
```

Pero como la disposición de la salida no es excesivamente legible disponemos, afortunadamente, de la función `stargazer{stargazer}`, que permite presen-

Estadísticos e-Books & Papers

Cuadro 10.19.: Coeficientes de regresión del modelo y su significatividad

```
=============================================
                      Dependent variable:
                  ---------------------------
                      middle         high
                       (1)            (2)
---------------------------------------------
science                0.024         0.046*
                      (0.021)        (0.025)

socst                 0.039**       0.082***
                      (0.020)        (0.024)

femalefemale          -0.817**       -0.849*
                      (0.391)        (0.448)

Constant              -1.912*       -5.970***
                      (1.127)        (1.438)

---------------------------------------------
Akaike Inf. Crit.     404.070        404.070
=============================================
Note:              *p<0.1; **p<0.05; ***p<0.01
```

tar como tablas prácticamente preparadas para la publicación a los coeficientes y su significatividad (cuadro 10.19)

```
stargazer(multi1, type="text")
```

Finalmente, al igual que ocurría con la regresión logística binomial, si queremos analizar las contribuciones relativas, es necesario recurrir al equivalente a los *odd ratios* que denominamos en la multinomial *risk ratios,* pero cuya definición es la misma, son simplemente e^B, siendo B los coeficientes no estandarizados que acabamos de obtener. Los calculamos de este modo y, de nuevo con la función **stargazer{stargazer}**, los hacemos legibles en el cuadro 10.20.

```
multi1.rrr = exp(coef(multi1))
stargazer(multi1, type="text", coef=list(multi1.rrr), p.auto=FALSE)
```

Con toda la información estamos en condiciones de interpretar los resultados. En primer lugar fijémonos en los coeficientes de regresión del cuadro 10.19. No olvidemos que el grupo de referencia es la clase baja. El coeficiente de regresión de las notas en ciencias sociales es significativo y positivo para las clases medias y altas en relación con la clase baja. Este signo de los coeficientes nos señala que un incremento en el rendimiento en esta área de conocimiento incrementa la probabilidad de estar en la clase media y en la clase alta respecto a estar en la baja. Fijémonos, sin embargo, que el signo de la variable sexo es negativo

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 10.20.: *Risk ratios* del modelo estimado

	Dependent variable:	
	middle (1)	high (2)
science	1.024	1.048*
	(0.021)	(0.025)
socst	1.040**	1.085***
	(0.020)	(0.024)
femalefemale	0.442**	0.428*
	(0.391)	(0.448)
Constant	0.148*	0.003***
	(1.127)	(1.438)
Akaike Inf. Crit.	404.070	404.070
Note:	*p<0.1; **p<0.05; ***p<0.01	

para las dos clases sociales y que estaba codificado como 1= mujer. Nos está señalando que, permaneciendo todo lo demás constante, ser mujer disminuye la probabilidad de estar en esas clases media y alta.

Para analizar el impacto relativo de las variables nos hemos de fijar en los *risk ratio.* Recordemos que aquellas variables con un coeficiente positivo tendrán *risk ratio* superiores a la unidad e inferiores a 1 en caso contrario. Tener esto en cuenta es importante para, de la lectura del cuadro 10.20, ver que el mayor impacto sobre la movilidad social —restricción a la misma en este caso— la tiene la variable sexo, puesto que ser mujer, hace 0,82 veces más probable estar en la clase media respecto a estar en la baja, pero, claro, multiplicar por 0,82 es una disminución de la probabilidad de ocurrencia.

Estadísticos e-Books & Papers

Estadísticos e-Books & Papers

11. Análisis de componentes principales

11.1. Introducción

En muchas ocasiones el investigador se enfrenta a situaciones en las que, para analizar un fenómeno, dispone de información de muchas variables que están correlacionadas entre sí en mayor o menor grado. Estas correlaciones son como un velo que impiden evaluar adecuadamente el papel que juega cada variable en el fenómeno estudiado. El análisis de componentes principales (PCA) permite pasar a un nuevo conjunto de variables —las componentes principales— que gozan de la ventaja de estar incorrelacionadas entre sí y que, además, pueden ordenarse de acuerdo con la información que llevan incorporada. Si estuviéramos realizando una regresión, el tener muchas variables independientes correlacionadas entre sí puede generarnos un problema de multicolinealidad que oscurezca la interpretación de los resultados. Trabajar con componentes principales eliminaría ese problema sin pérdida de información (asumiendo que utilizásemos todas las componentes principales).

Otras veces el objetivo no es tanto evitar la correlación, sino reducir datos, es decir, trabajar con una cantidad menor de variables que representen el mismo problema minimizando la pérdida de información que conlleva el no utilizar todas las variables originales. Si en lugar de trabajar con todas las variables, trabajamos con un número menor de componentes principales, garantizamos que la pérdida de información es la menor posible facilitando la interpretación, eso sí, siempre que seamos capaces de dotar de contenido a los componentes principales en función de las variables que agrupe.

Varias preguntas irán surgiendo a lo largo del tema: ¿cuánta información estamos dispuestos a perder para ganar facilidad de interpretación? ¿cuántos componentes principales deberían retenerse? ¿cómo se interpretan? A todas ellas daremos respuesta.

Una cuestión adicional es la fuerte confusión que suele existir entre el análisis de componentes principales y el análisis factorial exploratorio. Esta confusión se incrementa porque, como señala Sharma (1996), muchos programas estadísticos como SPSS ofrecen el análisis de componentes principales como una opción dentro del procedimiento de análisis factorial exploratorio. En el tema 12, cuando profundicemos en el análisis factorial, dedicaremos parte del tiempo a señalar las diferencias, pero para ello es importante tener una clara intuición de qué es el análisis de componentes principales y cómo funciona.

Estadísticos e-Books & Papers

11.2. La geometría del análisis de componentes principales

Esta sección está totalmente basada en la ilustración que realiza Sharma (1996) del PCA y es deudora de la brillantez y sencillez con que este autor es capaz de sintetizar un proceso geométricamente complejo. La ilustración se basa en un caso con dos variables pero permitirá en la sección siguiente la generalización a cualquier número de ellas. El cuadro 11.1 ofrece los valores originales de estas dos variables para 12 casos, los valores de estas mismas variables en desviaciones respecto a la media, sus varianzas, así como la matriz de la suma de cuadrados y productos cruzados **SCPC**, la matriz de varianzas y covarianzas entre las variables (**S**) y la matriz de correlaciones (**R**). Para análisis posteriores veremos que es importante ver que (a) la varianza total de las dos variables es $44,182 = 23,091 + 21,091$ y que ambas variables están correlacionadas $\rho = 0,746$. A la variable x_1 le corresponde el 52,26 % de la varianza total y a la x_2 el 47,74 %.

Vamos a proyectar las dos variables que están representadas en un espacio bidimensional sobre una recta que denominaremos X_1^* que formará, como muestra la figura 11.1, un ángulo θ con el eje original X_1. Por trigonometría básica, las proyecciones (coordenadas) de cualquier punto sobre ese nuevo eje X_1^* —representadas con un punto sobre la línea— vendrán dadas por la expresión:

$$x_1^* = \cos\theta \times x_1 + \sin\theta \times x_2 \tag{11.1}$$

donde x_1^* es la coordenada de una observación en el nuevo eje, mientras que x_1 y x_2 son las coordenadas de esa observación sobre los ejes originales. Para el ejemplo del gráfico se ha escogido un ángulo de θ =10°, por lo que la ecuación sería:

$$x_1^* = 0,985x_1 + 0,174x_2 \tag{11.2}$$

con lo que la coordenada en el nuevo eje del punto 12 ($x_1 = -8; x_2 = -3$) sería –8,399, que es el valor representado en la figura 11.1. El cuadro 11.2 ofrece las proyecciones de todas las observaciones sobre ese eje y la varianza de las mismas, que puede comprobarse que es un 68 % de la varianza total inicial ($28,659/44,182$), pero lo importante es que la varianza de las proyecciones sobre este eje es mayor que la varianza que recogía cada una de las variables originales por separado ($28,659 > 23,091; 28,659 > 21,091$).

Si en lugar de un ángulo de θ =10° utilizáramos otro ángulo, las proyecciones serían diferentes y también lo sería la varianza explicada por ese eje del total de la varianza. Obtener un componente principal consiste en ir variando el ángulo θ hasta que la nueva variable x_1^* explique el máximo valor posible de la varianza total de las dos variables originales (44,182). La combinación lineal para ese ángulo que maximiza la varianza explicada (minimiza la pérdida de información aunque hay una variable menos que interpretar) es la componente principal.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 11.1.: Datos originales y en desviaciones respecto a la media

Caso	x_1		x_2	
	Original	Desviación respecto media	Original	Desviación respecto media
1	16	8	8	5
2	12	4	10	7
3	13	5	6	3
4	11	3	2	-1
5	10	2	8	5
6	9	1	-1	-4
7	8	0	4	1
8	7	-1	6	3
9	5	-3	-3	-6
10	3	-5	-1	-4
11	2	-6	-3	-6
12	0	-8	0	-3
Media	8	0	3	0
Varianza	23,091	23,091	21,091	21,091

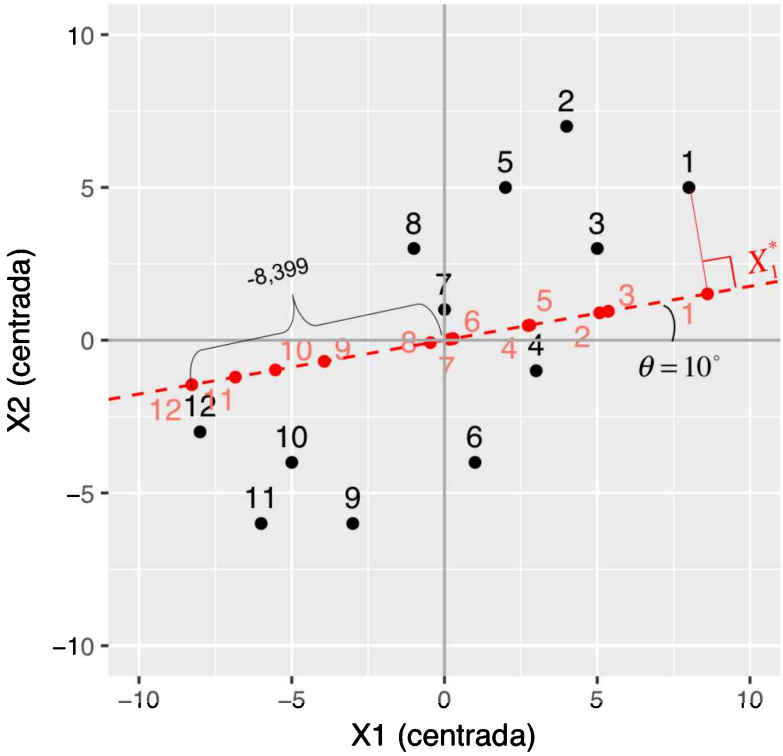
$$\mathbf{SSCP} = \begin{bmatrix} 254 & 181 \\ 181 & 232 \end{bmatrix} \quad \mathbf{S} = \begin{bmatrix} 23,091 & 16,455 \\ 16,455 & 21,091 \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} 1,000 & 0,746 \\ 0,746 & 1,000 \end{bmatrix}$$

Fuente: Sharma (1996, p. 59)

Estadísticos e-Books & Papers

Figura 11.1.: Datos originales centrados y su proyección sobre X_1^*



Fuente: Sharma (1996, p. 60).

© Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 11.2.: Datos originales centrados y nueva variable x_1^* para una rotación de θ =10°

Caso	Datos centrados		x_1^*
	x_1	x_2	
1	8	5	8,747
2	4	7	5,155
3	5	3	5,445
4	3	-1	2,781
5	2	5	2,838
6	1	-4	0,290
7	0	1	0,173
8	-1	3	-0,464
9	-3	-6	-3,996
10	-5	-4	-5,619
11	-6	-6	-6,951
12	-8	-3	-8,399
Media	0	0	0
Varianza	23,091	21,091	28,659

Fuente: Sharma (1996, p. 61).

La figura 11.2, basada en el cuadro 11.3, muestra el porcentaje de varianza explicada para distintos ángulos θ. El máximo se alcanza para θ =43,261°, por lo que el componente principal es:

$$x_1^* = \cos 43,261 \times x_1 + \sin 43,261 \times x_2 = 0,729x_1 + 0,685x_2 \tag{11.3}$$

Se puede comprobar en el cuadro 11.3 como el nuevo eje x_1^* no recoge toda la varianza, solo el 87,31 % de la misma, por lo tanto será posible identificar un segundo eje —al que denominaremos x_2^*— que recoja la varianza no contemplada por el primero y al que definiremos de manera ortogonal al anterior. Por lo tanto, si el ángulo entre x_1 y x_1^* era θ, también será θ el ángulo entre x_2 y x_2^*. La ecuación para el segundo eje será, por tanto:

$$x_2^* = -\sin\theta \times x_1 + \cos\theta \times x_2 \tag{11.4}$$

y como θ =43,261°, entonces la expresión anterior se convierte en:

$$x_2^* = -0{,}685x_1 + 0{,}728x_2 \tag{11.5}$$

Si calculamos ahora la proyección de las variables originales sobre el segundo eje con la expresión (11.5), el resultado que se muestra en el cuadro 11.4 y se ilustra en la figura 11.3 permite extraer conclusiones importantes:

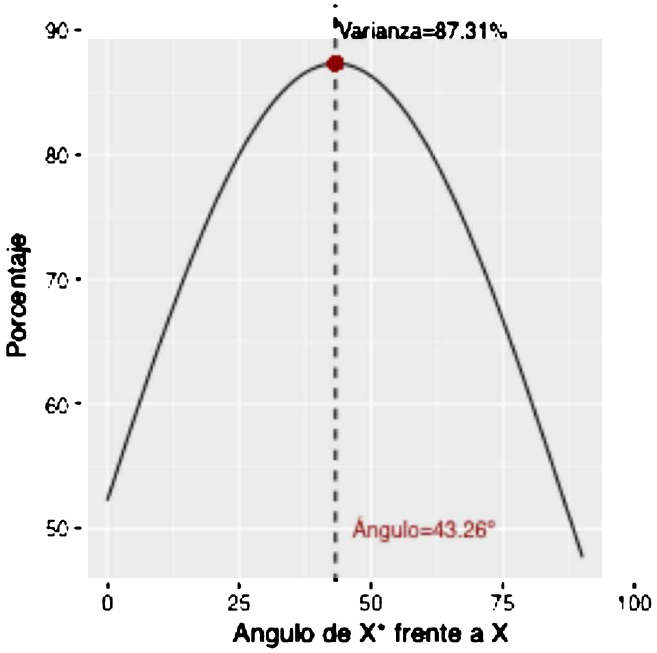
Estadísticos e-Books & Papers

Cuadro 11.3.: Varianzas explicadas por la nueva variable x_1^* para distintos ángulos de rotación θ

θ	Varianza total	Varianza de x_1^*	%
0	44,182	23,091	52,26
10	44,182	28,659	64,87
20	44,182	33,434	75,68
30	44,182	36,841	83,39
40	44,182	38,469	87,07
43,261	44,182	38,576	87,31
50	44,182	38,122	86,28
60	44,182	35,841	81,12
70	44,182	31,902	72,20
80	44,182	26,779	60,60
90	44,182	21,091	47,77

Fuente: Sharma (1996, p. 61).

Figura 11.2.: Datos originales centrados y su proyección sobre X_1^*



Fuente: Sharma (1996, p. 60)

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 11.4.: Datos originales centrados y nueva variable x_1^* para una rotación de θ =10°

Caso	Datos centrados		x_1^*	x_2^*
	x_1	x_2		
1	8	5	9,253	-1,841
2	4	7	7,710	2,356
3	5	3	5,697	-1,242
4	3	-1	1,499	-2,784
5	2	5	4,883	2,271
6	1	-4	-2,013	-3,598
7	0	1	0,685	0,728
8	-1	3	1,328	2,870
9	-3	-6	-6,297	-2,313
10	-5	-4	-6,382	0,514
11	-6	-6	-8,481	-0,257
12	-8	-3	-7,882	3,298
Media	0	0	0	0
Varianza	23,091	21,091	38,576	5,606

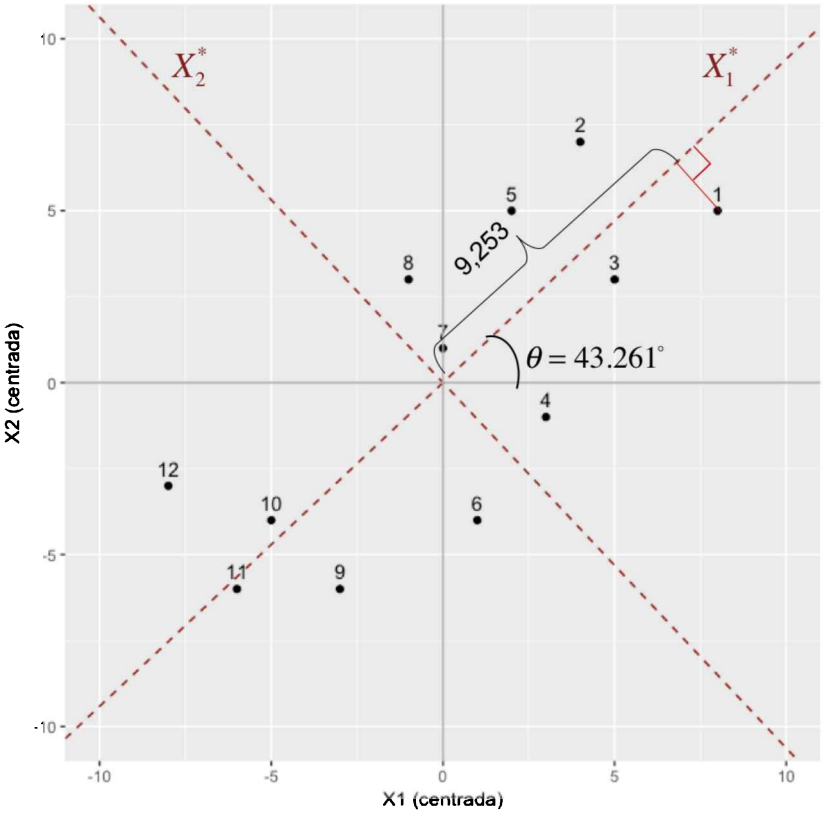
$$\mathbf{SSCP} = \begin{bmatrix} 424{,}334 & 0{,}000 \\ 0{,}000 & 61{,}666 \end{bmatrix} \quad \mathbf{S} = \begin{bmatrix} 38{,}576 & 0,000 \\ 0,000 & 5,606 \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} 1,000 & 0,000 \\ 0,000 & 1,000 \end{bmatrix}$$

Fuente: Sharma (1996, p. 62).

Estadísticos e-Books & Papers

Figura 11.3.: Datos originales y nuevos ejes (componentes principales)



Fuente: Sharma (1996, p. 63)

© Ediciones Paraninfo

Estadísticos e-Books & Papers

1. La orientación de los datos no cambia, solo que puede representarse tanto sobre los ejes originales como sobre los nuevos.

2. A los nuevos ejes se les denomina *componentes principales* y a las proyecciones de las variables sobre ellos, *puntuaciones sobre las componentes principales.*

3. Las nuevas variables son una combinación lineal de las originales y también están centradas en su media (esta es cero).

4. La suma de cuadrados totales (diagonal de la SSCP), $486 = 424,334 + 61,666$ es la misma en las variables originales y en las transformadas.

5. Las varianzas de x_1^* y x_2^* son 38,756 y 5,606, y la suma de ambas, 44,182. Esta suma es la misma que la de las variables originales, no cambia, lo que es lógico porque la orientación de los puntos en el espacio no ha cambiado.

6. Los porcentajes de la varianza recogidos por de x_1^* y x_2^* son, respectivamente, el $87,31\,\% = (38,576/44,182)$ y el $12,69\,\% = (5,606/44,182)$ y este es el principal cambio. La varianza recogida por la primera variable es *más grande que la varianza recogida por cualquiera de las variables originales.*

7. La correlación entre las nuevas variables es cero, x_1^* y x_2^* son ortogonales, están incorrelacionadas.

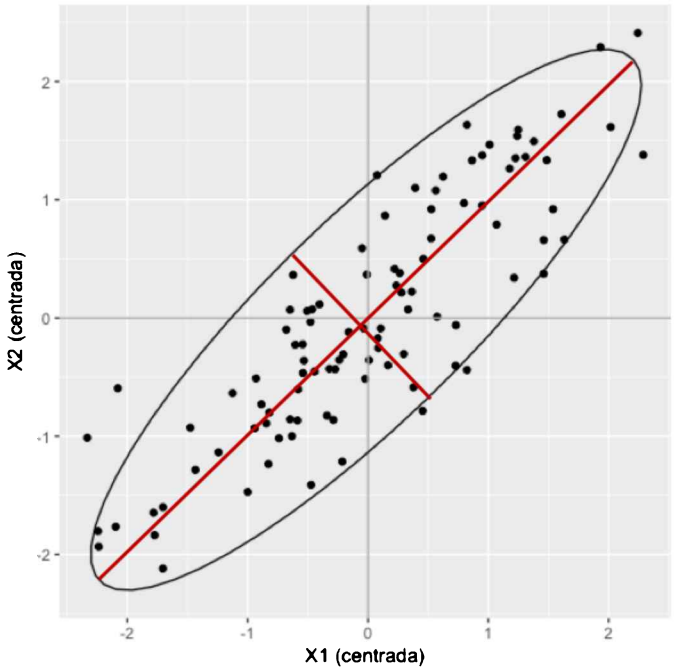
Esta ilustración geométrica del PCA es fácilmente generalizable a más de dos variables. Si tenemos p variables, estas podrán representarse en un espacio p-dimensional de tal forma que el primer componente principal recogerá el máximo de la varianza total, mientras que el segundo recogerá el máximo de la varianza que no haya recogido el primero y así sucesivamente. Todos los componentes estarán incorrelacionados. El número máximo de componentes principales coincidirá con el número original de variables p.

11.3. Componentes principales de dos variables

Analizada la interpretación geométrica del análisis de componentes principales, debemos proceder a su formalización matemática para trasladar la intuición a los procedimientos que luego vamos a ver reflejados en las salidas de los programas que usemos en la estimación. Lo desarrollaremos para el caso de dos variables y luego lo generalizaremos a p de ellas. Es verdad que el caso de dos variables es poco útil dado que una técnica de reducción de datos tiene poco que aportar cuando las variables iniciales son pocas, pero nos permitirá consolidar, con un desarrollo fácil de seguir y de realizar manualmente, la intuición de la sección anterior.

Estadísticos e-Books & Papers

Figura 11.4.: Interpretación gráfica de autovectores y autovalores



Comenzaremos por intentar trasladar la intuición geométrica de un concepto que es clave en el desarrollo del PCA y que no es sencillo de comprender, el de los **autovectores y autovalores**. Creemos que quien mejor ha trasladado esa intuición general ha sido Field (2005). Este autor plantea esta situación. Imaginemos, como es nuestro caso, que tenemos dos variables distribuidas normalmente y que están correlacionadas entre sí (si no lo estuvieran, los componentes principales serían los ejes originales y el PCA no tendría sentido). Al estar correlacionadas, su gráfico de dispersión formará una elipse como la de la figura 11.4. Podemos, por tanto, trazar dos líneas que midan la longitud y la altura de esa elipse. Estas líneas son los *autovectores* de la matriz de correlaciones original. Nótese que estas dos líneas son perpendiculares, como lo son siempre los autovectores. Si añadiéramos una tercera variable, la elipse se convertiría en algo parecido a un balón de rugby y tendríamos tres líneas perpendiculares, tres autovectores, y así sucesivamente.

Cada autovector tiene asociado un *autovalor*. El autovalor nos da una medida de la longitud del autovector y observando ese valor podemos tener una idea clara de lo homogénea o heterogéneamente que los datos están distribuidos. Veámoslo. Llamamos *condición* a la ratio entre el autovalor más grande y el más pequeño. Veamos dos casos extremos. Si no hubiera relación entre dos variables (panel a de la figura 11.5), el gráfico de dispersión mostrará —más o

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

menos— un círculo, y como la altura y la longitud —los autovectores— son los mismos, los mismos serían los autovalores que miden su longitud y esa relación daría 1. En el caso de correlación perfecta, el gráfico de dispersión forma una línea recta, la altura de la elipse será muy pequeña (0 en el caso extremo) y la división entre los autovalores tenderá hacia infinito.

Pues bien, en la ilustración que estamos siguiendo, como se comprueba en el cuadro 11.1, partíamos de la siguiente matriz de correlaciones que muestra que sí que existe un nivel de correlación razonable entre las variables que justifique la realización de un PCA:

$$\mathbf{R} = \begin{bmatrix} 1,000 & 0,746 \\ 0,746 & 1,000 \end{bmatrix}$$

La aplicación del procedimiento de componentes principales requiere calcular los autovalores y los autovectores de la matriz de covarianzas. En nuestra ilustración y, como también puede comprobarse en el cuadro 11.1, la matriz de varianzas y covarianzas es:

$$\mathbf{S} = \begin{bmatrix} 23,091 & 16,455 \\ 16,455 & 21,091 \end{bmatrix}$$

Calculando los autovalores y autovectores y estimando el PCA mediante la función de `R, sdv{stats}`, obtenemos los resultados del cuadro 11.5:

```
# Datos originales
x1<-c(16,12,13,11,10,9,8,7,5,3,2,0)
x2<-c(8,10,6,2,8,-1,4,6,-3,-1,-3,0)
# Datos centrados
x1_m<-x1-mean(x1) x2_m<-x2-mean(x2)
# Matriz datos originales
X<-matrix(c(x1,x2),ncol=2,nrow=12)
# Matriz datos centrados
Xm<-matrix(c(x1_m,x2_m),ncol=2,nrow=12)
# SSCP, S y R (datos centrados)
SSCP<-t(Xm) %*% (Xm)
S<-SSCP/(12-1)
R<-cov2cor(S)

svd(X)
```

La aplicación del procedimiento de componentes principales requiere calcular las raíces características y los vectores característicos de la matriz de covarianzas. Para la matriz $\mathbf{R}$ de nuestro ejemplo las raíces características que se obtienen son:

$$\lambda_1 = 38{,}576$$

Estadísticos e-Books & Papers

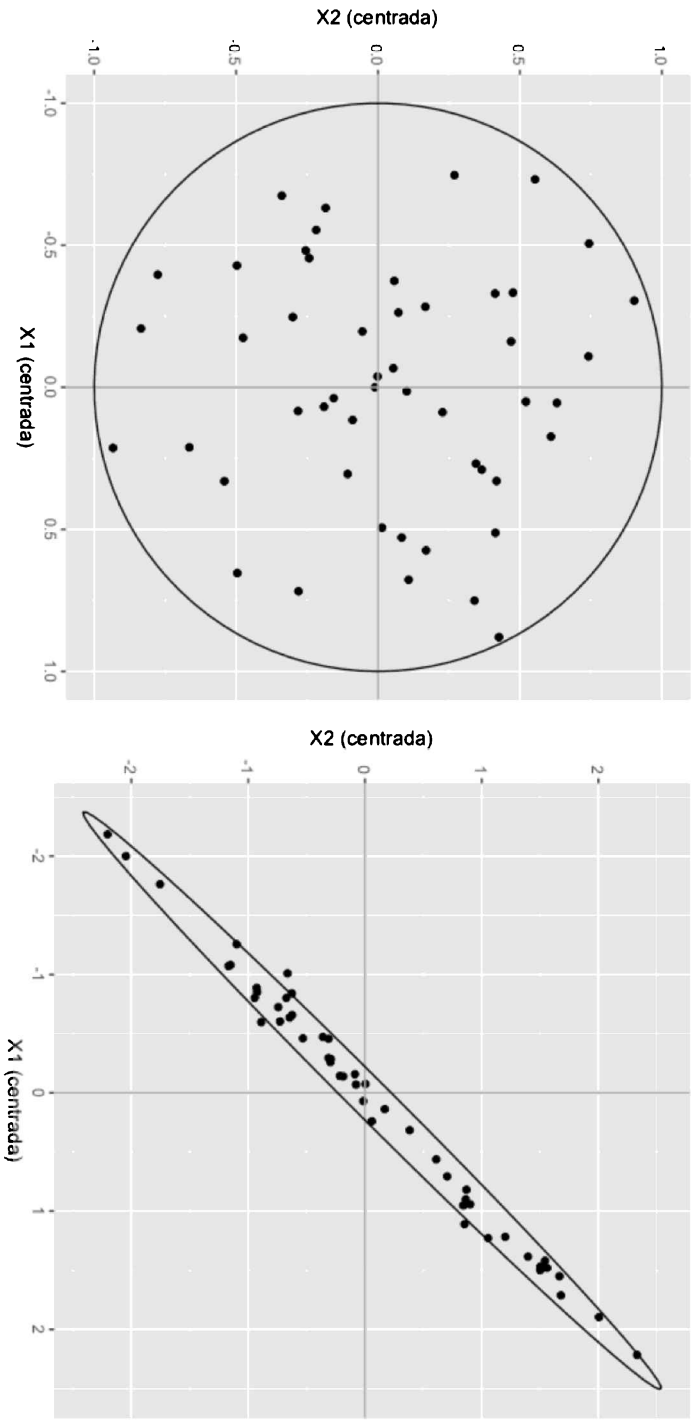


Figura 11.5.: Casos extremos de relación entre variables

© Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 11.5.: Autovalores y autovectores

```
> svd(S)
$d
[1] 38.575813  5.606005

$u
           [,1]       [,2]
[1,] -0.7282381 -0.6853242
[2,] -0.6853242  0.7282381

$v
           [,1]       [,2]
[1,] -0.7282381 -0.6853242
[2,] -0.6853242  0.7282381
```

$$\lambda_2 = 5{,}606$$

Vemos que, como hemos señalado, la importancia relativa del primer componente principal es mucho mayor $(38,576/(38,576+5,606) = 87,3\,\%)$ que la del segundo. Por lo tanto, si decidiéramos realizar el análisis con una sola variable en lugar de con dos (reducción de datos) perderíamos apenas un 13 % de la información total.

Si observamos los datos del cuadro 11.4, vemos como el autovalor coincide con la varianza de cada componente principal a la que va asociado, confirmando esa interpretación geométrica de que, cuanto mayor es su valor, más correlacionadas están las proyecciones en torno a su eje:

$$\mathbf{S} = \begin{bmatrix} 38{,}576 & 0,000 \\ 0,000 & 5,606 \end{bmatrix}$$

Cada autovalor tiene asociado un autovector, que, en el caso de nuestro ejemplo y como muestra el cuadro 11.5 son:

$$\mathbf{u_1} = \begin{bmatrix} u_{11} \\ u_{12} \end{bmatrix} = \begin{bmatrix} -0{,}728 \\ -0{,}685 \end{bmatrix}$$

$$\mathbf{u_2} = \begin{bmatrix} u_{21} \\ u_{22} \end{bmatrix} = \begin{bmatrix} -0{,}685 \\ 0{,}728 \end{bmatrix}$$

¿A qué se corresponden esos autovectores? Precisamente marcan el giro que se ha producido en los ejes, recordemos que x_2^* por ejemplo lo obtuvimos:

$$x_2^* = -\sin\theta \times x_1 + \cos\theta \times x_2$$

Estadísticos e-Books & Papers

$$x_2^* = -0{,}685x_1 + 0{,}728x_2$$

$$\sin\theta = 0,685 \rightarrow \arcsin(0,685) = 43,261$$

es decir, los coeficientes de los vectores $\mathbf{u_1}$ y $\mathbf{u_2}$ son los coeficientes que hay que aplicar a las variables originales para obtener las componentes principales. Así, genéricamente las componentes principales se expresan de la siguiente forma:

$$x_1^* = u_{11}X_1 + u_{12}X_2$$

$$x_2^* = u_{21}X_1 + u_{22}X_2$$

y han de cumplir las siguientes condiciones arbitrarias, pero necesarias para que los ejes sean ortogonales (11.7) y fijar las escalas de medida (11.6):

$$u_{11}^2 + u_{12}^2 = 1 \;\; ; \;\; u_{21}^2 + u_{22}^2 = 1 \tag{11.6}$$

$$u_{11}u_{21} + u_{21}u_{22} = 0 \tag{11.7}$$

Por lo tanto, con todo lo expuesto, para nuestro ejemplo, los **componentes principales** serían:

$$x_1^* = -0{,}728x_1 - 0{,}685x_2$$

$$x_2^* = -0{,}685x_1 + 0{,}728x_2$$

que se puede comprobar que cumplen que:

$$-0{,}728^2 - 0{,}685^2 = 0,530 + 0{,}470 = 1$$

$$-0{,}728 \times (-0{,}685) - 0{,}685 \times 0{,}728 = 0$$

Las proyecciones de los casos sobre esos ejes son lo que denominamos **puntuaciones de los componentes principales**. Es simplemente la aplicación de los componentes principales a las puntuaciones para cada caso de las variables originales centradas; por ejemplo, para el caso 1, en el que ($x_1 = 8$ y $x_2 = 5$), sus puntuaciones factoriales serían:

$$x_1^* = -0{,}728 \times 8 - 0,685 \times 5 = -9,2525$$

$$x_2^* = -0{,}685 \times 8 + 0,728 \times 5 = 1,8414$$

Si ejecutamos mediante `prcomp{stats}` el análisis de componentes principales, las puntuaciones factoriales para cada caso son las que se muestran en el

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 11.6.: Puntuaciones en las componentes principales

```
$x
              PC1        PC2
 [1,] -9.2525259  1.8414027
 [2,] -7.7102217 -2.3563703
 [3,] -5.6971632  1.2419065
 [4,] -1.4993902  2.7842106
 [5,] -4.8830971 -2.2705423
 [6,]  2.0130586  3.5982767
 [7,] -0.6853242 -0.7282381
 [8,] -1.3277344 -2.8700386
 [9,]  6.2966594  2.3134563
[10,]  6.3824874 -0.5136683
[11,]  8.4813738  0.2574838
[12,]  7.8818776 -3.2978790
```

cuadro 11.6:

```
prcomp(X)
```

En ocasiones las puntuaciones en los componentes principales se ofrecen estandarizadas para tener media 0 y desviación típica de 1, lo que se consigue simplemente dividiendo la puntuación factorial en el componente principal que acabamos de obtener por sus respectivas desviaciones típicas.

En el análisis de componentes principales es importante conocer la correlación de cada variable con las componentes para tener una medida de lo importante que es cada variable en la interpretación del componente principal. A este concepto se le denomina **carga factorial.** Su obtención es muy sencilla. Así, la carga factorial l_{hj} entre la componente h-ésima y la variable j-ésima viene dada por:

$$l_{hj} = \frac{u_{hj}}{\hat{s}_j}\sqrt{\lambda_h} \tag{11.8}$$

donde toda la notación es conocida salvo $\hat{s}_j$, que es la desviación típica de la variable j-ésima y que en nuestro ejemplo las tenemos en el cuadro 11.4. Así las cargas factoriales de x_1 y x_2 sobre la primera componentes serían:

$$l_{11} = \frac{-0,728}{\sqrt{23,091}}\sqrt{38,756} = -0,94$$

$$l_{12} = \frac{-0,685}{\sqrt{21,091}}\sqrt{38,756} = -0,93$$

Mientras que sobre el segundo componente principal serían:

Estadísticos e-Books & Papers

Cuadro 11.7.: Cargas factoriales. Matriz factorial.

```
$var$cor
        Dim.1      Dim.2
V1 0.9412618 -0.3376776
V2 0.9268425  0.3754503
```

$$l_{21} = \frac{-0,685}{\sqrt{23,091}}\sqrt{5,606} = -0,34$$

$$l_{22} = \frac{0,728}{\sqrt{21,091}}\sqrt{5,606} = -0,38$$

La función `prcomp{stats}` no ofrece estas correlaciones o cargas factoriales, pero, por ejemplo, la función `PCA{FactoMineR}` sí que lo hace (cuadro 11.7). En el caso de un ejemplo tan elemental la aportación de esta información es mínima y la contribución de las dos variables es muy similar a la formación del componente principal 1 y 2, pero cuando tomemos ejemplos más realistas, con más variables, esta información será central para la interpretación de los componentes. A esta matriz se la denomina **matriz de componentes** o **matriz factorial**, dependiendo del programa.

```
library(FactoMineR)
PCA(X = X, scale.unit = FALSE, ncp = 2, graph = TRUE)
```

11.4. Componentes principales para el caso general

En este epígrafe se va examinar con detalle cómo se obtiene, o se extrae, la primera componente principal. Posteriormente, se indicarán las reglas que se aplican para extraer el resto de las componentes. En cualquier caso, conviene señalar que la obtención de componentes principales es un caso típico de cálculo de autovalores y autovectores de una matriz simétrica. En los últimos apartados se examinarán algunas propiedades de las componentes principales. Concretamente, se obtendrán las varianzas de las componentes, las correlaciones entre las componentes y las variables originales, y las puntuaciones de las componentes. Este apartado puede omitirse si se ha entendido e intuido el apartado anterior en cuanto que es una generalización matemática, pero no va a aportar elementos nuevos que considerar más allá de los expuestos.

Antes de realizar la exposición analítica de los puntos que se acaban de enumerar, vamos a hacer un resumen de los **resultados más importantes que se van a obtener** y que son una extensión de los ya obtenidos en la sección anterior.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

1. Las componentes principales son combinaciones lineales de las variables originales.

2. Los coeficientes de las combinaciones lineales son los elementos de los autovectores asociados a la matriz de covarianzas de las variables originales.

3. La primera componente principal está asociada al mayor autovalor de la matriz de covarianzas de las variables originales.

4. La varianza de cada componente es igual al autovalor al que va asociada.

5. En el caso de que las variables estén tipificadas (se aplica el PCA sobre la matriz de correlaciones como haremos en el caso 11.1), la proporción de la variabilidad total de las variables originales captada por una componente es igual al autovalor correspondiente dividido por el número de variables originales.

6. La correlación entre una componente y una variable original se determina con el autovalor de la componente y el correspondiente elemento del autovector asociado, en el caso de que las variables originales estén tipificadas

11.4.1. Obtención de la primera componente

Consideremos que se dispone de una muestra de tamaño n acerca de las p siguientes variables $X_1, X_2, \ldots, X_p$ y que las observaciones están expresadas bien en desviaciones respecto a la media (centradas) o bien como variables tipificadas.

La primera componente, de la misma forma que el resto, se expresa como combinación lineal de las variables originales:

$$Z_{1i} = u_{11}X_{1i} + u_{12}X_{2i} + \cdots + u_{1p}X_{pi} \tag{11.9}$$

Puede comprobarse fácilmente que tanto si las X_j variables están expresadas en desviaciones como si están tipificadas se obtiene que la media muestral de Z_1 es igual a 0.

Para el conjunto de las n observaciones muestrales esta ecuación se puede expresar matricialmente de la siguiente forma:

$$\begin{bmatrix} Z_{11} \\ Z_{12} \\ \vdots \\ Z_{1n} \end{bmatrix} = \begin{bmatrix} X_{11} & X_{21} & \cdots & X_{p1} \\ X_{12} & X_{22} & \cdots & X_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ X_{1n} & X_{2n} & \cdots & X_{pn} \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{1p} \end{bmatrix} \tag{11.10}$$

o en notación matricial compacta:

$$\mathbf{z_1} = \mathbf{X}\mathbf{u_1} \tag{11.11}$$

Estadísticos e-Books & Papers

La primera componente se obtiene de forma que su varianza sea máxima, sujeta a la restricción de que la suma de los pesos u_{1j} al cuadrado sea igual a la unidad.

La varianza del primer componente, teniendo en cuenta que su media es 0, vendrá dada por:

$$var\,(Z_1) = \frac{\sum_{i=1}^{n} Z_{1j}^2}{n} = \frac{1}{n}\mathbf{z_1}'\mathbf{z_1} = \frac{1}{n}\mathbf{u_1}'\mathbf{X}'\mathbf{X}\mathbf{u_1} = \mathbf{u_1}'\left[\frac{1}{n}\mathbf{X}'\mathbf{X}\right]\mathbf{u_1} \qquad (11.12)$$

Si las variables están expresadas en desviaciones respecto a la media, $(1/n)\mathbf{X}'\mathbf{X}$ es la matriz de covarianzas muestral a la que denominaremos $\mathbf{V}$; para variables tipificadas $(1/n)\mathbf{X}'\mathbf{X}$ es igual a la matriz de correlaciones $\mathbf{R}$. Las componentes pueden obtenerse para ambos tipos de variables, aunque en los paquetes de ordenador para análisis multivariante se utiliza la matriz de correlación en la mayor parte de los casos. A efectos expositivos y para darle mayor generalidad vamos a suponer, sin embargo, que utilizamos la matriz de covarianzas. Por lo tanto, la varianza a maximizar es:

$$var(Z_1) = \mathbf{u_1}'\mathbf{V}\mathbf{u_1} \qquad (11.13)$$

La restricción señalada analíticamente de que la suma de los cuadrados de los pesos sea la unidad viene dada por:

$$\sum_{j=1}^{p} u_{1i}^2 = \mathbf{u_1}'\mathbf{u_1} = 1 \qquad (11.14)$$

En consecuencia, incorporando esta restricción se forma el siguiente lagrangiano:

$$L = \mathbf{u_1}'\mathbf{V}\mathbf{u_1} - \lambda(\mathbf{u_1}'\mathbf{u_1} - 1) \qquad (11.15)$$

Para maximizar el valor del lagrangiano derivamos respecto a $\mathbf{u_1}$ e igualamos a 0:

$$\frac{\partial L}{\partial \mathbf{u_1}} = 2V\mathbf{u_1} - 2\lambda\mathbf{u_1} = \mathbf{0} \qquad (11.16)$$

es decir:

$$(\mathbf{V} - \lambda\mathbf{I})\mathbf{u_1} = \mathbf{0} \qquad (11.17)$$

Al resolver la ecuación $|\mathbf{V} - \lambda\mathbf{I}| = 0$, se obtienen p autovalores. Si se toma la raíz característica mayor (λ_1), se halla el autovector asociado a la misma $\mathbf{u_1}$, aplicando la regla de normalización dada en (11.14). Así pues, el vector de ponderaciones que se aplica a las variables iniciales para obtener la primera

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

componente principal es el autovector asociado al mayor autovalor de la matriz $\mathbf{V}$.

11.4.2. Obtención de las restantes componentes

Una componente genérica expresada en forma matricial para todas las observaciones viene dada, de forma análoga a (11.11), por:

$$\mathbf{z_h} = \mathbf{X}\mathbf{u_h} \tag{11.18}$$

Para su obtención, además de la restricción de:

$$\mathbf{u_h'}\mathbf{u_h} = 1 \tag{11.19}$$

se imponen las restricciones adicionales ya expuestas de que:

$$\mathbf{u_h'}\mathbf{u_1} = \mathbf{u_h'}\mathbf{u_2} = \cdots = \mathbf{u_h'}\mathbf{u_{h-1}} = \mathbf{0} \tag{11.20}$$

Por lo tanto, se imponen las restricciones adicionales de que el vector característico asociado a la componente h-ésima sea ortogonal a todos los vectores característicos calculados previamente. Su obtención, aparte de tener en cuenta las restricciones (11.20), no plantea problemas conceptuales nuevos. En todo caso señalaremos que el vector $\mathbf{u_h}$ está asociado a la raíz h-ésima, una vez ordenadas de mayor a menor.

En definitiva, las p componentes principales que se pueden calcular son una combinación lineal de las variables originales, en la que los coeficientes de ponderación son los correspondientes vectores característicos asociados a la matriz $\mathbf{V}$.

11.4.3. Varianzas de las componentes

Si se tienen en cuenta las propiedades de los autovalores y como ilustramos en el caso de dos componentes, la varianza de la componente h-ésima coincide con su autovalor, es decir:

$$Var(Z_h) = \mathbf{u_h'}\mathbf{V}\mathbf{u_h} = \lambda_h \tag{11.21}$$

Adoptando como medida de la variabilidad de las variables originales la suma de sus varianzas, dicha variabilidad será igual a la traza de la matriz $\mathbf{V}$. Ahora bien, la traza de la matriz $\mathbf{V}$ se puede expresar así teniendo en cuenta :

$$\frac{\lambda_h}{\text{traza}\mathbf{V}} = \frac{\lambda_h}{\sum_{h=1}^{p} \lambda_h} \tag{11.22}$$

y en el caso particular en que los datos estuvieran tipificados y la matriz de covarianzas fuera la de correlación $\mathbf{R}$, entonces dado que traza($\mathbf{R}$) $= p$, la

Estadísticos e-Books & Papers

proporción de variabilidad correspondiente a la componente h-ésima es:

$$\frac{\lambda_h}{p} \tag{11.23}$$

11.4.4. Correlación entre las componentes principales y las variables originales

Antes de obtener el coeficiente de correlación de cada componente con cada variable, vamos a calcular la covarianza entre la variable X_j y el componente Z_h .

Definamos los vectores muestrales de la componente Z_h y variable X_j por:

$$\mathbf{x_j} = \begin{bmatrix} X_{j1} \\ X_{j2} \\ \vdots \\ X_{jn} \end{bmatrix} \qquad \mathbf{z_h} = \begin{bmatrix} Z_{h1} \\ Z_{h2} \\ \vdots \\ Z_{hn} \end{bmatrix}$$

La covarianza muestral entre X_j y Z_h viene dada por:

$$\text{cov}(X_j, Z_h) = \frac{1}{n}\mathbf{x_j}'\mathbf{z_h} \tag{11.24}$$

El vector $\mathbf{x_j}$ se puede expresar en función de la matriz $\mathbf{X}$, utilizando el vector de orden p, al que designaremos por δ, que tiene un 1 en la posición j-ésima y 0 en las posiciones restantes. Así,

$$\mathbf{x_j}' = \boldsymbol{\delta}'\mathbf{X} = \begin{bmatrix} 0 & \cdots & 1 & \cdots & 0 \end{bmatrix} \begin{bmatrix} X_{11} & \cdots & X_{1i} & \cdots & X_{1n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{j1} & \cdots & X_{ji} & \cdots & X_{jn} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{p1} & \cdots & X_{pi} & \cdots & X_{pn} \end{bmatrix} \tag{11.25}$$

Teniendo en cuenta la expresión anterior y (11.18), la covarianza dada en (11.24) puede expresarse de la siguiente forma:

$$\text{cov}(X_j, Z_h) = \frac{1}{n}\boldsymbol{\delta}'\mathbf{X}'\mathbf{X}\mathbf{u_h} = \boldsymbol{\delta}'\mathbf{V}\mathbf{u_h} = \boldsymbol{\delta}'\lambda_h\mathbf{u_h} = \lambda_h\boldsymbol{\delta}'\mathbf{u_h} = \lambda_h u_{hj} \tag{11.26}$$

En consecuencia, la correlación existente entre la variable X_j y la componente Z_h, que recordemos denominábamos **carga factorial**, es la siguiente:

$$l_{jh} = \frac{cov\,(X_j, Z_h)}{\sqrt{var(X_j)}\sqrt{var(Z_h)}} = \frac{\lambda_h u_{hj}}{\sqrt{var(X_j)}\sqrt{\lambda_h}} = \frac{u_{hj}}{\sqrt{var(X_j)}}\sqrt{\lambda_h} \tag{11.27}$$

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

expresión que coincide con la que derivamos en su momento (11.8). En el caso en que las variables estén tipificadas (varianza 1), la expresión anterior adopta la forma:

$$l_{hj} = u_{hj}\sqrt{\lambda_h} \tag{11.28}$$

que es la que se corresponde con las llamadas *matrices factoriales* —*factor matrix*— en la mayoría de programas estadísticos.

11.4.5. Puntuaciones sin tipificar y tipificadas

Una vez calculados los coeficientes u_{hj}, se pueden obtener las puntuaciones Z_{hi}, es decir, los valores de las componentes correspondientes a cada observación, a partir de la siguiente relación:

$$Z_{hi} = u_{h1}X_{1i}+u_{h2}X_{2i}+\cdots+u_{hp}X_{pi} \quad h = 1, 2, \cdots, p \quad i = 1, 2, \cdots, n \tag{11.29}$$

Si una componente se divide por su desviación típica se obtiene una componente tipificada. Así, designando por Y_h a la componente h-ésima tipificada, esta viene definida por:

$$Y_h = \frac{Z_h}{\sqrt{\lambda_h}} \tag{11.30}$$

de tal forma que las puntuaciones tipificadas serían sustituyendo en (11.29):

$$\frac{Z_{hi}}{\sqrt{\lambda_h}} = \frac{u_{h1}}{\sqrt{\lambda_h}}X_{1i} + \frac{u_{h2}}{\sqrt{\lambda_h}}X_{2i} + \cdots + \frac{u_{hp}}{\sqrt{\lambda_h}}X_{pi} \tag{11.31}$$

o alternativamente:

$$Y_{hi} = c_{hi}X_{1i} + c_{h2}X_{2i} + \cdots + c_{hp}X_{pi} \tag{11.32}$$

A la matriz formada por los coeficientes c definidos en (11.32) se la denomina **matriz de coeficientes para el cálculo de las puntuaciones en las componentes** en la mayoría de programas estadísticos.

A continuación abordaremos una serie de cuestiones que son centrales en la aplicación de un PCA, como qué criterios podemos seguir para establecer el número de componentes que debemos retener en la interpretación del PCA —recordemos que hay que buscar un equilibrio entre facilidad de interpretación, sin pérdida de excesiva información—, cómo interpretar esas componentes principales, efecto del tipo de datos (centrados, tipificados) sobre el PCA o cómo pueden utilizarse las puntuaciones en las componentes en análisis ulteriores. Pero, para ilustrar algunas de estas cuestiones y tener un caso completo de PCA, lo ilustraremos mediante el caso 11.1.

Estadísticos e-Books & Papers

Cuadro 11.8.: Variables de la base de datos de empleo

Variable	Etiqueta	Definición: porcentaje de empleados en...
x_1	Agricultura	Agricultura
x_2	Minería	Minería
x_3	Industria	Industria
x_4	Energía	Industrias de generación de energía
x_5	Construcción	Construcción
x_6	ServiciosInd	Servicios a la industria
x_7	Finanzas	Sector financiero
x_8	ServiciosPer	Servicios a la sociedad y a las personas
x_9	Transporte	Transporte y las comunicaciones

Fuente: Hand *et al.* (1994, p. 303)

Caso 11.1. Estructura sectorial del empleo en Europa
Hand *et al.* (1994) dan los datos de la distribución del empleo por sectores en una serie de países europeos[1]. Las variables son las recogidas en el cuadro 11.8. El objetivo es determinar si existen perfiles de empleo que sirvan además para categorizar a los países en función de los mismos.

11.5. Aspectos operativos en la estimación de un PCA

11.5.1. Efecto del tipo de datos sobre el análisis de componentes principales

A lo largo de la presentación del PCA hemos alternado las derivaciones de expresiones donde los datos estaban centrados en la media y cómo cambiaban las mismas cuando los datos están tipificados. Los resultados pueden diferir si la varianza de las variables es muy diferente. Como señala Sharma (1996), la varianza de las variables afecta a los resultados de un PCA y siempre tendrá una mayor influencia en la generación de un componente dado aquella variable con más varianza.

Esto hace que la recomendación sea la de que, si el investigador no tiene ninguna razón para querer atribuir más importancia a unas variables que a otras, se recurra siempre a la estimación basándose en valores tipificados, es decir, basándose en la matriz de correlaciones. Si solicitamos los descriptivos para las variables del caso 11.1, observamos en el cuadro 11.9 que la varianza es muy grande, variando desde 0,14 en la energía hasta 214,6 en la agricultura, lo que confirma la necesidad de trabajar con correlaciones.

[1]Puede accederse a los datos en formato electrónico en `http://www.stat.ncsu.edu/research/sas/sicl/data/`

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 11.9.: Descriptivos de las variables del caso 11.1

```
         Pais Agricultura   Mineria  Industria    Energía Construccion ServiciosInd
median     NA  14.4500000 0.8500000 27.5500000 0.80000000    8.4000000   14.5000000
mean       NA  18.5208333 1.1750000 26.8416667 0.88333333    8.3250000   13.3041667
SE.mean    NA   2.9903830 0.1932043  1.2920032 0.07770873    0.3188095    0.9270000
CI.mean    NA   6.1860786 0.3996735  2.6727123 0.16075276    0.6595077    1.9176456
var        NA 214.6173732 0.8958696 40.0625362 0.14492754    2.4393478   20.6238949
std.dev    NA  14.6498250 0.9465039  6.3294973 0.38069349    1.5618412    4.5413539
coef.var   NA   0.7909917 0.8055352  0.2358087 0.43097377    0.1876085    0.3413482
          Finanzas ServiciosPer Transporte
median   4.6500000   19.6500000  6.7000000
mean     3.8125000   20.5500000  6.5750000
SE.mean  0.4944369    1.3037849  0.2639753
CI.mean  1.0228206    2.6970846  0.5460746
var      5.8672283   40.7965217  1.6723913
std.dev  2.4222362    6.3872155  1.2932097
coef.var 0.6353406    0.3108134  0.1966859
```

```
library(pastecs)
stat.desc(datos,basic=FALSE)
```

Si solicitamos y analizamos la matriz de correlaciones (figura 11.6) observamos, por ejemplo, fuertes correlaciones negativas entre la agricultura y el resto de sectores (salvo la minería) lo que ya nos hace intuir un efecto de sustitución entre estos dos sectores y el resto. Países con más empleo agrícola serán probablemente países con menos empleo en servicios, por ejemplo. Esto debería manifestarse en un primer componente donde las cargas tuvieran pesos con signo contrario. También se observan fuertes correlaciones positivas entre los sectores de servicios (finanzas, personales y la industria).

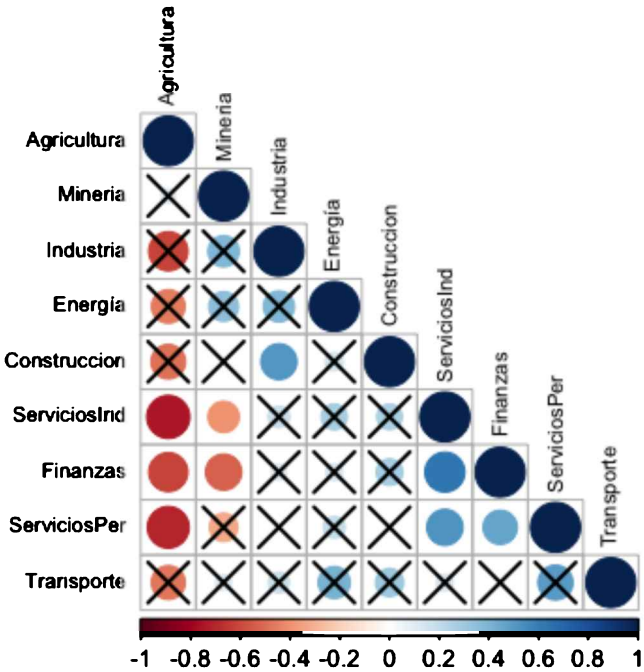
```
library(corrplot)
R<-cor(datos[2:10], method="pearson")
corrplot(R, p.mat = res1[[1]], sig.level=0.05,typ="lower")
```

Suele ser habitual en esta fase realizar el test de esfericidad de Bartlett (Bartlett, 1951) para descartar que la matriz de correlaciones sea la matriz identidad, test que ya presentamos en el tema 9. De ser la matriz identidad (covarianzas nulas), no habría agrupaciones de variables (correlaciones) que justificaran la existencia de un componente principal. El test de esfericidad de Bartlett rara vez no reachaza esta hipótesis nula. Así ocurre en nuestro caso (cuadro 11.10).

```
library(psych)
cortest.bartlett(R,n=24)
```


Estadísticos e-Books & Papers

Figura 11.6.: Matriz de correlaciones entre las variables del caso y significatividad (p<0,05).



Nota: Las aspas (X) señalan correlaciones no significativas

Cuadro 11.10.: Test de esfericidad de Bartlett

```
> cortest.bartlett(R,n=24)
$chisq
[1] 252.0011

$p.value
[1] 3.135063e-34

$df
[1] 36
```

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

11.5.2. Número de componentes principales que extraer

Como acabamos de comentar, la decisión a tomar es cuánta información (varianza) estamos dispuestos a sacrificar para ganar facilidad de interpretación. Obviamente es una cuestión de juicio del investigador y las reglas serán solo orientativas. Veamos las más habituales.

A. Criterio del autovalor superior a la unidad

Cuando se trabaja con variables tipificadas, el criterio de retener aquellas componentes cuyo autovalor sea superior a la unidad (Kaiser, 1960) es la opción por defecto de la mayoría de programas estadísticos. Recordemos que el autovalor es la varianza extraída por cada componente. Parece lógico que esta varianza sea, como mínimo, la correspondiente a la de una variable. Si una componente no es capaz de explicar más información que una variable, no va a facilitar la reducción de datos, es decir, facilitar la interpretabilidad de la información. Veámoslo de una manera analítica sencilla. El criterio planteado de manera general es retener aquellas componentes cuyos autovalores sean superiores a la media de los autovalores:

$$\lambda_h > \overline{\lambda} = \frac{\sum_{j=1}^{p} \lambda_j}{p} \tag{11.33}$$

pero, como cuando trabajamos con variables tipificadas, se verifica que la suma de los autovalores coincide con la traza de la matriz de correlaciones, que es p, entonces:

$$\sum_{j=1}^{p} \lambda_j = p \tag{11.34}$$

por lo que el criterio (11.33) se puede expresar como decíamos al principio:

$$\lambda_h > 1 \tag{11.35}$$

Los estudios de Cattell y Jaspers (1967), Browne (1968) y Linn (1968) demuestran que es un criterio que funciona bastante bien salvo con gran número de variables (>40) siendo especialmente preciso con un número pequeño de variables (10-15). La varianza de una variable explicada por el componente (comunalidad) también influye en la precisión del criterio. Cliff (1988) señala que es importante utilizar este criterio en combinación con otros para asegurarnos de que no estamos reteniendo demasiadas —o demasiadas pocas— componentes.

Si estimamos el PCA del caso 11.1 mediante la función `PCA{FactoMineR}`, el cuadro 11.11 nos ofrece los autovalores. De acuerdo con el criterio expuesto deberían retenerse los tres primeros.

Estadísticos e-Books & Papers

Cuadro 11.11.: Aplicación del criterio del autovalor >1

```
$eig
        eigenvalue percentage of variance cumulative percentage of variance
comp 1 3.6009354526           4.001039e+01                          40.01039
comp 2 2.1113460064           2.345940e+01                          63.46979
comp 3 1.2103485274           1.344832e+01                          76.91811
comp 4 0.8351719711           9.279689e+00                          86.19780
comp 5 0.5207887756           5.786542e+00                          91.98434
comp 6 0.3368747540           3.743053e+00                          95.72739
comp 7 0.2552492568           2.836103e+00                          98.56350
comp 8 0.1292414244           1.436016e+00                          99.99951
comp 9 0.0000438317           4.870189e-04                         100.00000
```

```
fit<-PCA(datos[2:10],scale.unit=TRUE,ncp=9,graph=TRUE)
head(fit)
```

B. Criterio del gráfico de sedimentación (*scree plot*)

Cattell (1966) propone representar en un gráfico los autovalores correspondientes a las distintas componentes. Como los autovalores siempre son decrecientes, lo será este gráfico de líneas. El criterio que propone Cattell (1966) es retener aquellas componentes anteriores a donde la línea comienza a nivelarse (si fuera la ladera de una montaña y cayesen rocas, donde estas comenzarían a pararse o sedimentar, de ahí el nombre del gráfico). Stevens (2009) señala que este criterio, si no se usa con precaución, pueden llevar a descartar componentes que, aunque añaden poca varianza explicada, pueden ser relevantes bajo la perspectiva de la relevancia teórica de las mismas. Hay distintos trabajos que han analizado la precisión del criterio. Tucker *et al.* (1969) encontraron que era preciso en 12 de 18 casos analizados, Linn (1968) en 7 de cada 10 y Cattell y Jaspers (1967) en 6 de 8. Un estudio posterior más extenso, el de Hakstian *et al.* (1982), concluye que tanto la regla del autovalor superior a la unidad como el gráfico de sedimentación son precisos para $N > 250$ y niveles altos de comunalidad ($\geq$ 0,60), pero que, cuando la comunalidad media cae por debajo de 0,30 o la ratio entre el número de factores y el número de variables es superior a 0,30, ambos son bastante imprecisos.

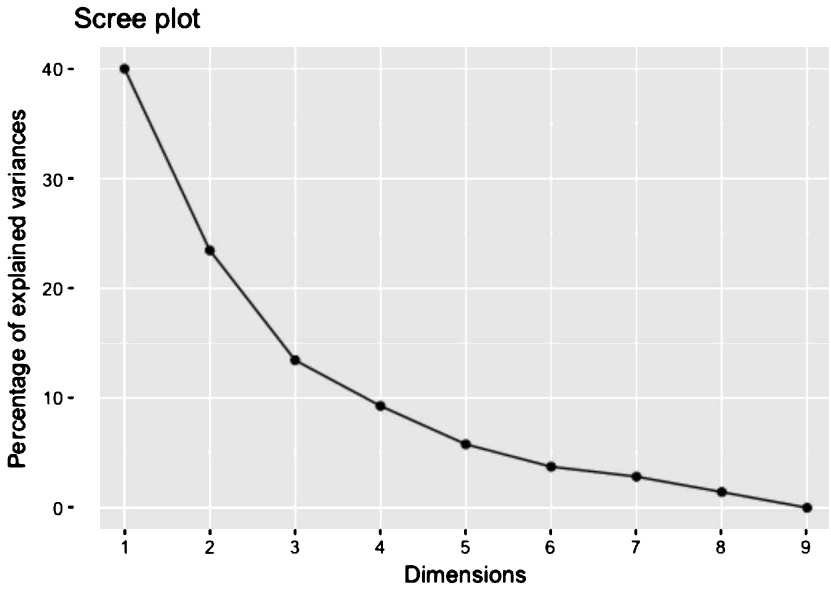
La estimación anterior que hemos realizado del PCA y que hemos guardado en el objeto `fit` nos ofrece directamente el gráfico de sedimentación, pero la función `fviz_eig {factoextra}` permite una representación más estética del mismo, que es la que mostramos en la figura 11.7.

```
library(factoextra)
fviz_eig(fit, geom="line")+
theme_grey()
```

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 11.7.: Gráfico de sedimentación



C. Método paralelo

Es obvio que hay mucha subjetividad en decidir dónde se produce el "codo" en la línea que marcaría el corte, sobre todo cuando la caída es progresiva. En la figura 11.7 no tendríamos muy claro si considerar tres o cuatro componentes al menos. Por esta razón Horn (1965) sugirió un procedimiento llamado análisis paralelo para objetivar este criterio, que ha sido operativizado en `R` por Dinno (2009) mediante la función `paran{paran}`. De acuerdo con Horn (1965), cuando se tienen datos totalmente incorrelacionados, un PCA debería generar autovalores iguales a 1 (cada autovalor explica la varianza de una única variable). Sin embargo, los errores de muestreo generan correlaciones espúreas que hace que unos autovalores sean superiores a la unidad y otros inferiores. La estrategia de Horn (1965) es contrastar los autovalores obtenidos mediante un PCA paralelo aplicado a muestras aleatorias de variables no correlacionadas con el mismo número de variables y observaciones que la muestra original que generarán autovalores que se han ajustado para evitar el error muestral. Solo se retendrían aquellas componentes con autovalores ajustados superiores a la unidad. El autovalor ajustado vendrá dado por la expresión:

$$\lambda_p - (\bar{\lambda}_p^r - 1) \tag{11.36}$$

donde λ_p es el autovalor p-ésimo de los datos originales y $\bar{\lambda}_p^r$ es la media de las estimaciones de ese mismo autovalor en las submuestras aleatorias generadas.

Estadísticos e-Books & Papers

Cuadro 11.12.: Análisis paralelo de Horn (1965)

```
Results of Horn's Parallel Analysis for component retention
5000 iterations, using the mean estimate

--------------------------------------------------
Component   Adjusted    Unadjusted    Estimated
            Eigenvalue  Eigenvalue    Bias
--------------------------------------------------
1           2.520492    3.600935      1.080443
2           1.447376    2.111346      0.663969
--------------------------------------------------

Adjusted eigenvalues > 1 indicate dimensions to retain.
(2 components retained)
```

Solicitando el análisis paralelo mediante la mencionada función `paran{paran}` del cuadro 11.12 y la figura 11.8 se confirma que el número de componentes que debería retenerse es dos.

```
library(paran)
paran(datos[2:10], iterations=5000,graph=TRUE,color=FALSE)
```

D. Test estadístico

Tanto Sharma (1996) como Stevens (2009) consideran que el test estadístico que presentaremos a continuación, que está basado en el test de Bartlett (1951) para analizar si la matriz de correlaciones es la identidad, adolece de las mismas limitaciones —básicamente ser muy sensible al tamaño muestral— y no recomiendan su uso pues suele aconsejar la retención de demasiadas componentes principales.

Se puede considerar que, digamos, los $p-m$ últimos autovalores poblacionales son iguales a 0. Si los autovalores muestrales que observamos correspondientes a estas componentes no son exactamente igual a 0, se debe a los problemas del azar. Por ello, bajo el supuesto de que las variables originales siguen una distribución normal multivariante, se pueden formular las siguientes hipótesis relativas a los autovalores poblacionales:

$$H_0: \ \lambda_{m+1} = \lambda_{m+2} = \cdots = \lambda_p = 0 \tag{11.37}$$

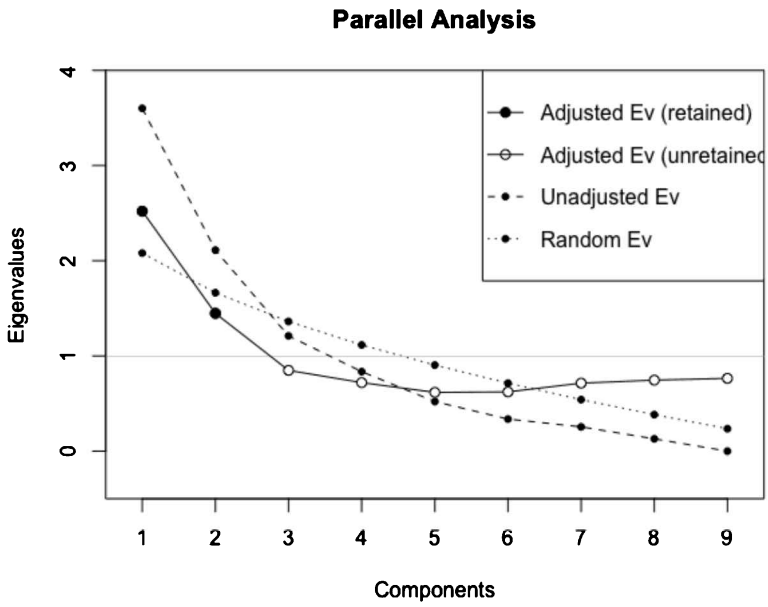
El estadístico para el contraste de esta hipótesis es el siguiente:

$$Q^* = \left\{ n - \frac{2p+11}{6} \right\} \left\{ (p-m) \ln \bar{\lambda}_{p-m} - \sum_{j=m+1}^{p} \ln \lambda_j \right\} \tag{11.38}$$

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 11.8.: Gráfico del análisis paralelo



Estadísticos e-Books & Papers

Cuadro 11.13.: Estadísticos de Bartlett (1950), Anderson (1963) y Lawley (1956)

```
  v       values     bartlett bartlett.chi bartlett.df   bartlett.p anderson.chi
1 1 3.6009354526 1.949601e-06    252.00115          36 3.135063e-34     315.5493
2 2 2.1113460064 1.258062e-05    208.74203          35 1.318235e-26     270.8005
3 3 1.2103485274 5.086216e-05    176.30731          27 7.298187e-24     237.2734
4 4 0.8351719711 1.229943e-04    154.55790          20 8.393238e-23     216.0809
5 5 0.5207887756 2.680282e-04    135.70291          14 5.058974e-22     197.3860
6 6 0.3368747540 4.604008e-04    121.65404           9 6.125556e-22     184.4019
7 7 0.2552492568 6.866135e-04    110.47004           5 3.259571e-22     174.8097
8 8 0.1292414244 1.355664e-03     95.75023           2 1.614739e-21     158.4831
9 9 0.0000438317 1.000000e+00      0.00000           0 1.000000e+00       0.0000
  anderson.df   anderson.p lawley.chi lawley.df     lawley.p
1          44 9.802519e-43  252.00115        36 3.135063e-34
2          35 3.100517e-38  209.02313        28 1.327676e-29
3          27 1.656624e-35  176.63469        21 1.337038e-26
4          20 7.219677e-35  155.00334        15 2.439725e-25
5          14 1.877130e-34  136.13522        10 2.606488e-24
6           9 6.097913e-35  122.16242         6 5.724939e-24
7           5 6.865097e-36  111.02336         3 6.608308e-24
8           2 3.853226e-35   96.01988         1 1.137357e-22
9           0 1.000000e+00    0.00000         0 1.000000e+00

$nFactors
bartlett anderson   lawley
       8        8        8
```

Bajo la hipótesis nula, el estadístico anterior se distribuye como una ji-cuadrado con $(p-m+2)(p-m+1)/2$ grados de libertad. Existen algunas variantes de este test básicamente referidas al cálculo de los grados de libertad, como son las de Anderson (1963) y Lawley (1956).

Una vez más, afortunadamente, porque el cálculo manual es complejo, la función `nBartlett{nFactors}` implementa el test original y sus variaciones con una sintaxis muy sencilla. El cuadro 11.13 ofrece los resultados y confirma la limitación a la que aludíamos en la medida en que recomienda retener 8 componentes en cualquiera de sus versiones.

```
library(nFactors)
nBartlett(R, N=24, alpha=0.01, cor=TRUE, details=TRUE)
```

Con todos los criterios expuestos nos inclinaremos por retener dos componentes principales que, como se aprecia en el cuadro 11.11, explicarían el 63,47 % de la información.

11.5.3. Interpretación de las componentes principales

Para interpretar las dos componentes extraídas es necesario fijarse en la contribución de cada variable. Al trabajar con datos tipificados, los valores estandarizados (cargas) y sin estandarizar coinciden. Cuanto mayor es la carga, mayor es la influencia que ha tenido esa variable en la formación de la componente. Por lo tanto podemos analizar cuáles son las cargas más altas y usar las variables a las que corresponden para dar una interpretación al eje. Si nos fijamos en el cuadro 11.14 vemos que la primera componente está muy correlacionada de manera negativa con la agricultura y en menor medida con la minería.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 11.14.: Correlaciones de las variables con las componentes (cargas)

```
$var$cor
                   Dim.1      Dim.2
Agricultura  -0.9755452 -0.0909600
Mineria      -0.2119704  0.8680503
Industria     0.4967865  0.6451644
Energía       0.4913828  0.5202499
Construccion  0.5166372  0.3259305
ServiciosInd  0.7857710 -0.3276420
Finanzas      0.6915111 -0.4852026
ServiciosPer  0.7029305 -0.3210492
Transporte    0.5093258  0.3325506
```

Las otras correlaciones son todas ellas positivas. Esto nos llevaría a interpretar esta componente como la que distingue a países con sistemas económicos muy basados en el sector primario frente a economías más industriales y de servicios.

La segunda componente correlaciona de manera positiva la minería, la industria, la energía, la construcción y el transporte y, de manera negativa, los servicios, ya sean a la industria, personales o financieros, por lo que nos llevaría a interpretarla como aquella que contrapone países con un mayor o menor desarrollo en su sector servicios.

Aunque el cuadro es bastante explicativo, puede facilitarse la interpretación representando las coordenadas estandarizadas de las variables sobre el mapa que marcan las componentes. La función `PCA{FactoMineR}` lo hace por defecto, pero siempre puede generarse un mapa más visual mediante el módulo de gráficos `ggplot2`, que extrae los datos directamente del objeto `fit` en el que se ha guardado la estimación. También la función `fviz_pca_var{factoextra}` es tremendamente interesante porque permite combinar el sentido de la correlación (carga) con la intensidad de la contribución de cada variable. Mostramos en la figura 11.9 las dos versiones que confirman la interpretación efectuada con una primera componente, que contrapone la agricultura al resto de sectores, y una segunda, que contrapone los servicios a los sectores industriales.

```
library(ggplot2)
datos.grafico2<-data.frame(fit$var$coord[,1:2])
ggplot(datos.grafico2)+
  geom_point(aes(x=Dim.1, y=Dim.2,colour="darkred"))+
  geom_text_repel(aes(x=Dim.1, y=Dim.2),
label=rownames(datos.grafico2))+
  geom_vline(xintercept = 0,colour="darkgray")+
  geom_hline(yintercept = 0,colour="darkgray")+
 labs (x="Dimension 1 (40.01%)", y="Dimension 2 (23.46%)")+
theme(legend.position="none")
```


Estadísticos e-Books & Papers

```
library(factoextra)
fviz_pca_var(fit, col.var="contrib")+
scale_color_gradient2(low="white", mid="blue",high="red",
midpoint=10.0)+
theme_gray()
```

Una vez interpretados los ejes, el último paso es representar a los objetos —los países— sobre ese mapa mediante las puntuaciones en las componentes que hemos derivado en (11.32). La estimación efectuada nos muestra estas coordenadas en el cuadro 11.15. Estos datos pueden visualizarse muy fácilmente en un gráfico, mostrando cómo efectivamente la ordenación de países sobre la componente 1 va desde los más agrícolas, como Turquía o Grecia, hasta los que tienen economías más desarrolladas en industria y servicios, pero el eje 2 ordena a estos últimos entre aquellos donde el peso de la industria es mayor (Hungría, Chequia, Rusia) y donde están más desarrollados los servicios (Suecia, Holanda, Dinamarca).

```
datos.grafico<-data.frame(fit$ind$coord[,1:2],datos$Pais)
colnames(datos.grafico)<-c("Dim.1","Dim.2","pais")
ggplot(datos.grafico)+
  geom_point(aes(x=Dim.1, y=Dim.2,colour="darkred"))+
  geom_text_repel(aes(x=Dim.1, y=Dim.2),
label=datos.grafico$pais)+
  geom_vline(xintercept = 0,colour="darkgray")+
  geom_hline(yintercept = 0,colour="darkgray")+
  labs (x="Dimension 1 (40.01%)", y="Dimension 2 (23.46%)")+
  theme(legend.position="none")
```

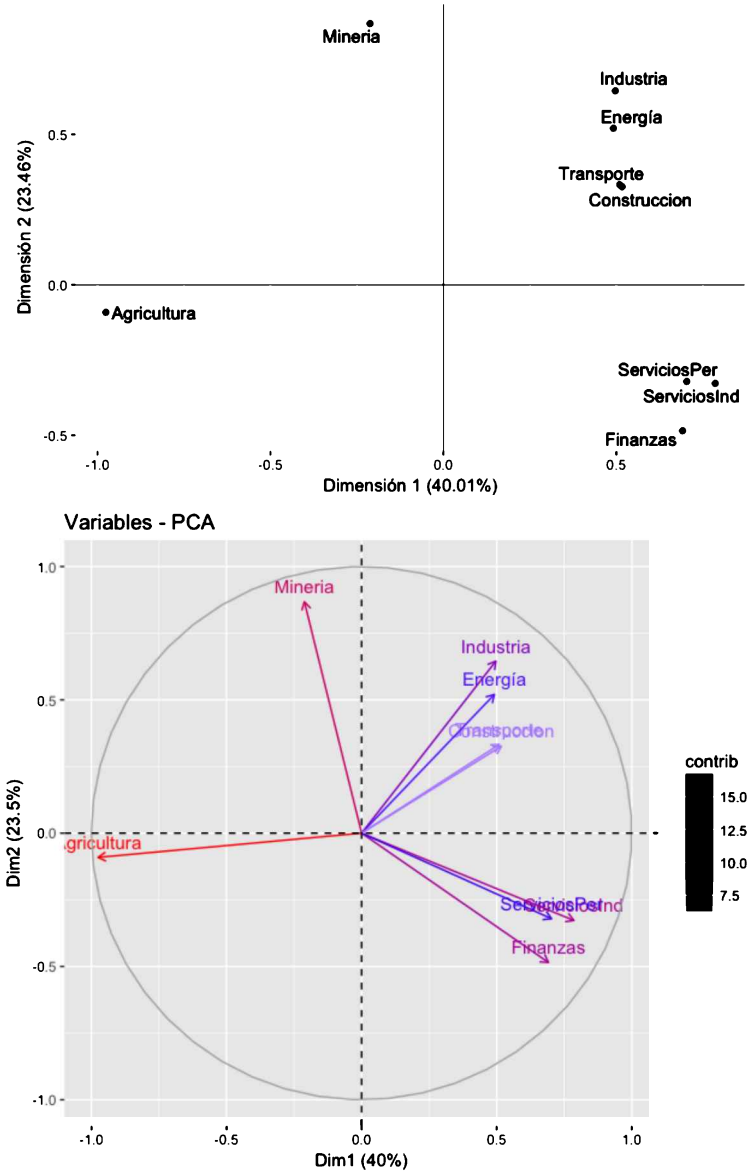
Parece obvio el interés de tener una representación conjunta de países y sectores sobre las dos componentes que nos facilite la interpretación sin tener que leer dos mapas por separado. La función `ggbiplot{ggbiplot}` nos lo permite con una sintaxis muy sencilla. La figura 11.11 nos ofrece este resultado, que no requiere de interpretación adicional a la ya expuesta.

```
library(ggbiplot)
ggbiplot(fit, obs.scale = 1, var.scale = 1)+
  scale_color_discrete(name = '')+
  expand_limits(x=c(-8,4), y=c(-2.5, 2.5))+
  labs (x="Dimension 1 (40.01%)", y="Dimension 2 (23.46%)")+
  geom_text_repel(aes(x=datos.grafico$Dim.1,
y=datos.grafico$Dim.2),
label=datos.grafico$pais,size=3)
```

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 11.9.: Gráfico de las cargas sobre las dos primeras componentes

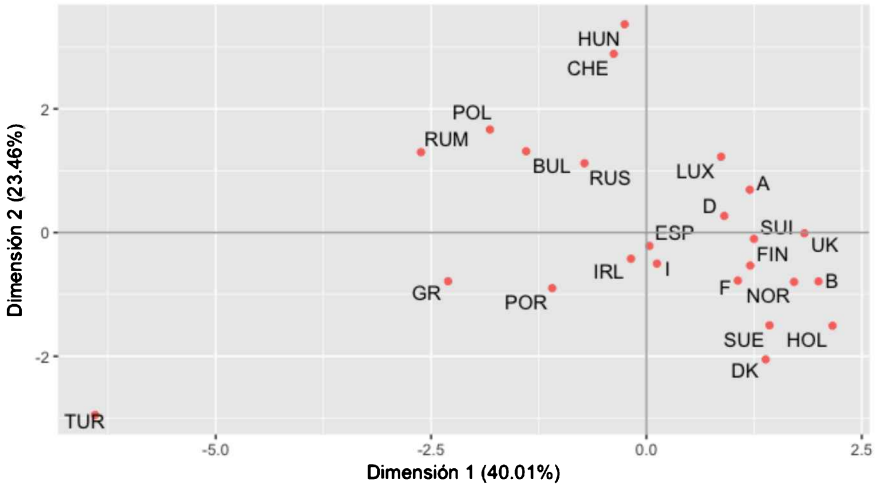


Estadísticos e-Books & Papers

Cuadro 11.15.: Coordenadas de los objetos (países) en las primeras dos componentes

```
$ind
$ind$coord
          Dim.1       Dim.2
1   1.99780226 -0.78982212
2   1.38457811 -2.05049829
3   1.06147821 -0.77650965
4   0.90375834  0.26924665
5  -0.17940569 -0.42351385
6   0.12172438 -0.50144570
7   0.86646019  1.22579378
8   2.16093232 -1.50684192
9   1.83416374 -0.01158207
10  1.19946200  0.69199606
11  1.20516629 -0.53409903
12 -2.30260616 -0.78756629
13  1.71397585 -0.79741293
14 -1.09454386 -0.89891361
15  0.03466063 -0.21484221
16  1.42863354 -1.49972760
17  1.24855216 -0.10230254
18 -6.40146995 -2.94888181
19 -1.39648891  1.31481253
20 -0.37964211  2.88945506
21 -0.25291724  3.36744945
22 -1.81688720  1.66510653
23 -2.61788461  1.29940635
24 -0.71950229  1.12069321
```

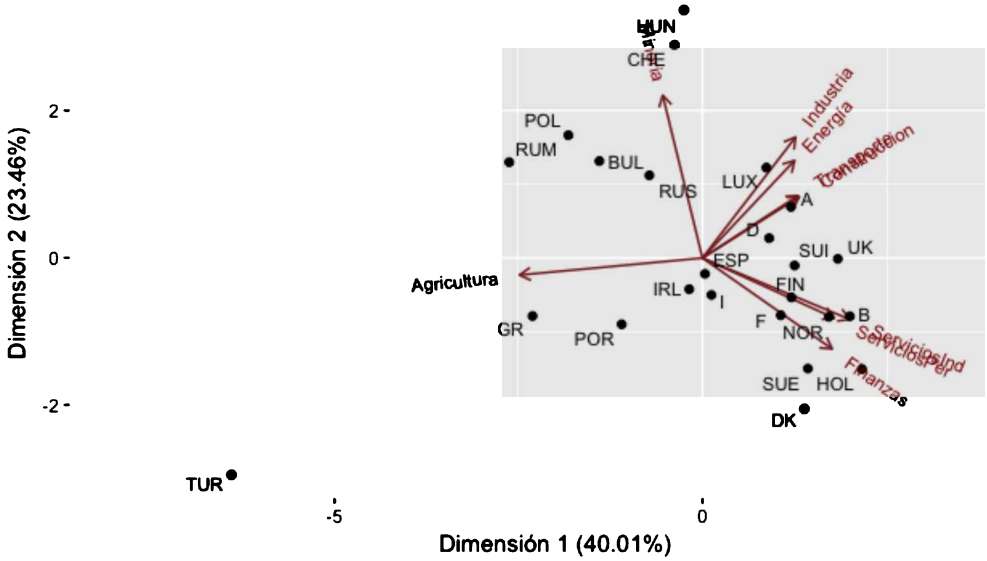
Figura 11.10.: Representación gráfica de los países sobre las primeras dos componentes



 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 11.11.: Representación gráfica conjunta de los países y de los sectores sobre las primeras dos componentes



Estadísticos e-Books & Papers

Estadísticos e-Books & Papers

12. Análisis factorial exploratorio

12.1. Introducción

Quizás la primera tarea para una buena presentación del análisis factorial exploratorio (EFA) sea que el lector sea capaz de distinguir la herramienta con nitidez del análisis de componentes principales (PCA) que presentamos en el tema anterior. Y la tarea no es sencilla porque hay mucha confusión al respecto. Gran parte de esta confusión nace del hecho de que la mayoría de programas estadísticos tiene a los componentes principales como la opción por defecto de extracción de factores y, como señalan Osborne y Banjanovic (2016), porque muchos investigadores utilizan ambas herramientas indistintamente o, matizan, usan PCA cuando la herramienta más apropiada es un EFA.

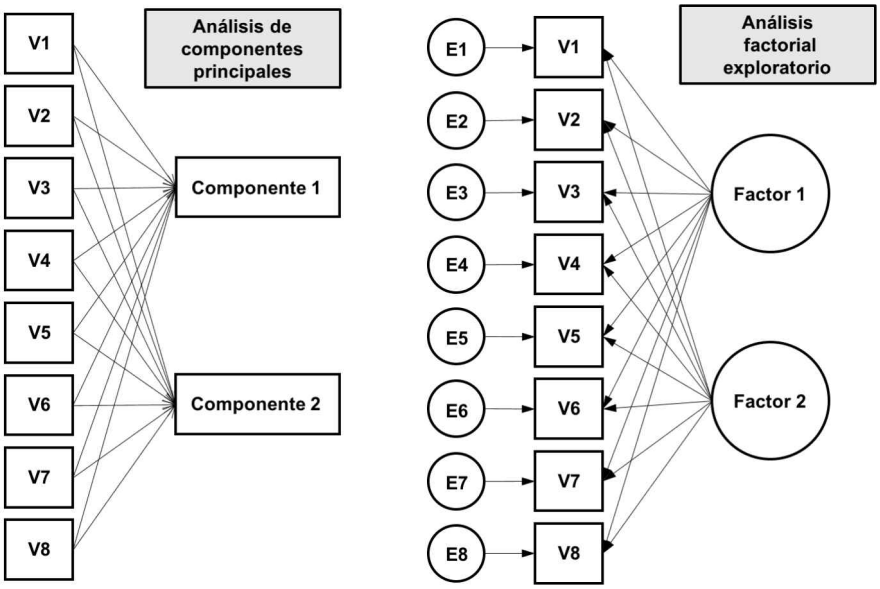
Aunque luego entraremos en cuestiones de fondo, el PCA es una versión computacionalmente más sencilla del EFA. El EFA fue desarrollado previamente al PCA (Hotelling, 1933) gracias a los trabajos de Spearman (1904). Pero en esta época previa a los ordenadores, el planteamiento del EFA era demasiado exigente para los cálculos manuales que se realizaban y el PCA se generó como una alternativa menos costosa en términos de esfuerzo de cálculo (Gorsuch, 1990).

Como hemos visto en el capítulo anterior, al PCA no le importa cuál es la estructura latente de las variables, es decir, si hay factores que están provocando que esas variables estén correlacionadas entre sí. Para el PCA las variables son en sí mismas el objeto del interés, no su estructura subyacente. En buena medida esto convierte al PCA en una herramienta similar a la regresión, al generar combinaciones lineales ponderadas de las variables. Como se ilustra en la figura 12.1, la dirección de las flechas va de las variables hacia las dos componentes extraídas, el 100 % de lo que contienen las componentes principales procede de las variables, están completamente definidas por ellas. En este sentido se dice que los parámetros del PCA pretenden reproducir las características de la muestra más que las de la población (Thompson, 2004).

El objetivo del EFA, sin embargo, es otro. Lo que busca es intentar detectar si hay variables latentes (no observadas) que explican por qué las variables manifiestas (indicadores) están correlacionadas entre sí y pueden agruparse en un proceso de reducción de datos. En la figura 12.1 vemos varias cosas: (a) los factores los representamos como círculos para ilustrar su carácter latente frente al manifiesto u observado (cuadrados) de las variables. Las flechas van desde los factores hacia los indicadores en la medida en que asumimos que hacen que se comporten (correlacionen) entre ellos de una manera determinada

Estadísticos e-Books & Papers

Figura 12.1.: Visión conceptual diferencial del PCA y del EFA



Fuente: Osborne y Banjanovic (2016).

(varianza común). Y también asumimos que los factores no explican todo el comportamiento de las variables, motivo por el cual incorporamos un término de error que recoge la parte del comportamiento de los indicadores no explicado por los factores (varianza específica o residual).

Puede seguirse a Osborne y Banjanovic (2016) para una aproximación al debate abierto sobre el uso de una u otra herramienta que ellos intentan sintetizar con la opinión de Widaman (1993, p. 263) "...[el] PCA nunca debería ser utilizado si un investigador quiere parámetros que reflejen variables latentes o factores" y concluyen que, aunque siga siendo la opción por defecto en muchos programas estadísticos para la reducción de datos, no es —en su opinión, remarcan— la opción conceptualmente más deseable y no tiene una ventaja clara que puedan destacar.

Como viene siendo habitual en capítulos precedentes, comenzaremos el capítulo mostrando el desarrollo matemático del EFA basándonos en una ilustración sencilla (caso 12.1) que facilite el cálculo manual de los distintos elementos que lo componen para, posteriormente, profundizar en los pasos operativos que hay que dar para estimar un EFA de manera adecuada, basándonos entonces en un caso más complejo (caso 12.2).

Caso 12.1. Tributo a Spearman (1904)

Como hemos apuntado con anterioridad, los orígenes del análisis factorial exploratorio cabe buscarlos en los esfuerzos de psicólogos como Charles Spearman

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 12.1.: Matriz de correlaciones de las notas de 220 estudiantes en 6 asignaturas

	Gaélico	Inglés	Historia	Aritmética	Álgebra	Geometría
Gaélico	1,000					
Inglés	,439	1,000				
Historia	,410	,351	1,000			
Aritmética	,288	,354	,164	1,000		
Álgebra	,329	,320	,190	,595	1,000	
Geometría	,248	,329	,181	,470	,464	1,000

Fuente: Lawley y Maxwell (1971)

o Karl Pearson para definir y medir la inteligencia. El objeto básico del EFA es describir las covarianzas entre variables observadas en función de otras no observables que subyacen bajo ellas, denominadas factores. Si un grupo de variables manifiestas guarda una fuerte correlación entre ellas pero a su vez la correlación con otro grupo de variables es relativamente baja, es sensato pensar que cada grupo pueda ser reflejo de un factor subyacente que cause ese comportamiento diferenciado.

Lawley y Maxwell (1971) presentan la matriz de correlaciones entre las notas de $n = 220$ estudiantes en $p = 6$ asignaturas, que recoge el cuadro 12.1. La cuestión que se plantea es si esas notas pueden estar causadas por un único factor subyacente, que sería lo que podría denominarse *inteligencia general,* o existen varios factores que pueden explicar distintos tipos de inteligencia.

12.2. Formulación del modelo de análisis factorial exploratorio

A lo largo de este capítulo vamos a considerar que las variables observables x_i son variables tipificadas, es decir, tienen media 0 y varianza 1. El modelo de análisis factorial exploratorio puede definirse del siguiente modo:

$$x_1 = \lambda_{11}\xi_1 + \lambda_{12}\xi_2 + \cdots \lambda_{1m}\xi_m + \varepsilon_1 \tag{12.1}$$

$$x_2 = \lambda_{21}\xi_1 + \lambda_{22}\xi_2 + \cdots \lambda_{2m}\xi_m + \varepsilon_2$$

$$\cdots$$

$$x_p = \lambda_{p1}\xi_1 + \lambda_{p2}\xi_2 + \cdots \lambda_{pm}\xi_m + \varepsilon_p$$

donde $\xi_1, \xi_2, \cdots, \xi_m$ son factores comunes, $\varepsilon_1, \varepsilon_2, ..., \varepsilon_p$ son factores únicos o específicos o errores, λ_{jh} es el peso del factor h en la variable j. A los coeficientes

Estadísticos e-Books & Papers

de este tipo se les denomina **cargas factoriales**.

En el modelo (12.1), cada una de las p variables observables es una combinación lineal de m factores comunes ($m < p$) y de un factor único. Así pues, todas las variables originales vienen influidas por todos los factores comunes, mientras que existe un factor único que es específico para cada variable. Debe tenerse en cuenta que tanto los factores comunes como los factores únicos no son observables.

Las ecuaciones del modelo se pueden expresar matricialmente de la siguiente forma:

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1m} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{pm} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix} \tag{12.2}$$

o en forma matricial compacta:

$$\mathbf{x} = \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\varepsilon} \tag{12.3}$$

12.2.1. Hipótesis del modelo

Para poder realizar inferencias a partir del modelo (12.3) es preciso formular hipótesis estadísticas sobre los factores comunes y sobre los factores únicos.

Las **hipótesis sobre los factores comunes** son las siguientes. La esperanza de cada uno de los factores comunes es nula:

$$E(\boldsymbol{\xi}) = \mathbf{0} \tag{12.4}$$

y la matriz de covarianzas de los factores comunes es la matriz identidad, lo que implica que la varianza de cada uno de los factores es 1 y que los factores están incorrelacionados entre sí, ya que todos los elementos de fuera de la diagonal principal son nulos.

$$E(\boldsymbol{\xi}\boldsymbol{\xi}') = \mathbf{I} \tag{12.5}$$

Así pues, los factores comunes son variables tipificadas de media 0 y varianza 1, y que además no están correlacionadas entre sí.

Las hipótesis sobre los factores únicos son las siguientes. La esperanza de cada uno de los factores únicos es nula, es decir:

$$E(\boldsymbol{\varepsilon}) = \mathbf{0} \tag{12.6}$$

La matriz de covarianzas de los factores es una matriz diagonal que denotaremos como $\boldsymbol{\Psi}$:

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \boldsymbol{\Psi} \tag{12.7}$$

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

La matriz de covarianzas de los factores únicos es una matriz diagonal, lo que implica que las varianzas de los factores únicos pueden ser distintas y también que los factores están incorrelacionados entre sí. Debe tenerse en cuenta que en una matriz diagonal todos los elementos de fuera de la diagonal principal son nulos.

La **hipótesis que se postula sobre la relación entre factores comunes y factores únicos** es la siguiente. La matriz de covarianzas entre los factores comunes y los factores únicos es nula:

$$E(\boldsymbol{\xi}\boldsymbol{\varepsilon}) = \mathbf{0} \tag{12.8}$$

Para poder realizar inferencias que permitan distinguir, para cada variable, entre los factores comunes y el factor único, es necesario postular que los primeros estén incorrelacionados con este último, tal como se establece en la hipótesis (12.8).

12.2.2. Propiedades del modelo

Dado que las variables x_i son variables tipificadas, su matriz de covarianzas es igual a la matriz de correlación poblacional $\mathbf{R_p}$, es decir:

$$E(x'x) = \begin{bmatrix} 1 & \rho_{21} & \cdots & \rho_{p1} \\ \rho_{21} & 1 & \cdots & \rho_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix} \tag{12.9}$$

Como se trata de variables tipificadas, la varianza de cada una de ellas es igual a 1. De acuerdo con el modelo (12.3) y teniendo en cuenta las hipótesis (12.4) a (12.8), la matriz de correlación poblacional (12.9) se puede descomponer de la siguiente forma[1]:

$$\mathbf{R_p} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi} \tag{12.10}$$

Como puede verse, en esta descomposición $\boldsymbol{\Lambda}'\boldsymbol{\Lambda}$ es la parte correspondiente a los factores comunes y $\boldsymbol{\Psi}$ es justamente la matriz de covarianzas de los factores únicos. La descomposición anterior puede expresarse de forma detallada de la siguiente forma:

[1] Demostración:
$R_p = E(\mathbf{x}\mathbf{x}') = E(\boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\varepsilon})(\boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\varepsilon})'$
$= \boldsymbol{\Lambda}E(\boldsymbol{\xi}\boldsymbol{\xi}')\boldsymbol{\Lambda}' + E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') + \boldsymbol{\Lambda}E(\boldsymbol{\xi}\boldsymbol{\varepsilon}') + E(\boldsymbol{\varepsilon}\boldsymbol{\xi}')\boldsymbol{\Lambda}'$
$= \boldsymbol{\Lambda}\mathbf{I}\boldsymbol{\Lambda}' + \boldsymbol{\Psi} + \mathbf{L}\mathbf{0} + \mathbf{0}\mathbf{L}'$
$= \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$

Estadísticos e-Books & Papers

$$\begin{bmatrix} 1 & \rho_{21} & \cdots & \rho_{p1} \\ \rho_{21} & 1 & \cdots & \rho_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix} =$$

$$\begin{bmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1m} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{pm} \end{bmatrix} \begin{bmatrix} \lambda_{11} & \lambda_{21} & \cdots & \lambda_{p1} \\ \lambda_{12} & \lambda_{22} & \cdots & \lambda_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_{1m} & \lambda_{2m} & \cdots & \lambda_{pm} \end{bmatrix} + \tag{12.11}$$

$$+ \begin{bmatrix} \varepsilon_1^2 & 0 & \cdots & 0 \\ 0 & \varepsilon_2^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \varepsilon_p^2 \end{bmatrix}$$

De acuerdo con (12.10), el primer elemento de la diagonal principal del primer miembro, que es la varianza de la variable tipificada de x_1, puede descomponerse de la siguiente forma:

$$1 = \lambda_{11}^2 + \lambda_{12}^2 + \cdots + \lambda_{1m}^2 + \varepsilon_1^2 \tag{12.12}$$

Análogamente, de forma genérica, la varianza de la variable tipificada x_j se puede descomponer de la siguiente forma:

$$1 = \underbrace{\lambda_{j1}^2 + \lambda_{j2}^2 + \cdots + \lambda_{jm}^2}_{\text{comunalidad}} + \varepsilon_j^2 \tag{12.13}$$

Vemos que (12.13) nos está diciendo que la varianza correspondiente a una variable x_j tiene dos partes, por un lado, la explicada por los factores comunes $\lambda_{j1}^2 + \lambda_{j2}^2 + \cdots + \lambda_{jm}^2$ y, por otra parte, la varianza no explicada, residual o específica ε_j^2. Pues bien, la parte de la varianza de la variable explicada por los factores se denomina **comunalidad.** Si la denotamos como h_j^2, entonces (12.13) puede ponerse como:

$$1 = h_j^2 + \varepsilon_j^2 \tag{12.14}$$

A veces, por paralelismo, el término ε_j^2 se denomina **especificidad**. El concepto de comunalidad es muy importante, nos indica —como hemos señalado— la parte de la varianza de cada variable que es explicada por todos los factores comunes. Es un concepto que distingue al EFA del PCA. En el EFA las comunalidades son siempre más pequeñas que 1 para cada variable porque, como es evidente de la expresión anterior, una parte de la varianza es residual, mientras que en el PCA ese residuo no existe e inicialmente son 1. El objeto del EFA será, entonces, ser capaces de descomponer la comunalidad entre los factores

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

comunes para saber la contribución específica de cada uno (λ_{jh}^2) a la varianza de la variable.

El coeficiente de correlación entre cualquiera dos variables puede expresarse en función, también, de las cargas factoriales, como se desprende de la expresión (12.11):

$$\rho_{hj} = \lambda_{h1}\lambda_{j1} + \lambda_{h2}\lambda_{j2} + \cdots + \lambda_{hm}\lambda_{jm} \tag{12.15}$$

Pues bien, llega el momento de ver los distintos procedimientos que existen para descomponer la comunalidad entre los distintos factores. Analizaremos también el problema que se deriva de la potencial ausencia de una solución a esta descomposición. Pero para que el seguimiento del proceso sea más sencillo lo ilustraremos con los datos del caso 12.1.

12.3. Métodos para la extracción de factores

El planteamiento del problema es teóricamente sencillo. Si observamos la expresión (12.10), todo se reduce a calcular las matrices Λ y Ψ a partir de las correlaciones muestrales de la matriz $\mathbf{R}$[2]. Pero lo que parece sencillo no siempre lo es, y en el caso del EFA, surgen dos problemas. Veámoslos sobre la estructura de nuestro ejemplo.

12.3.1. Limitaciones a la extracción de factores

El primer problema es la **limitación del número de grados de libertad** que puede provocar la imposibilidad de encontrar una solución al problema factorial. En el caso 12.1, la expresión (12.3), denotando como x_1 a la nota en gaélico, x_2 a la nota en inglés, x_3 a la de historia, x_4 a la de aritmética, x_5, a álgebra y x_6, a geometría se concreta en:

$$\begin{aligned}
x_1 &= \lambda_{11}\xi_1 + \lambda_{12}\xi_2 + \lambda_{13}\xi_3 + \lambda_{14}\xi_4 + \lambda_{15}\xi_5 + \lambda_{16}\xi_6 + \varepsilon_1\\
x_2 &= \lambda_{21}\xi_1 + \lambda_{22}\xi_2 + \lambda_{23}\xi_3 + \lambda_{24}\xi_4 + \lambda_{25}\xi_5 + \lambda_{26}\xi_6 + \varepsilon_2\\
x_3 &= \lambda_{31}\xi_1 + \lambda_{32}\xi_2 + \lambda_{33}\xi_3 + \lambda_{34}\xi_4 + \lambda_{35}\xi_5 + \lambda_{36}\xi_6 + \varepsilon_3\\
x_4 &= \lambda_{41}\xi_1 + \lambda_{42}\xi_2 + \lambda_{43}\xi_3 + \lambda_{44}\xi_4 + \lambda_{45}\xi_5 + \lambda_{46}\xi_6 + \varepsilon_4\\
x_5 &= \lambda_{51}\xi_1 + \lambda_{52}\xi_2 + \lambda_{53}\xi_3 + \lambda_{54}\xi_4 + \lambda_{55}\xi_5 + \lambda_{56}\xi_6 + \varepsilon_5\\
x_6 &= \lambda_{61}\xi_1 + \lambda_{62}\xi_2 + \lambda_{63}\xi_3 + \lambda_{64}\xi_4 + \lambda_{65}\xi_5 + \lambda_{66}\xi_6 + \varepsilon_6
\end{aligned}$$

Donde deberemos decidir cuántos de los 6 factores comunes cabe retener y estimar las cargas factoriales. La expresión (12.10), en general, tendrá los siguientes componentes:

[2]Estrictamente hablando, la expresión (12.10) contiene parámetros poblacionales desconocidos y nosotros debemos realizar su estimación, lo que nos obligaría a cambiar la notación, por ejemplo las correlaciones muestrales deberían denotarse como r_{12} por ejemplo o las cargas como $\hat{\lambda}_{12}$ y su matriz $\hat{\mathbf{\Lambda}}$. Pero, por simplicidad y para facilitar el seguimiento sin completar la notación, mantendremos la notación poblacional para representar también a los parámetros.

Estadísticos e-Books & Papers

- El segundo miembro tendrá $p \times p$ ecuaciones que se corresponderán con una ecuación para cada uno de los elementos de `R`, que es cuadrada. En nuestro ejemplo $6 \times 6 = 36$.
- Pero la matriz $\mathbf{R}$ es simétrica, es decir, no tiene 36 elementos distintos, dado que el triángulo superior coincide con el inferior, por lo tanto nos generará solo $p(p+1)/2$ elementos distintos, en nuestro ejemplo 21 correlaciones diferentes.
- Con esos datos habrá que estimar $p \times m$ cargas factoriales donde m es variable porque será el número de factores que finalmente retengamos (en nuestro ejemplo m puede variar desde 1 hasta 6) a los que hay que añadir p términos de error, en nuestro caso 6.

Para que el proceso de estimación pueda tener lugar se requiere que el número de ecuaciones sea mayor o igual al número de parámetros que hay que estimar, por tanto:

$$p(p+1)/2 \geq p \times m + p = p(m+1) \tag{12.16}$$

Como en nuestro ejemplo el número original de variables es $p = 6$, contamos con 21 correlaciones diferentes, por lo que a partir de $m = 3$ factores extraídos tendríamos un problema si utilizamos un método de extracción de factores donde la restricción de los grados de libertad se aplique, como pueden ser los de máxima verosimilitud. Sin embargo en aquellos en los que la extracción se basa en la obtención de los autovalores, como es el de componentes principales o el de ejes principales, esta restricción no aplicará.

Otro problema es el de la **no unicidad de la solución** debido al problema de la rotación de los ejes. Supongamos que nuestro problema del caso 12.1 tiene una solución adecuada de dos factores. Consideremos estas dos soluciones alternativas recogidas en el cuadro 12.2.

Bajo la columna `PC1` y `PC2` aparecen las estimaciones de las cargas factoriales para los dos factores extraídos, y `RC1` y `RC2` son las cargas factoriales en otra solución rotada obtenida a partir de la misma matriz de correlaciones. La columna `h2` es la comunalidad de cada variable explicada por la solución factorial y `u2` la especificidad. ¿Qué podemos comprobar? Que dos soluciones factoriales distintas generan la misma comunalidad total para cada variable, la misma unicidad y ambas soluciones proceden de la misma matriz de correlaciones. Si, además, interpretáramos la solución, veríamos que en el primer caso podríamos decantarnos por un único factor que podríamos denominar inteligencia general porque hay cargas muy altas sobre el primer factor, mientras que en la segunda solución parece que las notas en aritmética, álgebra y geometría se agrupan en el primer factor y las de gaélico, inglés e historia en el segundo, con lo que la interpretación diferiría.

Este es un problema sin solución, es decir, no habrá nunca una solución única a un problema de EFA. Como señala Sharma (1996), la solución a este problema

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 12.2.: Soluciones factoriales alternativas al caso 12.1

```
No rotada

     PC1   PC2   h2   u2 com
Gae 0.66  0.45 0.63 0.37 1.8
Eng 0.69  0.29 0.56 0.44 1.3
His 0.52  0.64 0.67 0.33 1.9
Ari 0.74 -0.41 0.72 0.28 1.6
Alg 0.74 -0.38 0.69 0.31 1.5
Geo 0.68 -0.36 0.59 0.41 1.5

Rotada

     RC1  RC2   h2   u2 com
Gae 0.22 0.76 0.63 0.37 1.2
Eng 0.35 0.66 0.56 0.44 1.5
His 0.00 0.82 0.67 0.33 1.0
Ari 0.83 0.15 0.72 0.28 1.1
Alg 0.81 0.18 0.69 0.31 1.1
Geo 0.75 0.16 0.59 0.41 1.1
```

pasa por la interpretación de cuál de las dos soluciones es más plausible, más coherente con la teoría o los a priori del investigador.

Planteadas las limitaciones a la solución de un EFA, estamos en condiciones de abordar distintos procedimientos para la extracción de los factores. Si se quiere una discusión más profunda de los procedimientos de extracción de factores, puede consultarse Harman (1976), Rummel (1970) o McDonald (1985).

12.3.2. Método de las componentes principales

La lógica del método de las componentes principales la presentamos en el tema anterior, por lo que vamos a dar otro enfoque para añadir valor al anterior. El método de las componentes principales puede considerarse un problema de autovalores y autovectores. Nosotros partimos de la matriz de correlaciones entre las variables $\mathbf{R}$. Las matrices de covarianzas y de correlaciones están entre las que pueden diagonalizarse (es decir, tomar valores en la diagonal y ceros en el resto de la matriz). En nuestro problema esto se expresaría:

$$\mathbf{L} = \mathbf{V}'\mathbf{R}\mathbf{V} \tag{12.17}$$

Donde $\mathbf{L}$ sería la mencionada matriz diagonal —a los valores de su diagonal se les denomina **autovalores**— y a las columnas de $\mathbf{V}$ se las denomina **autovectores**. Pero ¿qué tiene esto que ver con nuestro problema del EFA? La expresión (12.17) puede reorganizarse del siguiente modo:

Estadísticos e-Books & Papers

$$\mathbf{R} = \mathbf{VLV}' \tag{12.18}$$

es decir, la matriz de correlaciones puede ponerse como el producto de tres matrices, conteniendo autovalores y autovectores. Es inmediato que:

$$\begin{aligned} \mathbf{R} &= \mathbf{V}\sqrt{\mathbf{L}}\sqrt{\mathbf{L}}\mathbf{V}' \\ \mathbf{R} &= (\mathbf{V}\sqrt{\mathbf{L}})(\sqrt{\mathbf{L}}\mathbf{V})' \end{aligned} \tag{12.19}$$

Y si llamamos a $\mathbf{\Lambda} = \mathbf{V}\sqrt{\mathbf{L}}$, entonces:

$$\mathbf{R} = \mathbf{\Lambda}\mathbf{\Lambda}' \tag{12.20}$$

donde (12.20) coincide con la expresión (12.10) que sintetizaba nuestro problema del EFA con el matiz de no incorporar la varianza residual, aunque esta se puede obtener por diferencia entre la varianza total y la explicada por la solución factorial. Por lo tanto, la clave está en esta expresión:

$$\mathbf{\Lambda} = \mathbf{V}\sqrt{\mathbf{L}} \tag{12.21}$$

Obtener las cargas factoriales se reduce a obtener los autovectores $\mathbf{V}$ y los autovalores $\mathbf{L}$ de la matriz de correlaciones $\mathbf{R}$. Veámoslo para nuestro caso 12.1, pero no lo hagamos estimando un EFA, hagámoslo manualmente mediante el cálculo de autovalores y autovectores con `R` y luego veremos cómo el resultado coincide con la estimación obtenida mediante la función correspondiente de `R` para estimar EFA.

Definimos la matriz $\mathbf{R}$ y calculamos los autovalores y autovectores con la función **`svd{base}`**:

```
r <- c(1.000,
       .439,1.000,
       .410,.351,1.000,
       .288,.354,.164,1.000,
       .329,.320,.190,.595,1.000,
       .248,.329,.181,.470,.464,1.000)
#Convertimos el vector r en la matriz R
R<-lav_matrix_lower2full(r)
#Etiquetamos a las variables de la matriz
colnames(R) <- rownames(R) <-
c("Gae","Eng", "His","Ari","Alg","Geo")
#Autovalores y autovectores
sdv(R)
```

Para obtener la matriz con las cargas basta multiplicar la matriz que contiene los autovectores izquierdos (`$u` en el cuadro 12.3) por la raíz cuadrada de los autovalores (`$d` en el mismo cuadro), de acuerdo con la expresión (12.21). Dado que la función **`sdv{base}`** genera un vector con los autovalores y no una

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 12.3.: Autovalores y autovectores

```
$d
[1] 2.7328841 1.1297704 0.6151739 0.6012219 0.5247969 0.3961529

$u
           [,1]       [,2]          [,3]        [,4]       [,5]        [,6]
[1,] -0.3979211  0.4224739  0.2378797843 -0.44697886  0.6205763 -0.14725080
[2,] -0.4164293  0.2732052  0.6497850613  0.40590321 -0.3699525  0.16763850
[3,] -0.3129591  0.5996225 -0.6713466822  0.09898664 -0.2855052 -0.02217036
[4,] -0.4466126 -0.3885864 -0.0008312812 -0.23224145 -0.3517809 -0.68691430
[5,] -0.4499766 -0.3532281 -0.1360853358 -0.40242630 -0.1217941  0.69097829
[6,] -0.4103173 -0.3340032 -0.2279611918  0.64013367  0.5078616 -0.02051222

$v
           [,1]       [,2]          [,3]        [,4]       [,5]        [,6]
[1,] -0.3979211  0.4224739  0.2378797843 -0.44697886  0.6205763 -0.14725080
[2,] -0.4164293  0.2732052  0.6497850613  0.40590321 -0.3699525  0.16763850
[3,] -0.3129591  0.5996225 -0.6713466822  0.09898664 -0.2855052 -0.02217036
[4,] -0.4466126 -0.3885864 -0.0008312812 -0.23224145 -0.3517809 -0.68691430
[5,] -0.4499766 -0.3532281 -0.1360853358 -0.40242630 -0.1217941  0.69097829
[6,] -0.4103173 -0.3340032 -0.2279611918  0.64013367  0.5078616 -0.02051222
```

Cuadro 12.4.: Matriz $\mathbf{\Lambda}$ con las cargas factoriales

```
> LAMBDA
6 x 6 Matrix of class "dgeMatrix"
           [,1]       [,2]         [,3]        [,4]        [,5]        [,6]
[1,] -0.6578207  0.4490503  0.186576299 -0.34658069  0.44956326 -0.09268065
[2,] -0.6884175  0.2903916  0.509646048  0.31473125 -0.26800419  0.10551281
[3,] -0.5173663  0.6373427 -0.526557479  0.07675275 -0.20682814 -0.01395417
[4,] -0.7383147 -0.4130310 -0.000651999 -0.18007653 -0.25484019 -0.43234852
[5,] -0.7438760 -0.3754484 -0.106735839 -0.31203531 -0.08823116  0.43490643
[6,] -0.6783135 -0.3550142 -0.178796848  0.49635004  0.36790950 -0.01291053
```

matriz diagonal, como la que necesitamos en la expresión (12.21), generamos una matriz diagonal con los autovalores en la diagonal y cero en el resto de elementos mediante la función `Diagonal{Matrix}`. El cuadro 12.4 nos ofrece la matriz $\mathbf{\Lambda}$ con las cargas factoriales que buscábamos.

```
av<-Diagonal(6,auto$d)
LAMBDA=auto$u%*%sqrt(av)
```

Dado que todavía no hemos discutido un criterio para decidir cuántos de los 6 factores de la solución hemos de retener, asumamos que son 2 de ellos y veamos algunas de las propiedades que cumple la solución que ofrecimos al presentar el modelo del EFA y también los valores para elementos claves como la comunalidad, la especificidad o la correlación entre las variables. La solución derivada del cuadro 12.4 para dos factores puede escribirse (redondeando con dos decimales):

Estadísticos e-Books & Papers

Cuadro 12.5.: Solución de dos factores para el caso 12.1 con extracción por componentes principales

	Cargas		Comunalidad	Especificidad	Varianza compartida	
	ξ_1	ξ_2	h_i^2	ε_i	ξ_1	ξ_2
(x_1) Gaélico	-0.66	0.45	0.63	0.37	0.43	0.20
(x_2) Inglés	-0.69	0.29	0.56	0.44	0.47	0.08
(x_3) Historia	-0.52	0.64	0.67	0.33	0.27	0.41
(x_4) Aritmética	-0.74	-0.41	0.72	0.28	0.55	0.17
(x_5) Álgebra	-0.74	-0.38	0.69	0.31	0.55	0.14
(x_6) Geometría	-0.68	-0.36	0.59	0.41	0.46	0.13
Suma			3.86	2.14	2.73	1.13
Porcentaje			64 %	36 %	45 %	19 %

$$
\begin{aligned}
x_1 &= -0{,}66\xi_1 + 0{,}45\xi_2 + \varepsilon_1 \\
x_2 &= -0{,}69\xi_1 + 0{,}29\xi_2 + \varepsilon_2 \\
x_3 &= -0{,}52\xi_1 + 0{,}64\xi_2 + \varepsilon_3 \\
x_4 &= -0{,}74\xi_1 - 0{,}41\xi_2 + \varepsilon_4 \\
x_5 &= -0{,}74\xi_1 - 0{,}38\xi_2 + \varepsilon_5 \\
x_6 &= -0{,}68\xi_1 - 0{,}36\xi_2 + \varepsilon_6
\end{aligned}
$$

Nótese que ahora el término de error tiene todo el sentido, porque, por ejemplo, para x_1 contendría toda la varianza que correspondería a los factores $\xi_3 \cdots \xi_6$ que no hemos considerado en la solución de dos factores[3]:

$$\varepsilon_1 = 0{,}19\xi_3 - 0{,}35\xi_4 + 0{,}45\xi_5 - 0{,}09\xi_6$$

Veamos alguno de los cálculos de variables fundamentales para el análisis que hemos presentado anteriormente aplicados a este ejemplo y sintetizados en el cuadro 12.5.

La suma de los autovalores es la **varianza total del modelo**. Como las variables están tipificadas (proceden de una matriz de correlaciones), su varianza es 1 y como hay 6 variables la varianza total es 6. Con los autovalores que muestra el cuadro 12.3:

$$2{,}73 + 1{,}13 + 0{,}62 + 0{,}60 + 0{,}52 + 0{,}40 = 6{,}00$$

Una solución de dos factores recoge el 64 % de la varianza $(2{,}73 + 1{,}13)/6$ quedando el otro 36 % como residual al recoger los autovalores de factores no incluidos en la solución.

[3]Esto solo es estrictamente así en la descomposición mediante componentes principales porque, como vimos en la figura 12.1, no contemplan otro tipo de error residual, en otros procedimientos ese error no solo procedería de los factores no contemplados sino de otras fuentes.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

La comunalidad es la varianza de cada variable explicada por la solución factorial. Es simplemente la suma de los cuadrados de las cargas factoriales de cada variable en los factores de la solución. Por ejemplo para x_1:

$$h_1^2 = (-0{,}66)^2 + 0{,}45^2 = 0{,}64$$

es decir, la solución de dos factores explica el 64 % del comportamiento de la variable x_1. La especificidad o varianza residual no explicada por la solución factorial para la variable x_1 puede obtenerse de dos formas en la extracción por componente principales: como la diferencia hasta 1 de la comunalidad, $1 - 0{,}64 = 0{,}36$, o mediante las cargas factoriales que tenían los factores no incluidos en la solución (cuadro 12.3):

$$\varepsilon_1 = 0{,}19^2 + (-0{,}35)^2 + 0{,}45^2 + (-0{,}09)^2 = 0{,}36$$

Finalmente también puede comprobarse como la correlación entre cualquiera dos variables es igual al producto cruzado de sus respectivas estructuras de cargas. Así, por ejemplo en el cuadro 12.1 vemos como la correlación entre las variables x_1 y x_2 es $\rho_{12} = 0{,}439$, pues bien a partir de la estructuras de cargas factoriales de estas dos variables —contando todos los factores de $\xi_1 ... \xi_6$, no solo los dos de la solución, pueden verse en el cuadro12.4— obtendríamos:

$$\begin{aligned}\rho_{12} = (-0{,}66)(-0{,}69) + (0{,}45)(0{,}29) + (0{,}19)(0{,}51) + \\ +(-0{,}35)(0{,}31) + (0{,}45)(-0{,}27) + (-0{,}09)(0{,}11) = 0{,}439\end{aligned}$$

Un concepto importante posteriormente será el de **matriz de correlaciones reproducida**. Vemos que la correlación entre dos variables ha sido el producto cruzado de sus cargas, pero contando todos los factores posibles. Sin embargo, si consideramos una solución con solo dos factores de los seis posibles, la correlación reproducida será:

$$\rho_{12}^* = (-0{,}66)(-0{,}69) + (0{,}45)(0{,}29) = 0{,}58$$

La cercanía entre la matriz de correlaciones muestral y la reproducida (por ejemplo, la suma de las diferencias entre ellas o residuos), puede convertirse en un indicador del ajuste o calidad de la representación del modelo factorial.

Efectuado el cálculo manual de la extracción por componentes principales, veamos como obtenerlo directamente mediante la función `principal{psych}`, de momento asumiremos que la solución de dos factores es la adecuada, aunque más adelante analizaremos los criterios para tomar esta decisión:

```
fit.pca<-principal(R,nfactors=2,rotate="none") print(fit.pca)
```

Vemos en el cuadro 12.6, que toda la información coincide —redondeos decimales al margen— con la calculada manualmente mediante la descomposición de autovalores y autovectores. Las columnas `PC1` y `PC2` nos dan las cargas factoriales, `h2` la comunalidad y `u2` la especificidad —nótese que cada fila suma 1—.

Estadísticos e-Books & Papers

Cuadro 12.6.: Estimación del EFA mediante `principal{psych}`

```
Principal Components Analysis
Call: principal(r = R, nfactors = 2, rotate = "none")
Standardized loadings (pattern matrix) based upon correlation matrix
     PC1   PC2   h2   u2 com
Gae 0.66  0.45 0.63 0.37 1.8
Eng 0.69  0.29 0.56 0.44 1.3
His 0.52  0.64 0.67 0.33 1.9
Ari 0.74 -0.41 0.72 0.28 1.6
Alg 0.74 -0.38 0.69 0.31 1.5
Geo 0.68 -0.36 0.59 0.41 1.5

                       PC1  PC2
SS loadings           2.73 1.13
Proportion Var        0.46 0.19
Cumulative Var        0.46 0.64
Proportion Explained  0.71 0.29
Cumulative Proportion 0.71 1.00

Mean item complexity =  1.6
Test of the hypothesis that 2 components are sufficient.

The root mean square of the residuals (RMSR) is  0.11

Fit based upon off diagonal values = 0.9
```

`SS loadings` es la suma de los cuadrados de las cargas factoriales, es decir, la varianza explicada por cada factor que coincide con su autovalor, aparece luego el porcentaje que representa del total —`Proportion Var`— y el acumulado de los factores –`Cumulative var`–. Si en lugar de sobre la varianza total, hacemos los cálculos sobre la explicada por la solución factorial de dos factores para evaluar el peso relativo de cada uno, tenemos también los porcentajes por factor —`Proportion Explained`— y el acumulado —`Cumulative proportion`—. El resto de elementos de la salida se retomarán cuando se aborden cuestiones como el ajuste o la elección del número de factores para extraer.

12.3.3. Método de los ejes principales

El método de las *componentes principales iteradas*, al que también se denomina método de *ejes principales*, tiene una gran similitud con la extracción de factores por componentes principales. Este es un método iterativo, como lo son en general los métodos de obtención de factores, exceptuando el de componentes principales. Por ello, en algunas ocasiones no permite llegar a unas estimaciones adecuadas debido a los problemas de convergencia que pueden plantear los métodos iterativos.

Vamos a exponer los distintos pasos de este método de extracción de factores:

1. Se calcula la matriz de correlaciones muestral $\mathbf{R}$ dada en (12.9).

2. Se realiza una estimación inicial de las comunalidades de cada variable. Para ello, se calcula la regresión de cada variable sobre el resto de varia-

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

bles originales, estimándose la comunalidad de una variable mediante el coeficiente de determinación obtenido.

3. Se sustituye en la matriz $\mathbf{R}$ cada 1 de la diagonal principal por la estimación de la comunalidad correspondiente a cada variable. A la matriz $\mathbf{R}$ modificada de esta forma, a la que designaremos por $\mathbf{R}^*$, se la denomina matriz de correlación reducida:

$$R^* = \begin{bmatrix} \hat{h}_1^2 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & \hat{h}_2^2 & \cdots & \rho_{20} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & \hat{h}_p^2 \end{bmatrix} \tag{12.22}$$

Como la estimación de cada comunalidad es igual a 1 menos la estimación de la especificidad correspondiente, la matriz de correlación reducida se puede expresar de esta forma alternativa:

$$\mathbf{R}^* = \mathbf{R} - \mathbf{\Psi}$$

4. Se calculan los autovalores y autovectores de $\mathbf{R}^*$ y se obtienen las cargas factoriales estimadas como vimos en la subsección anterior.

5. Se determina el número de factores a retener m, con cualquiera de los procedimientos que veremos posteriormente.

6. Se calcula la comunalidad de cada variable con los m factores retenidos:

$$h_j^2 = \lambda_{j1}^2 + \lambda_{j2}^2 + \cdots \lambda_{jm}^2$$

y se calculan las varianzas residuales:

$$\varepsilon_j^2 = 1 - h_j^2$$

si alguna de estas varianzas fuera negativa, lo que puede ocurrir, el resultado de la estimación no sería admisible, por lo que hay que ser muy cauteloso en la interpretación de las salidas.

7. En el caso de que todos los valores sean positivos, se vuelve al paso 3 utilizando las nuevas comunalidades estimadas, y se repiten los pasos 4, 5 y 6. Un ciclo formado por todos estos pasos constituye una iteración.

El procedimiento se detiene cuando la diferencia entre la comunalidad estimada para cada variable entre dos iteraciones sucesivas sea menor de una cantidad prefijada. Se dice entonces que se ha alcanzado la convergencia.

La estimación mediante ejes principales (*PA-principal axis*) se realiza con la función `fa{psych}` y su sintaxis se explica en sí misma, solo hay que indicar el

Estadísticos e-Books & Papers

Cuadro 12.7.: Estimación del EFA con extracción por ejes principales

```
Factor Analysis using method =  pa
Call: fa(r = R, nfactors = 2, n.obs = 220, rotate = "none", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
     PA1   PA2   h2   u2 com
Gae 0.59  0.38 0.49 0.51 1.7
Eng 0.59  0.24 0.41 0.59 1.3
His 0.43  0.41 0.36 0.64 2.0
Ari 0.71 -0.33 0.62 0.38 1.4
Alg 0.70 -0.28 0.57 0.43 1.3
Geo 0.58 -0.18 0.38 0.62 1.2

                       PA1  PA2
SS loadings           2.22 0.59
Proportion Var        0.37 0.10
Cumulative Var        0.37 0.47
Proportion Explained  0.79 0.21
Cumulative Proportion 0.79 1.00

Mean item complexity =  1.5
Test of the hypothesis that 2 factors are sufficient.

The degrees of freedom for the null model are  15  and the objective function was
1.44 with Chi Square of  310.84
The degrees of freedom for the model are 4  and the objective function was  0.01

The root mean square of the residuals (RMSR) is  0.01
The df corrected root mean square of the residuals is  0.03

The harmonic number of observations is  220 with the empirical chi square  1.32
with prob <  0.86
The total number of observations was  220  with Likelihood Chi Square =  2.35
with prob <  0.67

Tucker Lewis Index of factoring reliability =  1.021
RMSEA index =  0  and the 90 % confidence intervals are  NA 0.079
BIC =  -19.22
Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                                   PA1  PA2
Correlation of scores with factors               0.90 0.73
Multiple R square of scores with factors         0.82 0.53
Minimum correlation of possible factor scores  0.64 0.0
```

número de factores que queremos extraer —vamos a seguir con dos, aunque no hemos visto todavía los criterios para determinar este número— `nfactors=2` y que el método de extracción ha de ser ejes principales (`fm="pa"`). El resultado se muestra en el cuadro 12.7 y su estructura e interpretación es la misma que dimos en el cuadro 12.6 para el método de componentes principales.

```
fit.pa<-fa(R,nfactors=2,fm="pa",rotate="none",n.obs=220)
print(fit.pa)
```

12.3.4. Método de máxima verosimilitud

Es otro proceso iterativo muy común en las herramientas que analizaremos en los próximos capítulos, como el análisis factorial confirmatorio o los modelos de ecuaciones estructurales. En estos temas daremos una visión más profunda del procedimiento, pero es necesario dejar aquí la intuición del mismo. Como hemos visto con anterioridad, una solución factorial determinada genera una

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

matriz de correlaciones reproducida que puede compararse con la matriz de correlaciones original. La matriz reproducida depende de los parámetros por estimar, como hemos visto también en la expresión (12.20). Pues bien, el procedimiento de máxima verosimilitud se basa en construir una función —la de máxima verosimilitud— que básicamente es la diferencia entre la matriz muestral y la reproducida y estimar los parámetros que minimicen esa diferencia en un proceso iterativo. Es un planteamiento muy paramétrico que exige normalidad multivariante.

La función de máxima verosimilitud que es necesario minimizar es la siguiente (se ofrece para valores no tipificados y, por ello, se plantea en términos de covarianzas y no de correlaciones):

$$F_{ML}\left(\mathbf{S};\Sigma^{*}\right)=tr\left(\mathbf{S}\Sigma^{*-1}\right)+\left[\log\left|\Sigma^{*}\right|-\log\left|\mathbf{S}\right|\right]-p \tag{12.23}$$

donde $\boldsymbol{S}$ es la matriz de covarianzas muestral, $\boldsymbol{\Sigma^*}$ es la matriz de covarianzas reproducida para un número de factores determinado y p que es el número de variables observadas. Con el símbolo $|\ |$ denotamos al determinante de la matriz de referencia. Como señala Long (1983b), cuanto más se aproximen las matrices $\mathbf{S}$ y $\boldsymbol{\Sigma^*}$, más se aproximará el producto $\mathbf{S}\boldsymbol{\Sigma^{*-1}}$ a la matriz identidad. Como la matriz identidad tiene rango $p \times p$, entonces, dado que la traza de esa matriz identidad es la suma de los p unos de la diagonal (o sea, p), el primer término de (12.23) se aproximará a p cuando las matrices estén próximas, compensándose con el término $-p$ del final de la expresión (12.23). Por otra parte, la diferencia de los logaritmos de los determinantes de $\mathbf{S}$ y Σ^* tenderá a 0, dado que, cuando las matrices estén próximas, también lo estarán sus determinantes. De esta forma, cuando las matrices sean iguales, la función de ajuste será cero.

Una vez más, la obtención de la solución es inmediata con la función `fa{psych}` y su sintaxis se explica, una vez más, en sí misma, solo hay que indicar el número de factores que queremos extraer —vamos a seguir con dos, aunque no hemos visto todavía los criterios para determinar este número— `nfactors=2` y que el método de extracción ha de ser máxima verosimilitud (`fm="ml"`).

```
fit.ml<-fa(R,nfactors=2,fm="ml",rotate="none",n.obs=220)
print(fit.ml)
```

12.3.5. Otros métodos de extracción

Otros métodos de extracción que describiremos de manera mucho más sucinta son los siguientes.

A. Extracción alpha

Pretende maximizar el coeficiente α de Cronbach (Cronbach, 1951) como estimador de la fiabilidad del factor. En el tema 14 profundizaremos en este coeficiente pero, de momento, podemos quedarnos con la idea de que es un

Estadísticos e-Books & Papers

Cuadro 12.8.: Estimación del EFA con extracción por máxima verosimilitud

```
Factor Analysis using method =  ml
Call: fa(r = R, nfactors = 2, n.obs = 220, rotate = "none", fm = "ml")
Standardized loadings (pattern matrix) based upon correlation matrix
     ML1   ML2   h2   u2 com
Gae 0.55  0.43 0.49 0.51 1.9
Eng 0.57  0.29 0.41 0.59 1.5
His 0.39  0.45 0.36 0.64 2.0
Ari 0.74 -0.27 0.62 0.38 1.3
Alg 0.72 -0.21 0.57 0.43 1.2
Geo 0.60 -0.13 0.37 0.63 1.1

                       ML1  ML2
SS loadings           2.21 0.61
Proportion Var        0.37 0.10
Cumulative Var        0.37 0.47
Proportion Explained  0.78 0.22
Cumulative Proportion 0.78 1.00

Mean item complexity =  1.5
Test of the hypothesis that 2 factors are sufficient.

The degrees of freedom for the null model are  15  and the objective function
was  1.44 with Chi Square of  310.84
The degrees of freedom for the model are 4  and the objective function was
0.01

The root mean square of the residuals (RMSR) is  0.01
The df corrected root mean square of the residuals is  0.03

The harmonic number of observations is  220 with the empirical chi square
1.34  with prob <  0.86
The total number of observations was  220  with Likelihood Chi Square =  2.33
with prob <  0.67

Tucker Lewis Index of factoring reliability =  1.021
RMSEA index =  0  and the 90 % confidence intervals are  NA 0.079
BIC =  -19.24
Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                                   ML1  ML2
Correlation of scores with factors               0.91 0.73
Multiple R square of scores with factors         0.82 0.53
Minimum correlation of possible factor scores  0.64 0.07
```

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

promedio de las correlaciones entre los indicadores. La diferencia fundamental con otros procedimientos (Osborne y Banjanovic, 2016) es que, en lugar de pretender generalizar la muestra a una población de individuos, pretende generalizar la muestra a una población de indicadores. Tiene la limitación de que cuando se produce la rotación, esta propiedad de generalización se pierde (Nunnally y Bernstein, 1994) y, como veremos al abordar la rotación, los resultados sin rotar son por lo general más difíciles de interpretar que los resultados rotados.

B. Mínimos cuadrados no ponderados

Utiliza una variación de la extracción por máxima verosimilitud que no hace requerir normalidad multivariante a los datos. Pretende minimizar el error operacionalizado como una suma de los cuadrados de los residuos entre la matriz reproducida y la muestral.

C. Mínimos cuadrados generalizados

Este método aplica el mismo criterio que el método anterior, pero ponderando las correlaciones con la inversa de la especificidad de las variables. De esta forma, las correlaciones entre variables con elevada especificidad tendrán menos peso en los resultados finales que las correlaciones entre variables con una baja especificidad.

12.3.6. ¿Qué método elegir?

Esta es, obviamente, la gran pregunta. Sharma (1996) afirma que, afortunadamente, las diferencias entre los resultados son mínimas, lo que hace que, en el fondo, no importe qué método de extracción utilizar. Sí que insiste, sin embargo, en que el método de componentes principales no descompone el modelo en varianza común y específica y que toda la varianza es común si se utilizan los p componentes. Sin embargo esto no ocurre con el planteamiento de ejes principales, lo que hace que sea preferido por muchos investigadores pues une esta visión más amplia a una facilidad elevada de cálculo computacional similar a los componentes principales.

Osborne y Banjanovic (2016) coinciden con el planteamiento anterior cuando la estructura factorial es clara y se cumple la normalidad multivariante. Si esta no se cumple, los ejes principales o los mínimos cuadrados no ponderados pueden ser una buena opción. En síntesis, el método no importa si se cumplen los supuestos pero sí en el caso en que no lo hagan. Hay bastante consenso en la literatura, señalan, de que la opción de máxima verosimilitud es la mejor opción cuando no se cumplen los supuestos y mínimos cuadrados no ponderados y ejes principales cuando sí (Fabrigar *et al.*, 1999; Nunnally y Bernstein, 1994).

En nuestro caso, si observamos la figura 12.2 donde hemos representado las cargas de los tres métodos de extracción, vemos que los resultados producidos son prácticamente idénticos.

Estadísticos e-Books & Papers

```
plot(fit.pca,labels=row.names(R),cex=.7, ylim=c(-.8,.8))
plot(fit.pa,labels=row.names(R),cex=.7, ylim=c(-.8,.8))
plot(fit.ml,labels=row.names(R),cex=.7, ylim=c(-.8,.8))
```

12.4. Determinación del número de factores que hay que retener

Sobre este tema ya hicimos una primera aproximación al tratar el análisis de componentes principales. Durante toda la exposición anterior hemos supuesto que retener dos factores era lo adecuado, pero nada hemos analizado para ver si esta decisión era o no correcta. Las reglas vistas en el capítulo anterior son de aplicación aquí y también sus ventajas y limitaciones: autovalor superior a la unidad, gráfico de sedimentación y método paralelo. Refresquemos brevemente estos criterios aplicados a nuestro caso. Remitimos al lector al capítulo anterior para un mayor detalle.

Kaiser (1960, 1970) propuso el criterio de solo retener aquellos factores cuyo **autovalor sea superior a la unidad**. La lógica de este criterio reside en el hecho de que el autovalor, recordemos de lo expuesto, es la suma de los cuadrados de las cargas asociadas al factor. Si este valor es alto, es que esas cargas lo eran y explicaban bien sus indicadores. Otra interpretación es que el autovalor representa la varianza. Como en una solución estandarizada la varianza con que una variable contribuye es 1, no tendría sentido retener factores que explican menos que una sola variable si estamos ante un procedimiento que pretende reducir datos. Este es un criterio que funciona bien con menos de 40 variables y muestras pequeñas, pero tiende a recoger demasiados factores con cifras superiores.

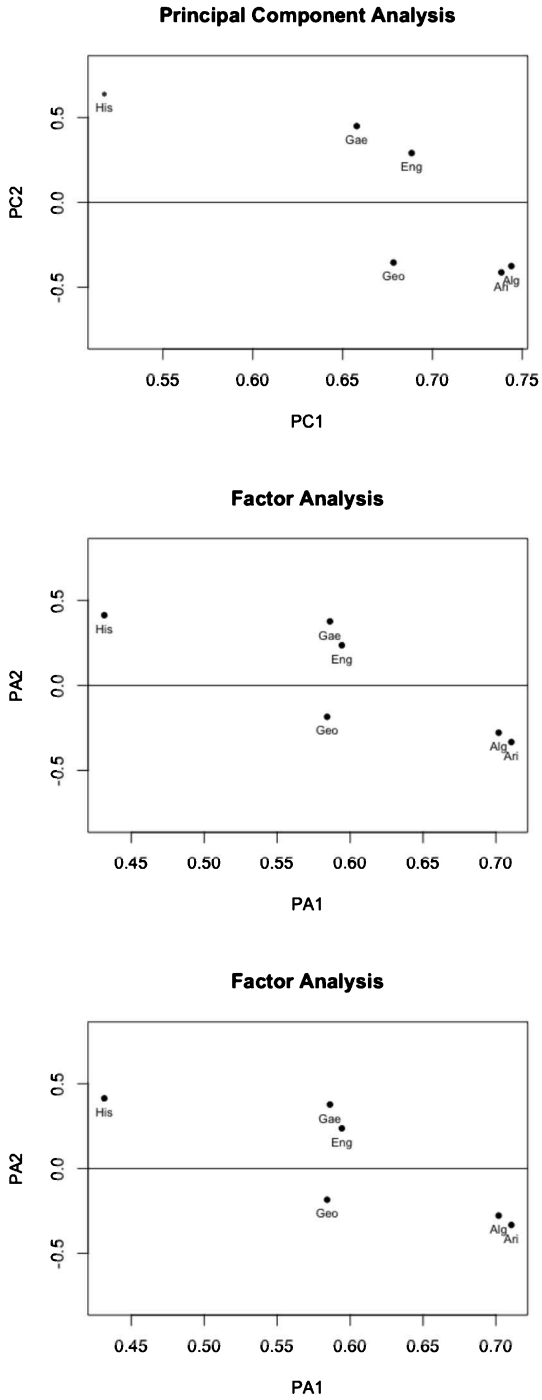
Cattell (1966) propuso el **gráfico de sedimentación** como elemento de decisión. Si se representan los autovalores en abscisas y el número de factores en ordenadas, los valores irán descendiendo —normalmente— hasta ser pequeños en los últimos factores. El punto donde el gráfico cambia de pendiente sería el que marcaría el número de factores que retener en la medida en que la contribución del resto es mayor en términos de complejidad de interpretación que no de varianza explicada. Es, como vimos, un criterio que necesita de la subjetividad de la interpretación del investigador y que, muchas veces, se encuentra con que los cambios de pendiente son tan progresivos que cuesta encontrar ese quiebro en el gráfico. Gorsuch (1983) encontró que los resultados del gráfico de sedimentación son más fiables con tamaños muestrales y comunalidades grandes.

Horn (1965) propuso el **análisis paralelo** como una alternativa a considerar solo los autovalores superiores a la unidad y como una forma de objetivar el gráfico de sedimentación. Primero se generan conjuntos de datos aleatorios con el mismo número de casos y variables que el original. Se realizan análisis PCA

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 12.2.: Representación gráfica de las tres estimaciones



Estadísticos e-Books & Papers

repetidos sobre cada uno de esos conjuntos aleatorios de datos anotándose los autovalores de cada análisis. Se calcula la media de esos autovalores en los conjuntos aleatorios de datos para cada factor y se comparan con los del conjunto real de datos. El criterio es retener solo aquellos autovalores cuyo promedio supere el aleatorio. Una ventaja de este criterio es que nos hace conscientes de que incluso los datos aleatorios pueden generar correlaciones espúreas que llevan a autovalores superiores a la unidad.

Finalmente Bartlett (1951) propuso un test estadístico que presentamos en el capítulo anterior. Recordemos que tanto Sharma (1996) como Stevens (2009) consideran que dicho test, que está basado en el test de Bartlett (1951) para analizar si la matriz de correlaciones es la identidad, adolece de las mismas limitaciones —básicamente ser muy sensible al tamaño muestral— y no recomiendan su uso pues suele recomendar la retención de demasiadas componentes principales. En cualquier caso lo mostraremos para ilustrar su obtención.

La siguiente sintaxis permite obtener todos los criterios expuestos de manera simultánea. La figura 12.3 confirma que, para el caso 12.1, retener dos factores es la solución más adecuada.

```
# autovalor>1, sedimentacion, paralelo
fa.parallel(R,fm="pa",n.obs=220,ylabel="Eigenvalues")
fa.parallel(R,fm="ml",n.obs=220,ylabel="Eigenvalues")
# test de Barlett
nBartlett(R, N=220, alpha=0.01, cor=TRUE, details=TRUE)
```

12.5. Rotación de la solución factorial

Cuando se aplica el análisis factorial se trata de que los factores comunes tengan una interpretación clara, porque de esa forma se analizan las interrelaciones existentes entre las variables originales. Sin embargo, en muy pocas ocasiones resulta fácil encontrar una interpretación adecuada de los factores iniciales, con independencia del método que se haya utilizado para su extracción.

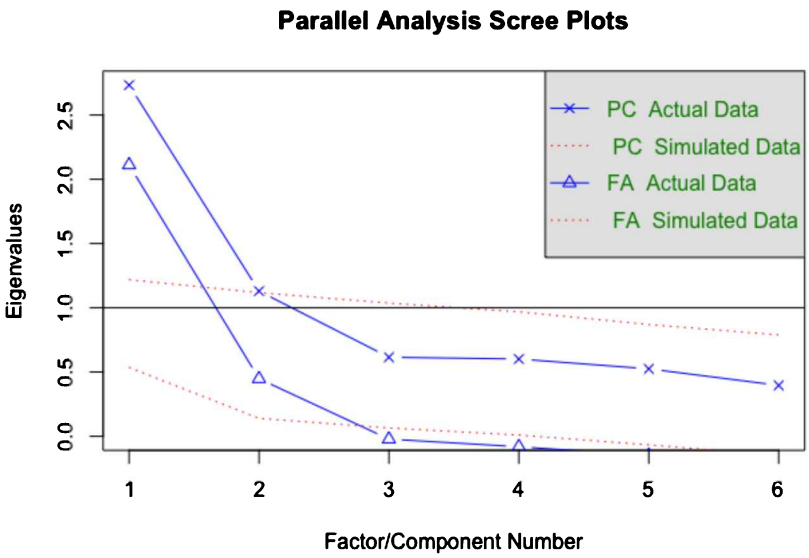
Precisamente los procedimientos de rotación de factores se han ideado para obtener, a partir de la solución inicial, unos factores que sean fácilmente interpretables. Veamos cuál es la esencia de estos procedimientos.

En la solución inicial cada uno de los factores comunes están correlacionados en mayor o menor medida con cada una de las variables originales. Pues bien, con los factores rotados se trata de que cada una de las variables originales tenga una correlación lo más próxima a 1 que sea posible con uno de los factores y correlaciones próximas a 0 con el resto de los factores. De esta forma, y dado que hay más variables que factores comunes, cada factor tendrá una correlación alta con un grupo de variables y baja con el resto de variables. Examinando las características de las variables de un grupo asociado a un determinado factor se pueden encontrar rasgos comunes que permitan identificar el factor y darle una

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 12.3.: Criterios para determinar el número de factores




```
> nBartlett(R, N=220, alpha=0.01, cor=TRUE, details=TRUE)
bartlett anderson   lawley
       2        2        2
```


Estadísticos e-Books & Papers

denominación que responda a esos rasgos comunes. Si se consigue identificar claramente estos rasgos, se habrá dado un paso importante, ya que con los factores comunes no solo se reducirá la dimensionalidad del problema, sino que también se conseguirá desvelar la naturaleza de las interrelaciones existentes entre las variables originales.

Existen dos formas básicas de realizar la rotación de factores: rotación ortogonal y rotación oblicua. Estas dos formas alternativas de rotación serán examinadas a continuación.

12.5.1. Rotación ortogonal

En la rotación ortogonal, los ejes se rotan de forma que quede preservada la incorrelación entre los factores. Dicho de otra forma, los nuevos ejes, o ejes rotados, son perpendiculares de igual forma que lo son los factores sin rotar. Por esta restricción, la rotación ortogonal se denomina también rotación rígida.

En la figura 12.4 pueden verse los resultados del caso 12.1 con los factores extraídos mediante ejes principales sin rotación y con una rotación ortogonal. Es bastante evidente que el nivel de correlación entre las variables y los ejes se ha incrementado pero de manera diferenciada para los dos grupos de variables. Ahora el nivel de correlación —no olvidemos que la carga indica la correlación entre la variable y el factor— entre las calificaciones en geometría, álgebra y aritmética con el nuevo factor 1 es muy elevado y es prácticamente nula con el factor 2. Por lo tanto sus cargas serán altas sobre el eje 1 y muy pequeñas sobre el 2. Este resultado es inverso para el segundo grupo de variables.

Si estimamos ahora el EFA mediante ejes principales pidiendo una rotación ortogonal que en la sintaxis aparece como `varimax` y cuya lógica explicaremos inmediatamente, vemos en el cuadro 12.9 que este comportamiento de la estimación de las cargas se confirma. La interpretación se facilita mucho ahora. Nos encontramos con un factor, la dimensión 2, que podríamos denominar factor de *habilidad matemática,* mientras que la dimensión 1, que agrupa a las calificaciones en gaélico, inglés e historia, podría denominarse factor de *habilidad verbal.* ¿Qué ha ocurrido con el factor 1 que en la solución sin rotar denominábamos *inteligencia general*? Pues que estaba ocultando estos dos factores que han emergido en la rotación. Si comparamos el cuadro 12.9 con el cuadro 12.7 que recoge la solución sin rotar, veremos que las comunalidades no cambian en la rotación ortogonal.

```
fit.pa<-fa(R,nfactors=2,fm="pa",rotate="varimax",n.obs=220)
```

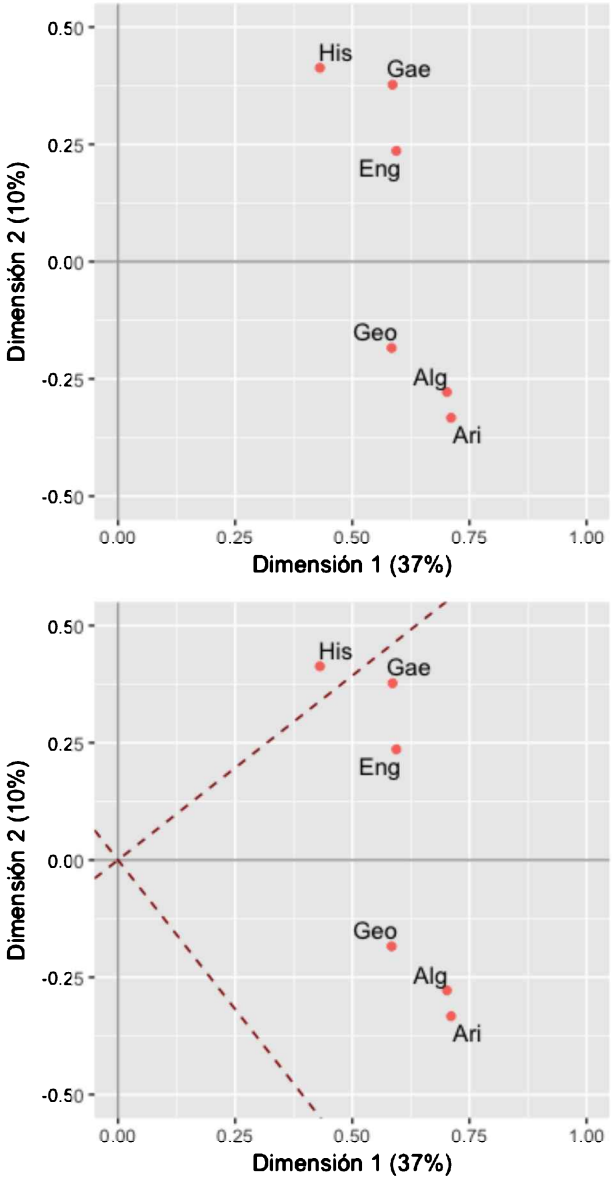
A. Varimax

La rotación ortogonal más conocida es la **rotación varimax**. Esta rotación determina el ángulo de giro de tal forma que se maximiza la suma de las varianzas de las cargas factoriales al cuadrado dentro de cada factor. Esto

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 12.4.: Solución rotada y sin rotar del caso 12.1 (ejes principales)



Estadísticos e-Books & Papers

Cuadro 12.9.: Cargas factoriales estimadas con una rotación ortogonal y extracción con ejes principales

```
Factor Analysis using method =  pa
Call: fa(r = R, nfactors = 2, n.obs = 220, rotate = "varimax", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
     PA1  PA2   h2   u2 com
Gae 0.23 0.66 0.49 0.51 1.2
Eng 0.32 0.55 0.41 0.59 1.6
His 0.08 0.59 0.36 0.64 1.0
Ari 0.76 0.18 0.62 0.38 1.1
Alg 0.72 0.22 0.57 0.43 1.2
Geo 0.57 0.22 0.38 0.62 1.3
```

se consigue maximizando la varianza del cuadrado de las cargas sujeta a la restricción de que la comunalidad de cada variable no cambia. Para un factor j determinado:

$$V_j = \frac{\sum_{h=1}^{p}\left(\lambda_{hj}^2 - \lambda_{\cdot j}^2\right)^2}{p} = \frac{p\sum_{h=1}^{p}\lambda_{hj}^4 - \left(\sum_{h=1}^{p}\lambda_{hj}^2\right)^2}{p^2} \tag{12.24}$$

donde V_j es la varianza de las comunalidades del factor j y $\lambda_{\cdot j}^2$ es el promedio de las cargas al cuadrado del factor j.

Por lo tanto, la suma de varianzas de las cargas factoriales al cuadrado dentro de cada factor será igual a (siendo m el número de factores seleccionados):

$$\begin{aligned} V &= \sum_{j=1}^{m} V_j = \sum_{j=1}^{m}\left(\frac{p\sum_{h=1}^{p}\lambda_{hj}^4 - \left(\sum_{h=1}^{p}\lambda_{hj}^2\right)^2}{p^2}\right) \qquad (12.25)\\ &= \frac{\sum_{j=1}^{m}\sum_{h=1}^{p}\lambda_{hj}^4}{p} - \frac{\sum_{j=1}^{m}\left(\sum_{h=1}^{p}\lambda_{hj}^2\right)^2}{p^2} \end{aligned}$$

Dado que el número de variables siempre es el mismo, la maximización de V es lo mismo que maximizar:

$$pV = \sum_{j=1}^{m}\sum_{h=1}^{p}\lambda_{hj}^4 - \frac{\sum_{j=1}^{m}\left(\sum_{h=1}^{p}\lambda_{hj}^2\right)^2}{p^2} \tag{12.26}$$

Una de las propiedades del método varimax es que, como hemos visto en la ilustración anterior para el caso 12.1, después de aplicado, queda inalterada tanto la varianza total explicada por los factores como la comunalidad de cada una de las variables. El método varimax, cuyo origen está relacionado con que la VARIanza se MAXimiza, facilita la interpretación de los factores.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

B. Quartimax

La rotación **quartimax** hace con las variables lo que varimax hace con los factores. Las simplifica incrementando la dispersión de las cargas dentro de las variables entre los factores. Mientras que varimax opera en las columnas de la matriz de cargas, quartimax opera en las filas. No es tan utilizada porque un investigador quiere normalmente factores más sencillos, no variables más sencillas.

La maximización de la varianza de las cargas en los factores sujeta a la restricción de que la comunalidad de cada variable permanece constante se operativiza del siguiente modo. Supongamos que para una variable h dada, definimos:

$$Q_h = \frac{\sum_{j=1}^{m} \left(\lambda_{hj}^2 - \lambda_{i\cdot}^2\right)^2}{m} \tag{12.27}$$

donde Q_h es la varianza de las comunalidades de la variable h; λ_{hj}^2 es el cuadrado de la carga de la variable h-ésima en el factor j; $\lambda_{h\cdot}^2$ es la media del cuadrado de la carga de la variable h, y m, el número de factores. Podemos poner:

$$Q_h = \frac{m \sum_{j=1}^{m} \lambda_{hj}^4 - \left(\sum_{j=1}^{m} \lambda_{hj}^2\right)^2}{m^2} \tag{12.28}$$

Y la varianza total para el conjunto de las variables quedaría:

$$Q = \sum_{h=1}^{p} Q_h = \sum_{h=1}^{p} \left[\frac{m \sum_{j=1}^{m} \lambda_{hj}^4 - \left(\sum_{j=1}^{m} \lambda_{hj}^2\right)^2}{m^2}\right] \tag{12.29}$$

En una rotación quartimax, la matriz de giro es aquella que maximiza la expresión anterior sujeta a la condición de que la comunalidad de cada variable permanezca constante. Una vez obtenida la solución factorial inicial, el número de factores m permanece constante. El segundo término de la expresión $\sum_{j=1}^{m} \lambda_{hj}^2$ es la comunalidad de la variable y, por lo tanto, también constante. Por lo que la maximización se ve reducida a:

$$Q = \sum_{h=1}^{p} \sum_{j=1}^{m} \lambda_{hj}^4 \tag{12.30}$$

C. Otras rotaciones ortogonales

Un planteamiento híbrido entre los dos anteriores sería intentar simplificar simultáneamente variables y factores, es decir, buscar una rotación que maximizara la varianza fila y columna, es decir:

$$Z = \alpha Q + \beta p V \tag{12.31}$$

Estadísticos e-Books & Papers

Cuadro 12.10.: Matriz de giro

```
           [,1]      [,2]
[1,]  0.7861450 0.6180421
[2,] -0.6180421 0.7861450
```

donde Q viene dada por (12.30) y pV por (12.26). Consideremos la siguiente ecuación:

$$\sum_{h=1}^{p}\sum_{j=1}^{m}\lambda_{hj}^{4}-\gamma\frac{\sum_{j=1}^{m}\left(\sum_{h=1}^{p}\lambda_{hj}^{2}\right)^{2}}{p^{2}} \tag{12.32}$$

donde $\gamma=\beta/(\alpha+\beta)$. Distintos valores de γ resultarán en distintos tipos de rotación. De hecho si $\gamma=0$ (por ejemplo porque $\alpha=1$ y $\beta=0$), entonces tenemos una rotación quartimax. Si $\gamma=1$ (por ejemplo porque $\alpha=0$ y $\beta=1$), tenemos una rotación varimax. Si $\gamma=m/2$, tenemos lo que se denomina una rotación **equimax** y, si $\gamma=0{,}5$ ($\alpha=1$; $\beta=1$), la conocida como **biquartimax.**

Sin embargo Mulaik (1972) demuestra que, a no ser que el investigador acierte en la determinación del número de factores, las rotaciones equimax y biquartimax tienen un comportamiento bastante errático.

Al final, cualquier rotación ortogonal se obtendrá multiplicando la matriz de cargas factoriales por una matriz de transformación que recoge el ángulo que los procedimientos anteriores han determinado como óptimo. Si denominamos $\mathbf{T}$ a esta matriz de giro, su relación con el ángulo de giro θ será la siguiente cuando el giro sea en el sentido de las agujas del reloj:

$$\mathbf{T}=\begin{bmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{bmatrix} \tag{12.33}$$

o la siguiente cuando el giro sea en el sentido contrario:

$$\mathbf{T}=\begin{bmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{bmatrix} \tag{12.34}$$

Los resultados de la estimación calculan siempre esta matriz de giro, basta solicitar en la sintaxis su impresión dentro del objeto creado con la solución, en nuestro ejemplo `fit.pa`. El cuadro 12.10 recoge la matriz y puede comprobarse fácilmente como aplicando esta matriz de giro a los coeficientes de la solución no rotada del cuadro 12.7, obtenemos los de la solución rotada del cuadro 12.9:

```
fit.pa$rot.mat
```

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

$$\begin{bmatrix} 0{,}59 & 0{,}38 \\ 0{,}59 & 0{,}24 \\ 0{,}43 & 0{,}41 \\ 0{,}71 & -0{,}33 \\ 0{,}70 & -0{,}28 \\ 0{,}58 & -0{,}18 \end{bmatrix} \begin{bmatrix} 0{,}7861450 & 0{,}6180241 \\ -0{,}6180421 & 0{,}7861450 \end{bmatrix} = \begin{bmatrix} 0{,}23 & 0{,}66 \\ 0{,}32 & 0{,}55 \\ 0{,}08 & 0{,}59 \\ 0{,}76 & 0{,}18 \\ 0{,}72 & 0{,}22 \\ 0{,}57 & 0{,}22 \end{bmatrix}$$

12.5.2. Rotación oblicua

Con la denominación de rotación oblicua se indica que los ejes no son ortogonales, es decir, que no son perpendiculares.

Cuando se realiza una rotación oblicua, los factores ya no estarán incorrelacionados, con lo que se pierde una propiedad que en principio es deseable que cumplan los factores. Sin embargo, en ocasiones puede compensarse esta pérdida, si, a cambio, se consigue una asociación más nítida de cada una de las variables con el factor correspondiente.

En la figura 12.5 se ilustra la rotación oblicua. Como puede verse, con esta rotación se consigue que las cargas factoriales del grupo de variables de geometría, álgebra y aritmética sean todavía más pequeñas en el factor 2 que cuando se aplicaba una rotación ortogonal.

El método de rotación oblicua más conocido es el denominado **oblimin**. Existen algoritmos que permiten controlar el grado de no ortogonalidad. Si el lector necesita profundizar en estos algoritmos o en el concepto de rotación oblicua, puede consultarse a Harman (1976), Clarkson y Jennrich (1988) o Basilevsky (1994).

Conviene advertir que, tanto en la rotación ortogonal como en la rotación oblicua, la comunalidad de cada variable no se ve modificada. Pero no todas las propiedades del modelo permanecen inalteradas.

12.5.3. Cambios que provoca la rotación oblicua en las propiedades del modelo

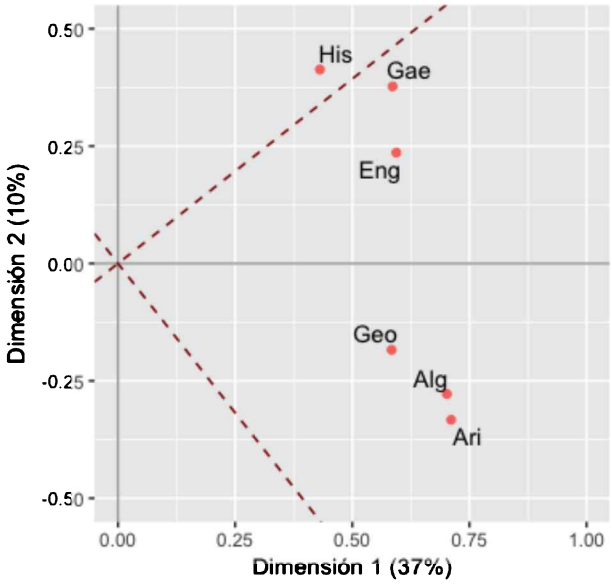
El cambio más evidente de cualquier rotación y al que ya hemos hecho alusión anteriormente afecta a la **matriz factorial** (llamada habitualmente *pattern matrix*) que contiene las cargas factoriales y que son distintas para la solución original y para la rotada.

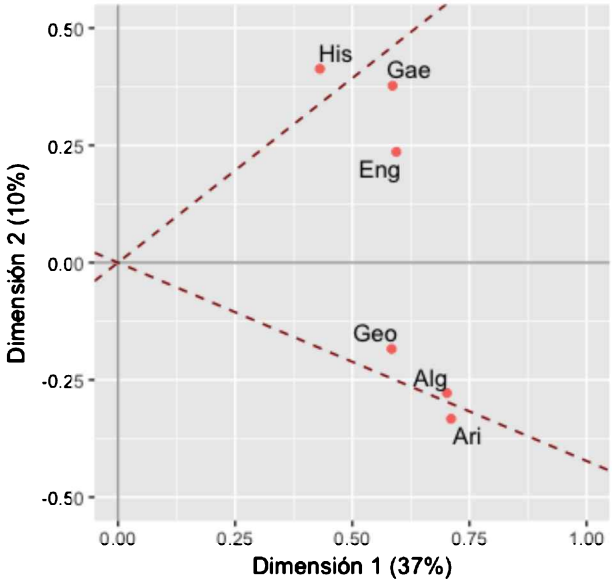
Sin embargo, hasta este momento, no habíamos hecho alusión alguna a la denominada **matriz de estructura** porque en cualquier rotación ortogonal coincide con la matriz factorial rotada. Esto deja de ser cierto cuando la rotación es oblicua.

Retomemos el cuadro 12.5. La ecuación que recoge, por ejemplo, la relación de la variable x_1 con los factores ξ_1 y ξ_2 es la siguiente:

Estadísticos e-Books & Papers

Figura 12.5.: Rotación ortogonal (varimax) y oblicua (oblimin) del caso 12.1 (ejes principales)





 © Ediciones Paraninfo

Estadísticos e-Books & Papers

$$x_1 = -0{,}66\xi_1 + 0{,}45\xi_2 \tag{12.35}$$

donde –0,66 y 0,45 son las cargas factoriales, de manera que la matriz factorial será para el conjunto de variables:

$$\begin{bmatrix} -0{,}66 & 0{,}45 \\ -0{,}69 & 0{,}29 \\ -0{,}52 & 0{,}64 \\ -0{,}74 & -0{,}41 \\ -0{,}74 & -0{,}38 \\ -0{,}68 & -0{,}36 \end{bmatrix}$$

El valor –0,66 es la correlación individual entre la variable x_1 y el factor ξ_1, mientras que 0,45 es la correlación individual entre la variable x_2 y el factor ξ_2. La ecuación (12.35) es una regresión en la que la nota es la variable dependiente, y ξ_1 y ξ_2, las independientes, siendo las cargas factoriales los coeficientes de regresión. Como en cualquier análisis de regresión, los coeficientes de regresión coinciden con las correlaciones individuales solo si las variables independientes están incorrelacionadas entre sí. Como en una rotación ortogonal este es el caso, la matriz que recoge las cargas factoriales —matriz factorial— coincide con la matriz que recoge las correlaciones individuales entre el factor y la variable a la que se denomina **matriz de estructura**. Algunos autores (Sharma, 1996) llaman a esas correlaciones **cargas estructurales**.

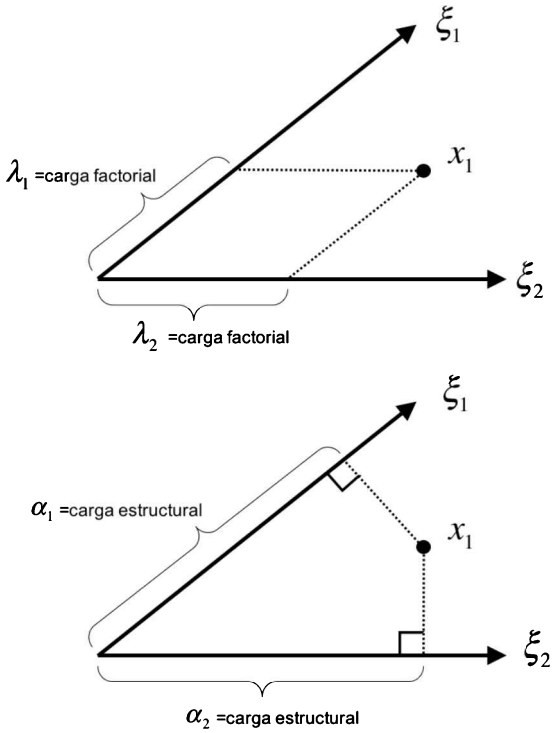
Veamos las diferencias bajo una perspectiva geométrica en la figura 12.6. Observamos que la *carga factorial* es la proyección de la variable manteniendo líneas paralelas a los ejes (panel a), mientras que las *cargas estructurales* son la proyección perpendicular a los ejes. La **interpretación del cuadrado de la carga estructural** es que este mide la varianza recogida por el efecto del factor y también por la interacción del factor con todos los demás factores. Esto hace que las cargas estructurales y, por ello, la matriz de estructura no sea muy útil para interpretar la estructura factorial y se recomiende siempre recurrir a la matriz factorial. El cuadro 12.11 ofrece las matrices de estructura para el caso de extracción por ejes principales con la rotación varimax (donde se puede comprobar que coincide con la matriz factorial) y con la rotación oblimin (donde esta coincidencia no se da).

12.6. Indicadores de bondad de la solución factorial

Los contrastes y estadísticos del análisis factorial, que se examinan en este epígrafe, se han agrupado en dos bloques, según que se apliquen previamente a la extración de los factores o que se apliquen *ex post*. Con la aplicación de los contrastes incluidos en el primer bloque se trata de analizar la pertinencia de aplicación del análisis factorial a un conjunto de variables observables. En

Estadísticos e-Books & Papers

Figura 12.6.: Interpretación geométrica de las cargas factoriales y las cargas de estructura



Fuente: Sharma (1996, p. 140).

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 12.11.: Matrices de estructura y factoriales para una rotación oblicua y una ortogonal en el caso 12.1

```
Estimación con rotación varimax-Pattern matrix
Standardized loadings (pattern matrix) based upon correlation matrix
     PA1  PA2   h2   u2 com
Gae 0.23 0.66 0.49 0.51 1.2
Eng 0.32 0.55 0.41 0.59 1.6
His 0.08 0.59 0.36 0.64 1.0
Ari 0.76 0.18 0.62 0.38 1.1
Alg 0.72 0.22 0.57 0.43 1.2
Geo 0.57 0.22 0.38 0.62 1.3

Estimación con rotación varimax-Structure matrix
Loadings:
    PA1   PA2
Gae 0.228 0.658
Eng 0.321 0.553
His       0.592
Ari 0.764 0.177
Alg 0.724 0.215
Geo 0.573 0.216

Estimación con rotación oblimin-Pattern matrix
Standardized loadings (pattern matrix) based upon correlation matrix
      PA1   PA2   h2   u2 com
Gae  0.03  0.68 0.49 0.51 1.0
Eng  0.17  0.53 0.41 0.59 1.2
His -0.11  0.65 0.36 0.64 1.1
Ari  0.80 -0.03 0.62 0.38 1.0
Alg  0.74  0.02 0.57 0.43 1.0
Geo  0.57  0.07 0.38 0.62 1.0

Estimación con rotación oblimin-Structure matrix
Loadings:
    PA1   PA2
Gae 0.391 0.696
Eng 0.454 0.622
His 0.234 0.591
Ari 0.784 0.389
Alg 0.755 0.414
Geo 0.610 0.372
```


Estadísticos e-Books & Papers

Cuadro 12.12.: Test de esfericidad de Barlett

```
$chisq
[1] 310.8409

$p.value
[1] 3.110695e-57

$df
[1] 15
```

el análisis *ex post* se pretende evaluar el modelo factorial estimado.

Dentro del primer bloque se va examinar el contraste de esfericidad de Bartlett y la medida de adecuación muestral de Kaiser, Meyer y Olkin.

En el segundo bloque se incluye el cálculo de las diferencias entre los coeficientes de correlación observados y los reproducidos, así como un contraste formal para medir la bondad del ajuste. Conviene señalar que las posibilidades de aplicación de este último son limitadas.

12.6.1. Contraste de esfericidad de Bartlett

La cuestión esencial, previa a la realización del análisis factorial, que se plantea es la siguiente: ¿están correlacionadas entre sí las variables originales? Si no lo estuvieran, no existirían factores comunes y, por lo tanto, no tendría sentido aplicar el análisis factorial. Para dar respuesta a esta cuestión se suele utilizar el contraste de esfericidad de Bartlett.

La hipótesis nula que hay que contrastar es que todos los coeficientes de correlación teóricos entre cada par de variables son nulos. El desarrollo de este contraste ya se ofreció en el capítulo 7 y su expresión venía dada por (7.21). Para el caso que nos ocupa vemos que, como ocurre casi siempre, la hipótesis nula de que la matriz de correlaciones es la identidad puede rechazarse para cualquier nivel de significatividad (cuadro 12.12).

```
cortest.bartlett(R,n=220)
```

12.6.2. Medidas de adecuación muestral de Kaiser-Meyer-Olkin

Los estadísticos Kaiser, Meyer y Olkin (Kaiser, 1970; Cerny y Kaiser, 1977) propusieron una medida de adecuación de la muestra al análisis factorial, que es conocida por las iniciales de sus nombres (KMO).

Antes de definir este estadístico conviene examinar la interpretación que puede darse a los coeficientes de correlación parcial entre cada par de variables originales en el contexto de la elaboración de un modelo factorial.

Un coeficiente de correlación parcial mide la correlación existente entre dos variables, una vez que se han descontado los efectos lineales de otras variables.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

En un modelo factorial se pueden interpretar esos efectos de otras variables como los correspondientes a los factores comunes. Por lo tanto, el coeficiente de correlación parcial entre dos variables sería equivalente, en este contexto, al coeficiente de correlación entre los factores únicos de dos variables. De acuerdo con el modelo de análisis factorial, los coeficientes de correlación teóricos calculados entre cada par de factores únicos son nulos por hipótesis. Si los coeficientes de correlación parcial constituyen una aproximación a dichos coeficientes teóricos, deben estar próximos a 0.

Una vez hechas estas precisiones, la medida KMO se define de la siguiente forma:

$$KMO = \frac{\sum_{h=1}^{p}\sum_{j=1}^{p}\rho_{jh}^2}{\sum_{h=1}^{p}\sum_{j=1}^{p}\rho_{jh}^2 + a_{jh}^2} \tag{12.36}$$

En la expresión anterior, ρ_{jh} son coeficientes de correlación observados entre variables originales, mientras que a_{jh} son coeficientes de correlación parcial entre variables originales.

En el caso de que exista adecuación de los datos a un modelo de análisis factorial, el término del denominador que recoge los coeficientes a_{jh} será pequeño y, consecuentemente, la medida KMO estará próxima a l. Según indican Kaiser (1974) y Kaiser y Rice (1974), el *benchmark* para considerar adecuado el indicador sería (no traducimos los adjetivos para mantener los matices del autor):

$\geq 0{,}9$	marvelous
$[0{,}8; 0{,}9)$	meritorious
$[0{,}7; 0{,}8)$	middling
$[0{,}6; 0{,}7)$	mediocre
$[0{,}5; 0{,}6)$	miserable
$< 0{,}5$	unacceptable

Basada en la KMO, se puede calcular también una **medida de adecuación muestral individual** para cada una de las variables. Esta medida, denominada MSA (Measure of Sampling Adequacy), se define de la siguiente forma:

$$MSA = \frac{\sum_{h\neq j}\rho_{jh}^2}{\sum_{h\neq j}\rho_{jh}^2 + \sum_{h\neq j}a_{jh}^2} \tag{12.37}$$

Un valor próximo a 1 de MSA_j indicará que la variable x_j es adecuada para su tratamiento en el análisis factorial con el resto de las variables. El cuadro 12.13 nos muestra la salida de los dos estadísticos KMO, el general y el

Estadísticos e-Books & Papers

Cuadro 12.13.: Medidas KMO de adecuación muestral general e individual

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = R)
Overall MSA =  0.77
MSA for each item =
 Gae  Eng  His  Ari  Alg  Geo
0.77 0.81 0.76 0.74 0.75 0.83
```

individual. En el primer caso estaría cerca de ser "*meritorious*" y en el segundo todos los valores individuales de las variables son también superiores a 0,70.

```
KMO(R)
```

12.6.3. Diferencias entre las correlaciones observadas y reproducidas

En el caso en que se utilizara una solución con tantos factores como variables —independientemente de que su utilidad sería nula— la matriz con las correlaciones que implicaría el modelo coincidiría con las muestrales. En la medida en que esto no sea así, y tomemos menos factores, las matrices diferirán. Cuanto más lo hagan, mayores serán los residuos (diferencia entre cada correlación) en la matriz residual.

La expresión (12.20) nos indicaba que la matriz de correlaciones reproducida se calculaba a partir de la matriz factorial como sigue:

$$\mathbf{R}^* = \mathbf{\Lambda}\mathbf{\Lambda}' \tag{12.38}$$

Por lo tanto, para el caso 12.1 con método de extracción por ejes principales y con rotación varimax, como vemos en el cuadro 12.11, la matriz de correlaciones reproducida sería:

$$R^* = \begin{bmatrix} 0{,}23 & 0{,}66 \\ 0{,}32 & 0{,}55 \\ 0{,}08 & 0{,}59 \\ 0{,}76 & 0{,}18 \\ 0{,}72 & 0{,}22 \\ 0{,}57 & 0{,}22 \end{bmatrix} \begin{bmatrix} 0{,}23 & 0{,}32 & 0{,}08 & 0{,}76 & 0{,}72 & 0{,}57 \\ 0{,}66 & 0{,}55 & 0{,}59 & 0{,}18 & 0{,}22 & 0{,}22 \end{bmatrix} =$$

$$= \begin{bmatrix} 0{,}4885 & 0{,}4366 & 0{,}4708 & 0{,}2936 & 0{,}3108 & 0{,}2763 \\ 0{,}4366 & 0{,}4049 & 0{,}3501 & 0{,}3422 & 0{,}3514 & 0{,}3034 \\ 0{,}4078 & 0{,}3501 & 0{,}3545 & 0{,}1670 & 0{,}1874 & 0{,}1754 \\ 0{,}2936 & 0{,}3422 & 0{,}1670 & 0{,}6100 & 0{,}5868 & 0{,}4728 \\ 0{,}3108 & 0{,}3514 & 0{,}1874 & 0{,}5868 & 0{,}5668 & 0{,}4588 \\ 0{,}2763 & 0{,}3034 & 0{,}1754 & 0{,}4728 & 0{,}4588 & 0{,}3733 \end{bmatrix}$$

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 12.14.: Matriz residual de la extracción con ejes principales y rotación varimax

```
> fit.pa$residual
            Gae          Eng          His          Ari          Alg          Geo
Gae  0.514451620  0.001577792  0.001397479 -0.002982844  0.022398434 -0.025040405
Eng  0.001577792  0.590916185 -0.003095620  0.010433601 -0.031388943  0.025312432
His  0.001397479 -0.003095620  0.643084487 -0.004922854  0.002105783  0.005056754
Ari -0.002982844  0.010433601 -0.004922854  0.384135768  0.003652524 -0.006531126
Alg  0.022398434 -0.031388943  0.002105783  0.003652524  0.430067350  0.002696246
Geo -0.025040405  0.025312432  0.005056754 -0.006531126  0.002696246  0.624690251
```

Y bastaría restar a la matriz original $\mathbf{R}$ del cuadro 12.1 la matriz $\mathbf{R}^*$ para tener la matriz residual. Afortunadamente esto no es necesario en la medida en que la función `fa{psych}` que hemos utilizado para la estimación del EFA calcula directamente esta información y la guarda en el objeto `fit.pa` junto con el resto de elementos de la estimación. Para recuperarlo solo es necesario solicitárselo:

```
fit.pa$residual
```

A partir de aquí pueden calcularse algunos indicadores de ajuste de la solución. Uno muy sencillo y que está implementado por defecto en `fa{psych}` es el RMSR (*root mean square of the residuals*) que se obtiene, simplemente, elevando al cuadrado los residuos para evitar la compensación de positivos y negativos, calculando la media de los mismos para obtener después la raíz cuadrada de esa media; `fa{psych}` lo aplica solo a los residuos fuera de la diagonal. Normalmente RMSR inferiores a 0,08 se consideran adecuados. Se extrae de la solución de manera muy sencilla con esta sintaxis, y su valor en esta aplicación es 0,01415.

```
fit.pa$rms
```

12.7. Puntuaciones factoriales

En algunas ocasiones el investigador puede estar interesado en conocer la puntuación de los individuos en los ejes factoriales. Por ejemplo, si quería realizar una regresión utilizando como variables independientes las variables x_1 a x_6 de nuestro ejemplo y se encontró con un problema de multicolinealidad, si efectúa un EFA con una rotación ortogonal y guarda las puntuaciones factoriales, éstas estarán incorrelacionadas y podrá utilizarlas como variables independientes en su regresión sin ningún problema —siempre que haya sido capaz de dar una interpretación razonable a los factores, claro—. Sin embargo, a diferencia del PCA, en donde las puntuaciones factoriales se derivaban directamente del modelo, como vimos, en el EFA es necesario estimarlas.

Estadísticos e-Books & Papers

Sigamos a Sharma (1996) para ver el procedimiento. La puntuación factorial para el individuo i en el factor j puede representarse como sigue:

$$\hat{F}_{ij} = \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots \hat{\beta}_p x_{ip} \tag{12.39}$$

donde $\hat{F}_{ij}$ es la estimación de la puntuación factorial del individuo i en el factor j. De forma matricial podemos representar esa ecuación como sigue:

$$\hat{\mathbf{F}} = \mathbf{X}\hat{\mathbf{B}} \tag{12.40}$$

donde ahora $\hat{\mathbf{F}}$ es una matriz $n \times m$ de m puntuaciones factoriales para los n individuos de nuestra base de datos. $\boldsymbol{X}$ es una matriz $n \times p$ que contiene las variables observadas —esto es, nuestra base de datos— y $\hat{\mathbf{B}}$ es una matriz $p \times m$ con los coeficientes de regresión para poder efectuar la estimación. En el caso en que las variables de $\boldsymbol{X}$ estén estandarizadas, llamaremos $\mathbf{Z}$ a la matriz que las contiene, por tanto reescribiremos:

$$\hat{\mathbf{F}} = \mathbf{Z}\hat{\mathbf{B}} \tag{12.41}$$

La ecuación anterior puede reescribirse del siguiente modo:

$$\frac{1}{n}\mathbf{Z}'\hat{\mathbf{F}} = \frac{1}{n}\mathbf{Z}'\mathbf{Z}\hat{\mathbf{B}} \tag{12.42}$$

o lo que es lo mismo:

$$\boldsymbol{\Lambda} = \mathbf{R}\hat{\mathbf{B}} \tag{12.43}$$

en la medida en que como hemos visto a lo largo del tema:

$$\frac{1}{n}(\mathbf{Z}'\hat{\mathbf{F}}) = \boldsymbol{\Lambda} \qquad \text{y} \qquad \frac{1}{n}(\mathbf{Z}'\mathbf{Z}) = \mathbf{R} \tag{12.44}$$

Por lo tanto, los coeficientes de regresión para estimar puntuaciones factoriales pueden calcularse como sigue:

$$\hat{\mathbf{B}} = \mathbf{R}^{-1}\boldsymbol{\Lambda} \tag{12.45}$$

y las puntuaciones factoriales:

$$\hat{\mathbf{F}} = \mathbf{Z}\mathbf{R}^{-1}\boldsymbol{\Lambda} \tag{12.46}$$

es decir, las puntuaciones factoriales no son otra cosa que una combinación de las variables originales tipificadas y de la matriz factorial. Obviamente, no podemos ilustrar el cálculo en el caso 12.1 porque, al tener solamente la matriz de correlaciones, no disponemos de datos individuales, esto es, no disponemos de $\mathbf{Z}$. Lo ilustraremos en el caso siguiente. El problema que tiene la estimación es que, como hemos visto, las indeterminación que genera la rotación —genera muchas matrices factoriales $\boldsymbol{\Lambda}$ diferentes— genera también distintos juegos de puntuaciones factoriales. Esto lleva a que los investigadores hayan diseña-

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 12.15.: Matriz $\mathbf{B}$ con los coeficientes de regresión para la estimación de las puntuaciones factoriales

```
$weights
          PC1        PC2
Gae 0.2407057  0.3974704
Eng 0.2519015  0.2570359
His 0.1893115  0.5641347
Ari 0.2701595 -0.3655885
Alg 0.2721945 -0.3323228
Geo 0.2482043 -0.3142357
```

do diferentes procedimientos de estimación cuya profundización está fuera del alcance de este manual, pero que pueden seguirse con detalle en Grice (2001).

En cualquier caso, la forma de solicitar esta información en el proceso de estimación es sencilla y la veremos en el siguiente caso. Los métodos disponibles en el paquete `fa{psych}` son los de Thurstone, tenBerge, Anderson, Bartlett, Harman y el método de componentes. Nótese, sin embargo, que, aunque no podamos calcular las puntuaciones factoriales a partir de una matriz de correlaciones sin datos individuales, sí que podemos calcular la matriz con los coeficientes de regresión, porque la expresión (12.45) solo requiere de esa matriz de correlaciones y la matriz factorial (cuadro 12.15).

```
factor.scores(R,fit.pca,method="Thurstone")
```

12.8. Un ejemplo de aplicación del análisis factorial exploratorio

Sharma (1996) plantea el siguiente caso.

Caso 12.2. Percepciones de los consumidores sobre marcas de refrescos
Una empresa de estudios de mercado ha realizado una encuesta a 95 consumidores para determinar sus percepciones sobre seis marcas de refrescos que compiten entre sí. Las marcas son las siguientes[4]: (1) Pepsi; (2) Coca Cola; (3) Gatorade; (4) Allsport; (5) Lipton; (6) Nestea. Para ello los entrevistados respondieron en una escala donde 1=Totalmente en desacuerdo hasta 7=Totalmente de acuerdo a las preguntas que aparecen en el cuadro 12.16. El objetivo es evaluar los factores que subyacen en la configuración de la imagen de las marcas y, a la vez, obtener el mapa perceptual de las seis marcas.

Después de lo expuesto durante el tema, optaremos por una extracción de factores por ejes principales con una rotación varimax. Pero antes hemos de ser capaces de determinar (a) si los datos son factorizables, es decir, si la matriz de correlaciones es distinta de la identidad y (b) si la muestra se adecua global

[4]Sharma (1996) señala que, aunque las marcas son reales, los datos son ficticios.

Estadísticos e-Books & Papers

Cuadro 12.16.: Preguntas para la medición de las percepciones

Etiqueta	Enunciado / Descripción
X1	La marca X tiene un sabor refrescante
X2	Prefiero X porque tiene menos calorías que otras marcas
X3	La marca X apaga mi sed inmediatamente
X4	Me gusta el sabor dulce de la marca X
X5	Prefiero X después del ejercicio porque me da energía
X6	Prefiero X porque viene en un envase respetuoso con el medioambiente
X7	X tiene minerales y vitaminas y me recupera además de quitar la sed
X8	X tiene un sabor muy distinto a las demás
X9	X tiene la combinación justa de minerales y vitaminas para ser saludable
X10	Prefiero tomar la marca X cuando tengo sed
Marca	1=Pepsi; 2=Coca Cola; 3=Gatorade; 4=Allsport; 5=Lipton; 6=Nestea
ID	Etiqueta para identificar al entrevistado

e ítem a ítem a la realización de un EFA (test KMO). Además de leer la base de datos que acompaña al libro que tiene, esta vez sí, datos individuales, no tenemos más remedio que calcular también la matriz de correlaciones porque algunos test, como el KMO, el test de esfericidad de Bartlett o el test de Bartlett para determinar el número de factores, solo pueden calcularse con ella en `R`. El cuadro 12.17 nos muestra que la matriz es factorizable en la medida en que el test de Barlett rechaza claramente que sea la matriz identidad. También que los datos de manera global son adecuados. Sin embargo, nótese que la pregunta X6 que aludía al envase ecológico tiene una MSA baja (0,59) y probablemente esté mal representada en la solución factorial, lo que nos obligará a hacer un seguimiento de su comportamiento —por ejemplo analizando su comunalidad y viendo en qué medida puede estar distorsionando la solución factorial—.

```
datos<-Datos_12_2_Caso
R<-cor(datos[,3:12],use="complete.obs",method="pearson")

cortest.bartlett(R,n=95)
KMO(R)
```

El paso siguiente es determinar el número de factores que deberemos extraer. Es necesario realizar primero este análisis con los criterios que en su momento señalamos, por ejemplo el análisis paralelo, en la medida en que la sintaxis de **`fa{psych}`** nos obliga a indicarle el número de factores que debe extraer antes de realizar la estimación. Utilizaremos, para alternar herramientas, la función **`paran{paran}`** para el análisis paralelo esta vez.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 12.17.: Resultados del test de esfericidad de Barlett y el test KMO

```
> cortest.bartlett(R,n=95)
$chisq
[1] 1054.053

$p.value
[1] 1.886679e-191

$df
[1] 45

> KMO(R)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = R)
Overall MSA =  0.81
MSA for each item =
  X1   X2   X3   X4   X5   X6   X7   X8   X9  X10
0.85 0.81 0.85 0.73 0.82 0.59 0.78 0.77 0.84 0.87
```

```
# autovalor>1, sedimentacion, paralelo
paran(datos[3:12], iterations=5000,graph=TRUE,color=FALSE)

# test de Bartlett
nBartlett(R, alpha=0.01, N=95, details=TRUE)
```

El análisis paralelo recomienda una solución de tres factores (cuadro 12.18). Si nos fijamos en el criterio del autovalor superior a la unidad (puede verse la línea =1 en la figura 12.7), también hay tres autovalores superiores a la unidad (la línea *Unadjusted EV* se corresponde con el gráfico de sedimentación del EFA). En el cuadro 12.18 se ofrece el test de Barlett para determinar el número de factores. Ya insistimos reiteradamente durante la presentación en su escasa fiabilidad al adolecer de las mismas limitaciones que el test de esfericidad. Su recomendación de cinco factores no se compadece con ninguno de los criterios anteriores.

Determinada que la mejor solución es la de tres factores, solicitamos con `fa{psych}` la estimación de un EFA mediante extracción por ejes principales `(fm=pa)`, con rotación `varimax (rotate=varimax)`, de los mencionados tres factores `(nfactors=3)` y, dado que ahora sí que contamos con datos individuales, que estime las puntuaciones factoriales mediante el procedimiento de regresión `(scores=regression)`. La opción `fa.sort(fit.pa)` simplemente ordena las cargas factoriales en cada factor por tamaño.

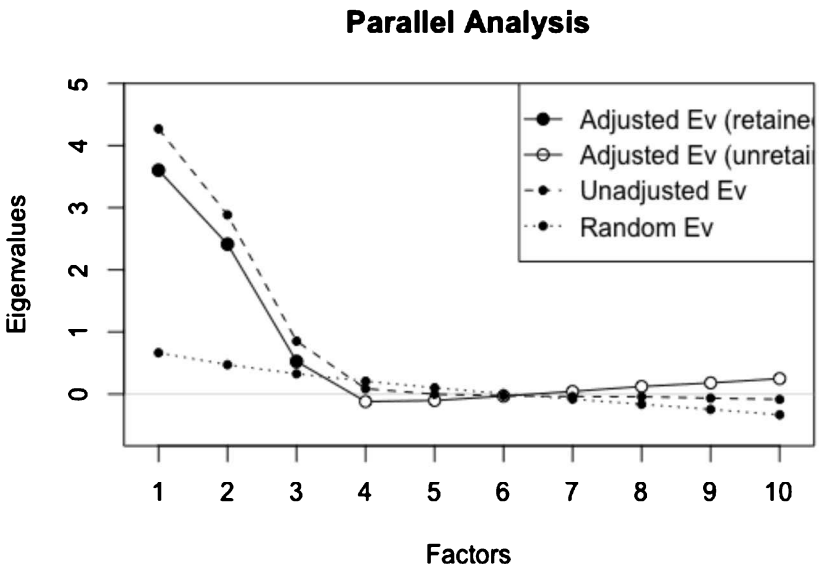
```
fit.pa<-fa(datos[,3:12],nfactors=3,fm="pa",rotate="varimax",
scores="regression")
fa.sort(fit.pa)
```


Estadísticos e-Books & Papers

Cuadro 12.18.: Resultados del análisis paralelo y el test de Barlett para la determinacion del número de factores que se deben extraer

```
Using eigendecomposition of correlation matrix.
Computing: 10%  20%  30%  40%  50%  60%  70%  80%  90%  100%


Results of Horn's Parallel Analysis for factor retention
5000 iterations, using the mean estimate

--------------------------------------------------
Factor      Adjusted    Unadjusted    Estimated
            Eigenvalue  Eigenvalue    Bias
--------------------------------------------------
No components passed.
--------------------------------------------------
1           3.603525    4.268841      0.665316
2           2.412755    2.885142      0.472387
3           0.525651    0.852166      0.326515
--------------------------------------------------

Adjusted eigenvalues > 0 indicate dimensions to retain.
(3 factors    retained)

> nBartlett(R, alpha=0.01, N=95, details=TRUE)
bartlett anderson   lawley
       5        5        5
```

Figura 12.7.: Resultado gráfico del análisis paralelo



© Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 12.19.: Matriz de cargas rotadas

```
      PA3   PA2   PA1    h2    u2 com
X2   0.93  0.09  0.21 0.919 0.081 1.1
X5   0.92  0.15  0.26 0.943 0.057 1.2
X9   0.92  0.11  0.27 0.940 0.060 1.2
X6  -0.14  0.06 -0.05 0.027 0.973 1.6
X4  -0.01  0.94 -0.09 0.886 0.114 1.0
X8   0.01  0.94 -0.22 0.922 0.078 1.1
X1   0.10  0.86 -0.22 0.794 0.206 1.2
X7   0.25 -0.23  0.91 0.950 0.050 1.3
X3   0.32 -0.17  0.86 0.875 0.125 1.4
X10  0.30 -0.20  0.86 0.862 0.138 1.4

                        PA3  PA2  PA1
SS loadings            2.86 2.66 2.60
Proportion Var         0.29 0.27 0.26
Cumulative Var         0.29 0.55 0.81
Proportion Explained   0.35 0.33 0.32
Cumulative Proportion  0.35 0.68 1.00
```

Los resultados del cuadro 12.19 muestran, como cabía esperar tras un KMO tan alto, una fuerte asociación de los indicadores a los factores, con una clara excepción que ya esperábamos, el indicador X6 (envase ecológico) que está muy mal representado por la solución factorial con una comunalidad muy baja ($h_6^2 = 0{,}027$). Ante una situación de este tipo caben dos opciones: (a) eliminar ese indicador o (b) reestimar el modelo sin él y solo eliminarlo en el caso en que hacerlo nos reduzca la complejidad del resultado, por ejemplo, ofreciendo una solución de dos factores en lugar de tres. Esto no va a ocurrir en este caso porque vemos que no existe un factor que esté formado solo por este indicador y los otros tres están muy asociados a sus respectivas variables. Por lo tanto lo mantendremos en la solución.

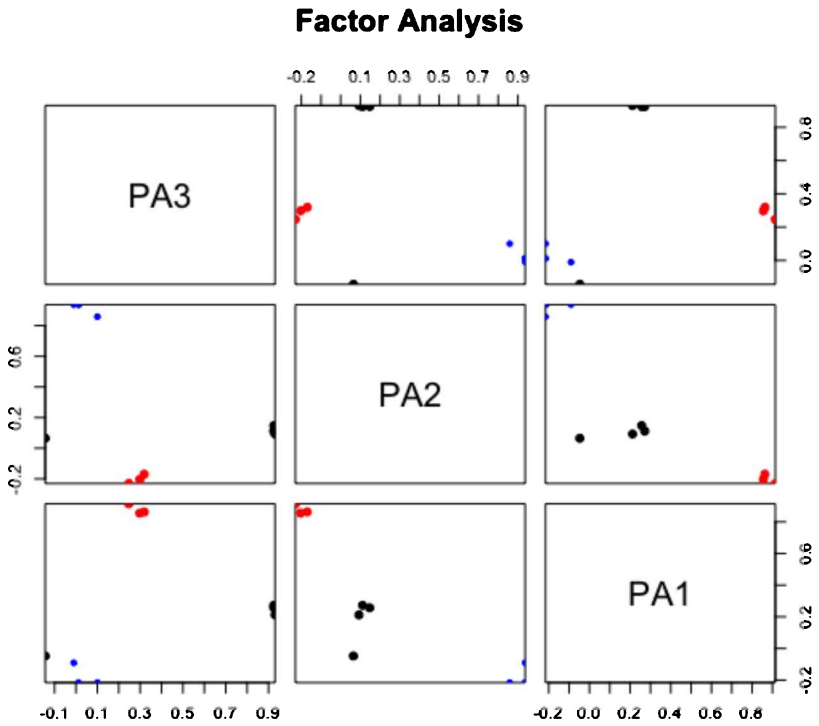
Vemos también que la solución de tres factores está recogiendo el 81 % de la información que contenían las 10 variables originales (`Cumulative Var` = 0,81) y que los tres ejes tienen una contribución similar a explicar la varianza (`Proportion Explained` es prácticamente la misma en los tres factores).

La figura 12.8 muestra las variables analizadas proyectadas sobre los factores (cargas factoriales) obtenidas mediante la opción `plot(fit.pa)`. Al haber retenido tres factores hemos de analizar tres gráficos de dispersión. Lo importante es ver que en cada par de factores las variables están separadas (agrupadas las que corresponden al mismo factor) a excepción de una de ellas que donde mejor se visualiza es en el gráfico de PA2 y PA3 que está aislada de las demás de su grupo y no se asocia a otro (es la mencionada variable X6 mal representada por la solución factorial).

Esta solución lo que nos está diciendo es que el factor PA3 recoge los *aspectos funcionales* de la bebida (pocas calorías, energía, minerales y vitaminas), PA2 recoge las características de su *sabor* (dulce, único, refrescante), mientras que

Estadísticos e-Books & Papers

Figura 12.8.: Resultado gráfico del análisis paralelo



 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 12.20.: Puntuaciones factoriales

```
> fit.pa$scores
               PA3          PA2          PA1
 [1,] -0.495125021 -1.366175134 -0.017137232
 [2,] -1.141231682 -0.538645457 -0.312603047
 [3,]  0.525880196 -0.638449790  0.632691684
 [4,]  0.460044509 -0.375229703  0.694110593
 [5,]  1.362596938  0.827181699  0.261718309
 [6,] -0.957723797  0.926644177  0.752913545
 [7,] -0.274921355  1.236038941  0.457614004
 [8,] -1.347489743  0.346133539  0.762545791
 [9,]  0.122255875 -1.077666639  1.591334252
[10,]  0.884309650 -0.483708382  0.862776624
```

PA1 engloba a los ítems que se relacionan con la *capacidad para apagar la sed.* Esos tres serían los ejes que subyacen en la valoración de las bebidas por parte de los consumidores.

Para realizar el mapa perceptual necesitamos extraer las puntuaciones factoriales que hemos solicitado con la opción `scores = regression`. Para ello basta extraerlas del objeto `fit.pa` que recoge los resultados del EFA mediante:

```
fit.pa$scores
```

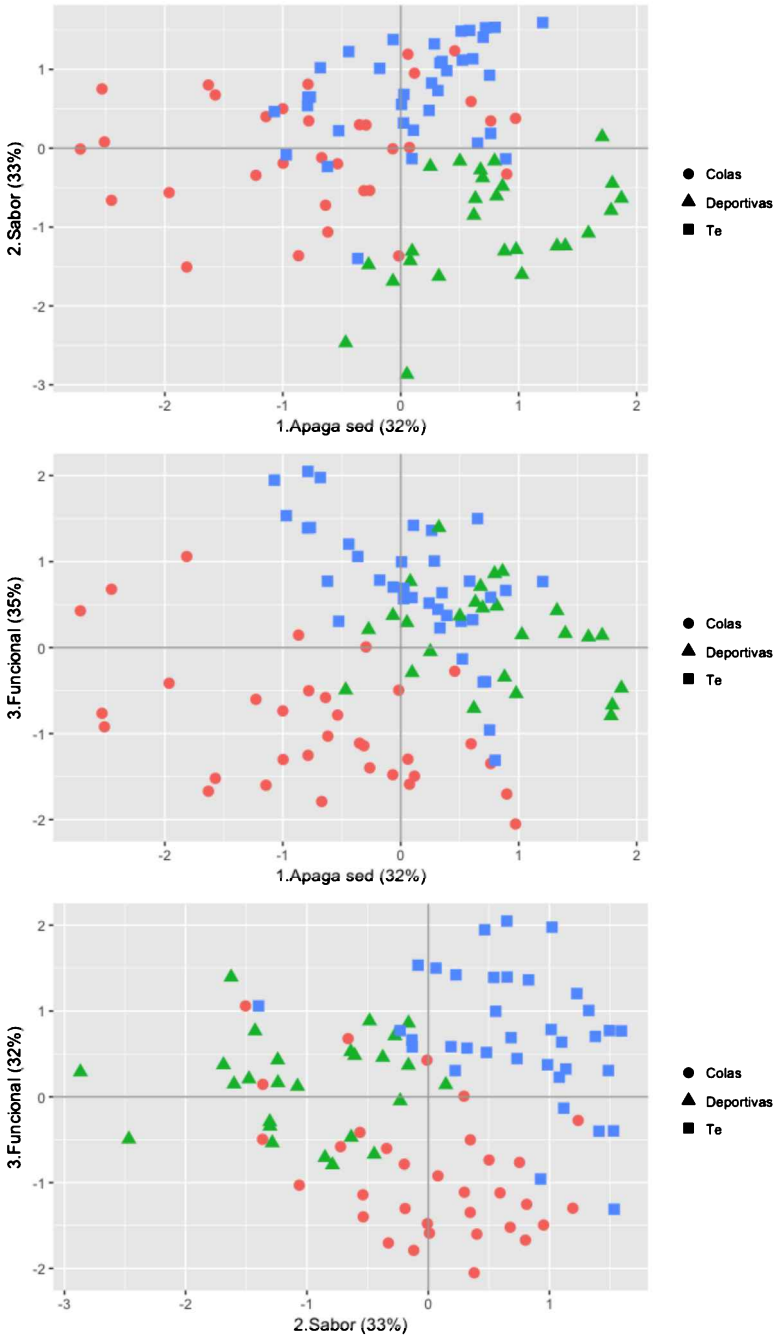
El cuadro 12.20 muestra las puntuaciones factoriales en los tres factores para una selección de los casos (los 10 primeros). Lo importante no es analizar las puntuaciones, sino utilizarlas para visualizar cómo se relacionan los factores con las valoraciones que ha hecho de las marcas cada individuo. Como tenemos muchas marcas, pero pueden agruparse en tres tipos, colas, deportivas y basadas en el té, las hemos agrupado para facilitar la interpretación que marca la figura 12.9.

Vemos que cada par de factores sirve para separar un grupo de bebidas del resto. Así la combinación del sabor (PA2) y apagar la sed (PA1) sirve para apartar a las bebidas deportivas, que están en el cuadrante inferior derecho del panel (a) del gráfico. El signo de las puntuaciones factoriales no debe interpretarse como, por ejemplo, buena o mala puntuación en el sabor, para eso hemos preparado el cuadro 12.21, que ofrece las medias de las variables asociadas a cada factor y nos permite concluir que las bebidas deportivas se distinguen del resto fundamentalmente por su buena capacidad para apagar la sed con un nivel intermedio de diferenciación de su sabor.

```
datos.medias<-data.frame(datos,datos.grafico$tipo.bebida)
datos.medias<-rename(datos.medias,
c("datos.grafico.tipo.bebida"="bebida"))
aggregate(cbind(X1,X2,X3,X4,X5,X7,X8,X9)~bebida,
data=datos.medias, mean, na.rm=TRUE)
```


Estadísticos e-Books & Papers

Figura 12.9.: Mapa perceptual de las bebidas



 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 12.21.: Promedio de las variables por tipo de bebida

```
      bebida       X1       X2       X3       X4       X5       X7       X8       X9
1      Colas 5.181818 2.484848 3.393939 5.030303 2.242424 3.424242 5.303030 2.454545
2 Deportivas 3.500000 4.423077 5.884615 3.269231 4.269231 5.961538 3.538462 4.500000
3         Te 6.138889 5.277778 4.916667 6.083333 5.333333 4.805556 5.972222 5.361111


      bebida       F1       F2       F3
1      Colas 3.434343 5.171717 2.393939
2 Deportivas 5.923077 3.435897 4.397436
3         Te 4.851852 6.064815 5.324074
```

```
aggregate
(cbind(F1=((X3+X7+X10)/3),F2=((X1+X4+X8)/3),F3=((X2+X5+X9)/3))
~bebida, data=datos.medias, mean, na.rm=TRUE)
```

El panel (b) nos muestra que las colas se separan del resto por la acción conjunta de apagar la sed (PA1) y sus características funcionales (PA3). Haciendo uso de nuevo del cuadro 12.21 vemos que a las colas las diferencia una capacidad intermedia para apagar la sed unida a pocas características funcionales.

Finalmente el panel (c) muestra como las bebidas basadas en el té vienen separadas del resto por la acción de los factores sabor (PA2) y características funcionales (PA3), siendo unas bebidas que puntúan alto en ambos elementos.

Podemos concluir, por tanto, que los factores subyacentes tienen una elevada capacidad para explicar las diferencias de percepción que tiene el consumidor de estos tipos de bebidas.

Estadísticos e-Books & Papers

Estadísticos e-Books & Papers

13. Modelos de ecuaciones estructurales: análisis factorial confirmatorio

13.1. Introducción

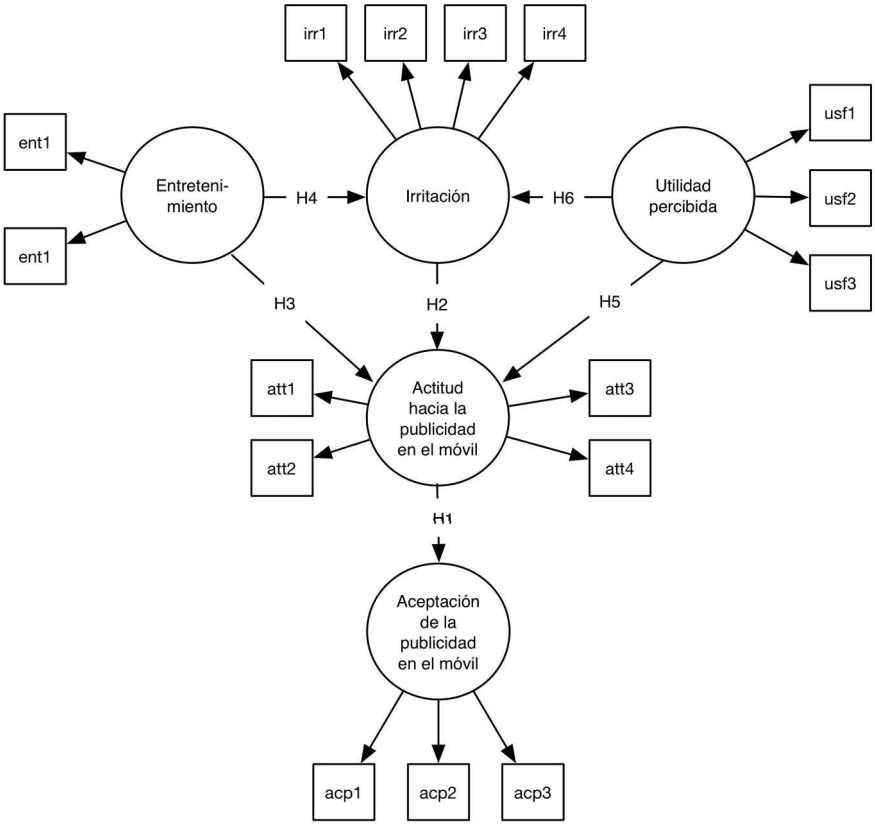
La primera advertencia que debe hacerse al lector es que los siguientes cuatro temas que forman parte de este manual deben considerarse una unidad integrada. Los temas 13, 14 y 15 son tres etapas en el proceso de estimación de un modelo estructural cuya lógica presentaremos inmediatamente, mientras que el capítulo 16 representa un enfoque distinto al anterior —modelos basados en varianzas frente a modelos basados en covarianzas— pero siempre para resolver un mismo problema: la estimación de relaciones múltiples en variables que, por la complejidad de los conceptos que recogen, necesitan de múltiples indicadores para su medición.

La figura 13.1 representa un ejemplo típico de modelo estructural utilizado habitualmente en investigación en psicología, sociología, *management* y marketing. Concretamente se corresponde con el trabajo de Aldás *et al.* (2013) que usaremos posteriormente como caso para el desarrollo de la herramienta. Por un lado, tenemos cinco variables latentes, los factores *entretenimiento, irritación, utilidad percibida, actitud hacia la publicidad en el móvil* y *aceptación de la publicidad en el móvil*, cuyas relaciones entre ellos no podrían estimarse con los procedimientos vistos hasta el momento, por ejemplo la relación lineal múltiple, porque no existe una única variable dependiente sobre la que influyen varias independientes, sino que factores como la *irritación* actúan simultáneamente como antecedentes de la *actitud* y como consecuentes de la *utilidad percibida* y el *entretenimiento*. La otra razón por la que no podríamos emplear una regresión para estimar el modelo planteado es que no existe una única variable manifiesta para medir cada factor, sino que la complejidad conceptual del mismo —pensemos en variables habituales en el campo del marketing como la calidad del servicio, la confianza, el riesgo percibido o la orientación al mercado— exigen de varios indicadores (una escala) para recoger todos los matices. Así, para mediar la *irritación* se ha elegido, como se verá con posterioridad, una escala de cuatro preguntas cuyos indicadores hemos representado como *irr1, irr2, irr3* e *irr4*.

Por lo tanto, el objetivo principal del investigador es contrastar las hipótesis que recoge la **parte estructural del modelo** (figura 13.2). Sin embargo, antes

Estadísticos e-Books & Papers

Figura 13.1.: Ejemplo de modelo de ecuaciones estructurales

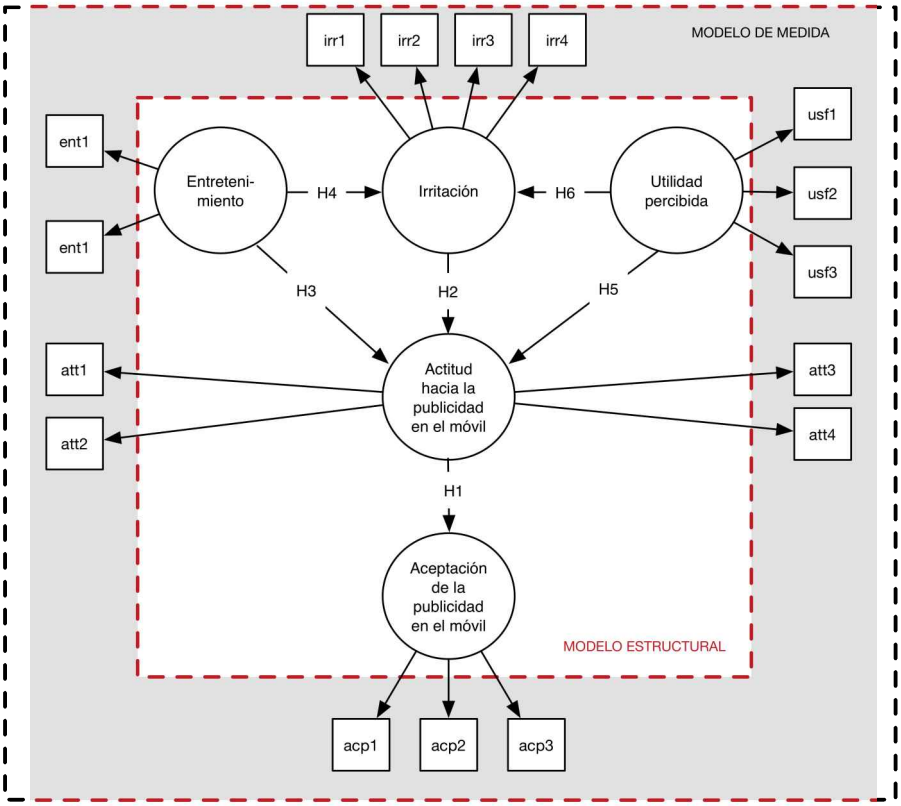


Fuente: Aldás *et al.* (2013)

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 13.2.: Ejemplo de modelo de ecuaciones estructurales



Fuente: Aldás *et al.* (2013)

de analizar —por ejemplo— la relación entre *irritación* y *actitud*, debe asegurarse que el **instrumento de medida** que ha utilizado para medir cada una de esas variables latentes es adecuado, cumple una serie de propiedades que le son exigibles y que en su momento abordaremos con detalle (p.ej. fiabilidad, validez convergente, validez discriminante, validez de contenido). Nos encontramos, por lo tanto, ante un proceso con varias etapas que sirven para estructurar esta serie de capítulos: en primer lugar habrá que validar el instrumento de medida, para lo que habrá que estimar un análisis factorial confirmatorio (capítulo 13) y con la información que nos proporciona, aplicarle una serie de criterios de análisis de la calidad de ese instrumento de medida (capítulo 14) para, solo entonces, estimar el modelo estructural y contrastar las hipótesis (capítulo 15).

Estadísticos e-Books & Papers

Cuadro 13.1.: Matriz de correlaciones entre las notas de los 275 estudiantes

	L	FSF	H	M	FSC	Q	Desv.típica
L	1.000						1.090
FSF	0.493	1.000					0.590
H	0.401	0.314	1.000				0.980
M	0.278	0.347	0.147	1.000			1.100
FSC	0.317	0.318	0.183	0.587	1.000		0.410
Q	0.284	0.327	0.179	0.463	0.453	1.000	1.110

13.2. Formalización matemática del análisis factorial confirmatorio (CFA)

Aunque las siglas naturales para el análisis factorial confirmatorio serían AFC, es bastante habitual referirse al mismo por las siglas que corresponderían a la misma terminología en inglés (CFA-*Confirmatory Factor Analysis*). Como en distintas salidas de los paquetes de `R` que utilizaremos aparece esa sigla, mantendremos el término CFA para la formalización matemática.

Para explicar de manera didáctica la formalización matemática de los CFA y el proceso de estimación, utilizaremos un ejemplo sencillo, el caso 13.1, que pretende ser un tributo a Spearman, con quien nació el concepto de factores latentes. Posteriormente, para consolidar lo expuesto y mostrar toda la potencia del paquete de `R` que utilizaremos en los procesos de estimación de los capítulos 13, 14 y 15 (`Lavaan`), lo aplicaremos a un caso complejo, que es el que hemos ilustrado en la figura 13.1.

Caso 13.1. Componentes de la inteligencia

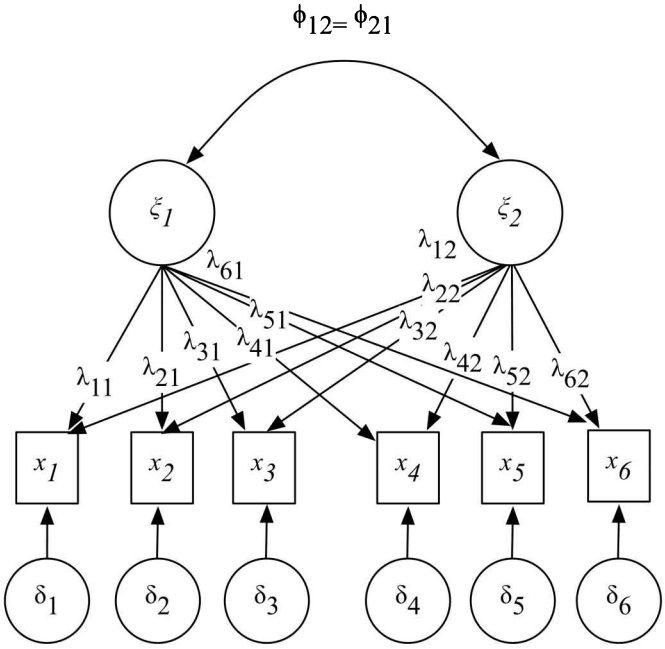
De acuerdo con el trabajo de Spearman (1904), tal y como lo presenta Sharma (1996), supongamos que un investigador ha recogido las notas de 275 alumnos de secundaria en seis asignaturas: lengua (L), filosofía (FSF), historia (H), matemáticas (M), física (FSC) y química (Q). En el cuadro 13.1 se recogen las correlaciones entre estas seis variables. Nuestro investigador se plantea una cuestión a la que quiere dar respuesta. Asumiendo que las notas de un alumno miden su inteligencia (I), desearía saber si estas se agrupan en un único factor (la inteligencia) o, por el contrario, miden distintos aspectos de la misma, por ejemplo, la inteligencia cuantitativa (IQ) y la inteligencia verbal (IV).

Si aplicamos la lógica que vimos en el capítulo 12 y asumimos que el investigador no tiene una hipótesis a priori acerca de qué estructura es la adecuada (un único componente de la inteligencia o dos), decidirá efectuar un análisis factorial exploratorio para ver cuántos factores obtiene. Su planteamiento aparece recogido gráficamente en la figura 13.3. Las **variables observadas** o **manifiestas** o **indicadores**, es decir, aquellas que se han medido (las notas de los alumnos en nuestro ejemplo), aparecen insertadas en un cuadrado para señalar que tienen este carácter y las hemos denotado como $X_1, ..., X_6$. Las **variables**

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 13.3.: Modelo de análisis factorial exploratorio



latentes, esto es las no observables o subyacentes (por ejemplo, factores como la inteligencia, en general, o la inteligencia verbal o cuantitativa, en particular), aparecen rodeadas por círculos. Una flecha recta desde una variable latente a una variable observada indica una relación de influencia directa. Así el factor ξ_1 está "causando" las notas de los alumnos en las seis asignaturas, es decir, la mayor o menor inteligencia "cuantitativa" provoca que los alumnos tengan notas diferentes. El término λ que aparece en cada una de las relaciones causales o *paths* es el parámetro que mide la intensidad de la relación, esto es, el término que hemos denominado **carga factorial** en una análisis factorial exploratorio y cuya denominación mantendremos en el confirmatorio.

Las variables latentes son de dos tipos: los mencionados **factores comunes** (ξ), que son comunes en cuanto que sus efectos son compartidos por más de una variable observada, y los **factores específicos o errores** (δ). Como se comprueba en la figura 13.1, cada uno de estos factores afecta solamente a una variable observada, y son errores aleatorios que se pueden haber producido en la medida de la variable observada —siguiendo nuestro ejemplo, a lo mejor no toda la nota se explica por uno u otro tipo de inteligencia, el estudiante puede ser bueno copiando en el examen o el profesor equivocarse al corregir— Finalmente, la flecha curva con dos puntas que une a los factores comunes indica que estas variables pueden estar **correlacionadas** con una intensidad

Estadísticos e-Books & Papers

ϕ.

Planteados los convenios de representación y los términos empleados en el CFA que son comunes a los de los modelos de estructuras de covarianza, que se examinarán en el próximo tema, nos restaría por señalar las diferencias del análisis factorial confirmatorio con respecto al análisis factorial exploratorio, examinado en el tema 12.

Volviendo a nuestro ejemplo, el investigador quiere saber si las notas están midiendo un único componente de la inteligencia o, por el contrario, reflejan el efecto de varios componentes. Como en un planteamiento exploratorio no tiene establecida una hipótesis a priori, su análisis factorial ha de contemplar como plausibles todas las posibilidades. Un caso extremo consistiría en que todas las variables carguen de forma significativamente intensa sobre un solo factor y muy poco sobre el otro. Un caso intermedio, aunque puede haber otras muchas combinaciones, consistiría en que un grupo de variables cargue significativamente sobre un factor y el resto de variables lo haga sobre un segundo factor. La figura 13.3 recoge todas las posibilidades y, en concreto, estos dos casos. En el primer caso —el factor único— $\lambda_{11}, \lambda_{21}, ..., \lambda_{61}$ serían significativos, mientras que $\lambda_{12}, \lambda_{22}, ..., \lambda_{62}$ no lo serían o, siéndolo, su tamaño de carga sería claramente inferior. En el segundo caso —dos factores—, λ_{11}, λ_{21} y λ_{31} tendrían un valor significativo, y λ_{41}, λ_{51} y λ_{61}, no (las notas en literatura, filosofía e historia cargan sobre un factor, inteligencia verbal, y no sobre el otro); por otra parte, λ_{12}, λ_{22} y λ_{32} no tendrían un valor significativo, y λ_{42}, λ_{52} y λ_{62}, sí (las notas en matemáticas, física y química cargan sobre un factor, la inteligencia cuantitativa). El investigador debe efectuar un análisis factorial exploratorio con objeto de averiguar cuál de las dos posibilidades (o cualquiera de las otras muchas que sugiere la figura 13.3) es más verosímil de acuerdo con los datos.

Ahora bien, el investigador, basándose en estudios previos o en una revisión de la literatura existente, puede considerar la hipótesis, por ejemplo, de que no existe una medida global de la inteligencia sino dos tipos alternativos de la misma: inteligencia verbal (que explicaría las calificaciones en lengua, filosofía e historia) e inteligencia cuantitativa (que explicaría las obtenidas en matemáticas, física y química). Si este es el caso, el análisis exploratorio ya no tiene sentido, ya que el investigador lo que pretende es confirmar o no la verosimilitud de su hipótesis. Su planteamiento aparece recogido ahora en la figura 13.4.

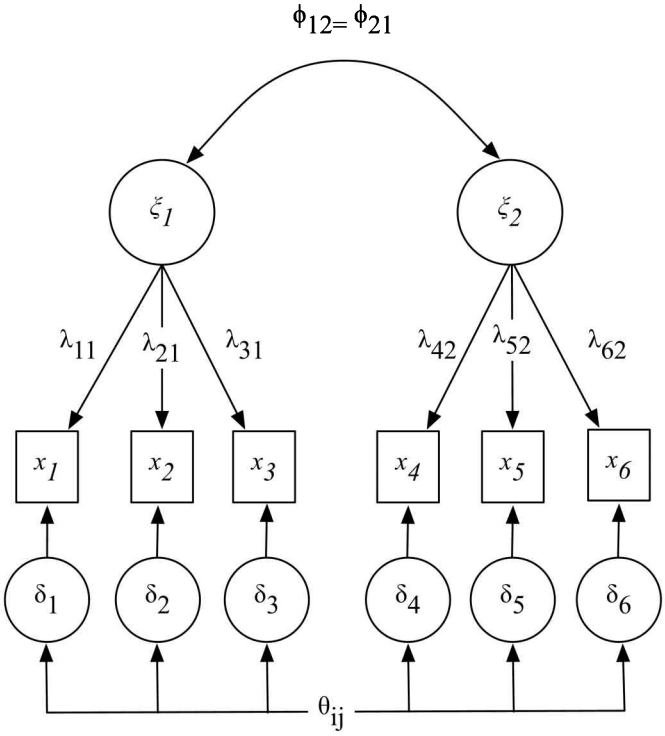
Ha de notarse que este es el planteamiento que seguimos en la modelización estructural. Primero definimos las variables latentes y luego buscamos indicadores —escalas— para medirlas. No tiene sentido que nos planteemos sobre quién deben cargar los indicadores, pero sí debemos confirmar que el modelo en que cargan sobre los factores que definen encaja con los datos muestrales, es plausible.

Pues bien, a partir del sencillo ejemplo que estamos siguiendo y que se ilustra en la figura 13.4, presentaremos a continuación la formalización del mismo siguiendo la notación Jöreskog y Sorborm (1996). La relación entre las variables observadas y las latentes de la figura 13.4 puede expresarse:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 13.4.: Modelo de análisis factorial confirmatorio



Estadísticos e-Books & Papers

$$\begin{aligned} x_1 &= \lambda_{11}\xi_1 + \delta_1 \\ x_2 &= \lambda_{21}\xi_1 + \delta_2 \\ x_3 &= \lambda_{31}\xi_1 + \delta_3 \\ x_4 &= \lambda_{42}\xi_2 + \delta_4 \\ x_5 &= \lambda_{52}\xi_2 + \delta_5 \\ x_5 &= \lambda_{62}\xi_2 + \delta_6 \end{aligned} \tag{13.1}$$

Si recurrimos a la notación matricial, la anterior expresión adoptaría la forma:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \end{bmatrix} \tag{13.2}$$

o de manera compacta:

$$\mathbf{x} = \Lambda\xi + \delta \tag{13.3}$$

donde, en general y no solo para este modelo, $\mathbf{x}$ es un vector $q \times 1$ que contiene las q variables observadas o indicadores, ξ es un vector $s \times 1$ que contiene los s factores comunes, Λ es una matriz $q \times s$ que contiene las cargas factoriales y δ es el vector $q \times 1$ de los factores específicos o errores. Asumimos que —por lógica— el número de variables observadas será siempre mayor que el de factores comunes o, lo que es lo mismo, que $q > s$.

Tanto las variables latentes como las observadas de la expresión 13.3 vienen expresadas como desviaciones sobre la media, con lo que la esperanza de cada vector es otro vector de ceros:

$$E(\mathbf{x}) = \mathbf{0};\ E(\xi) = \mathbf{0};\ \mathbf{E}(\delta) = \mathbf{0}$$

Este desplazamiento respecto al origen no afecta a las covarianzas entre las variables. Si denotamos como Σ a la matriz de varianzas y covarianzas entre las variables observadas (vector $\mathbf{x}$), de acuerdo con la expresión 13.3, resulta que:

$$\Sigma = E(\mathbf{x}\mathbf{x}') = E\left[(\Lambda\xi + \delta)(\Lambda\xi + \delta)'\right] \tag{13.4}$$

Teniendo en cuenta que la traspuesta de una suma de matrices es la suma de las traspuestas, y que la traspuesta de un producto es el producto de las traspuestas en orden inverso, tenemos que:

$$\Sigma = E\left[(\Lambda\xi + \delta)(\xi'\Lambda' + \delta')\right] \tag{13.5}$$

y teniendo en cuenta la propiedad distributiva y calculando la esperanza:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

$$\begin{aligned}\Sigma &= E\left[\Lambda\xi\xi'\Lambda' + \Lambda\xi\delta' + \delta\xi'\Lambda' + \delta\delta'\right] \\ &= E\left[\Lambda\Sigma = \Lambda'\right] + E\left[\Lambda\xi\delta'\right] + E\left[\delta\xi'\Lambda'\right] + E\left[\delta\delta'\right]\end{aligned} \tag{13.6}$$

Dado que la matriz Λ no contiene variables aleatorias, al ser constantes los parámetros poblacionales, se tiene que:

$$\Sigma = \Lambda E\left[\xi\xi'\right]A' + \Lambda E\left[\xi\delta'\right] + E\left[\delta\xi'\right]\Lambda' + E\left[\delta\delta'\right] \tag{13.7}$$

Si denotamos:

$$\Phi = E\left[\xi\xi'\right]$$

$$\Theta = E\left[\delta\delta'\right]$$

y asumimos que δ y ξ están incorrelacionados, la expresión 13.7 puede escribirse del siguiente modo:

$$\Sigma = \Lambda\Phi\Lambda' + \Theta \tag{13.8}$$

Es muy importante, para desarrollos posteriores, analizar el contenido de la expresión 13.8. Así, en el primer miembro aparece una matriz que contiene $q(q+1)/2$ varianzas y covarianzas distintas de las variables observadas[1]. En el segundo miembro aparecen $q \times s$ cargas factoriales (Λ), $s(s+1)/2$ varianzas y covarianzas entre los factores comunes (ξ) y $q(q+1)/2$ varianzas y covarianzas entre los errores (δ). Por lo tanto, la expresión 13.8 expresa los $q(q+1)/2$ elementos distintos de la matriz de varianzas y covarianzas poblacionales que representa al modelo teórico de la figura 13.4 en función de los $[qs + s(s+1)/2 + q(q+1)/2]$ parámetros desconocidos que habrá que estimar de las matrices Λ, Θ y Φ. Así pues, los parámetros que se deberían estimar aparecen vinculados mediante la expresión 13.8 a los valores de las varianzas y covarianzas poblacionales de las variables observadas.

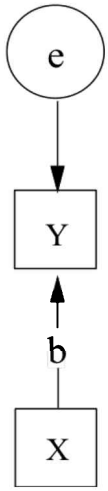
Aplicado a nuestro ejemplo, las matrices que contienen los parámetros que se deben estimar adoptarán la forma siguiente:

$$\Lambda = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix} \quad \Phi = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{bmatrix}$$

[1] Para determinar el número de varianzas y covarianzas distintas, téngase en cuenta que Σ es una matriz $q \times q$ simétrica.

Estadísticos e-Books & Papers

Figura 13.5.: Modelo estructural elemental



$$\Theta = \begin{bmatrix} \theta_{11} & \theta_{21} & \theta_{31} & \theta_{41} & \theta_{51} & \theta_{61} \\ \theta_{21} & \theta_{22} & \theta_{32} & \theta_{42} & \theta_{52} & \theta_{62} \\ \theta_{31} & \theta_{32} & \theta_{33} & \theta_{43} & \theta_{53} & \theta_{63} \\ \theta_{41} & \theta_{42} & \theta_{43} & \theta_{44} & \theta_{54} & \theta_{64} \\ \theta_{51} & \theta_{52} & \theta_{53} & \theta_{54} & \theta_{55} & \theta_{65} \\ \theta_{61} & \theta_{62} & \theta_{63} & \theta_{64} & \theta_{65} & \theta_{66} \end{bmatrix}$$

Nótese que la matriz Θ tiene $6(6+1)/2 = 21$ elementos distintos por estimar y no $6 \times 6 = 36$, dado que las correlaciones entre los errores cumplen que $\theta_{ij} = \theta_{ji}$, por lo que los elementos del triángulo inferior se han repetido en el superior.

Visto de manera muy intuitiva, ¿a qué se reduce, a grandes rasgos, el método CFA? La finalidad de este método es obtener estimaciones para los parámetros que contienen las matrices Λ, Θ y Φ que hagan que la matriz de varianzas y covarianzas poblacional Σ estimada, obtenida a partir de ellas, sea lo más parecida posible a la matriz de varianzas y covarianzas muestral $\mathbf{S}$ que se obtiene a partir de los valores muestrales de las variables observadas. Veámoslo con un ejemplo todavía más sencillo, el más sencillo que podemos plantear, una regresión con una variable explicativa que, al fin y al cabo, no deja se ser un caso particular de un modelo estructural con una variable latente que solo tiene un indicador. La representación está en la figura 13.5.

Ilustración 13.1

¿De qué datos dispondríamos en nuestra base de datos en una regresión de la variable dependiente Y sobre la variable independiente X? Supongamos, como ilustración, que las del cuadro 13.2. Dado que, X e Y las tenemos medidas,

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 13.2.: Datos simulados para la ilustración

X	Y	X	Y
1	1.1	11	10.4
2	2.0	12	12.4
3	3.1	13	13.3
4	3.9	14	13.9
5	4.5	15	14.7
6	5.8	16	16.1
7	6.9	17	16.9
8	7.7	18	17.8
9	8.9	19	19.2
10	9.8	20	20.5

$Var(X)$	$Var(Y)$	$Cov(X,Y)$
35.000	35.935	35.429

podríamos calcular fácilmente la varianza de X $[Var(X)]$, la varianza de Y $[Var(Y)]$ y la covarianza entre X e Y $[Cov(X,Y)]$. Esas estimaciones serían números obtenidos de nuestra muestra en la matriz de varianzas y covarianzas muestrales **S**. Por ejemplo:

$$S = \begin{bmatrix} 35{,}935 & 35{,}429 \\ 35{,}429 & 35{,}000 \end{bmatrix} \tag{13.9}$$

Por su parte, el modelo teórico adopta la expresión elemental:

$$Y = a + bX + e$$

resultando elemental derivar las expresiones de las varianzas y covarianzas teóricas:

$$Var\,(Y) = b^2 Var\,(X) + Var\,(e)$$

$$Cov\,(X,Y) = bVar\,(X)$$

por lo que la matriz teórica de varianzas y covarianzas Σ adoptaría la expresión:

$$\Sigma = \begin{bmatrix} b^2 Var\,(X) + Var\,(e) & bVar\,(X) \\ bVar\,(X) & Var\,(X) \end{bmatrix} \tag{13.10}$$

donde los parámetros por estimar son b y $Var(e)$.

Se trata de estimar esos parámetros de tal forma que las dos matrices coincidan, dado que no hay grados de libertad (en el caso del CFA se tratará de estimar los parámetros de manera que las dos matrices se parezcan al máximo, pues siempre habrá grados de libertad y no podrán estimarse de manera

Estadísticos e-Books & Papers

Cuadro 13.3.: Resultados de la estimación

```
Call:
lm(formula = Y ~ X, data = datos)

Residuals:
     Min       1Q   Median       3Q      Max
-0.55113 -0.15774 -0.01436  0.18111  0.43857

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1837     0.1281  -1.433    0.169
X             1.0123     0.0107  94.632   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 0.2758 on 18 degrees of freedom
Multiple R-squared:  0.998,   Adjusted R-squared:  0.9979
F-statistic:  8955 on 1 and 18 DF,  p-value: < 2.2e-16
```

determinista como en el ejemplo). De las expresiones (13.9) y (13.10) es fácil obtener el coeficiente de regresión igualando la celda (2,1) de S con la (2,1) de Σ:

$$bVar\,(X) = b \times 35{,}000 = 35{,}429 \rightarrow b = 1{,}0123$$

y análogamente para la $Var(e)$. Puede comprobarse como, si realizamos efectivamente la regresión en `R` mediante la función `lm{stats}`, el resultado es exactamente ese (cuadro 13.3).

```
summary(lm(Y~X,datos))
```

Pero para poder entrar en el procedimiento de estimación es necesario abordar previamente el problema de la identificación que se plantea en el método CFA.

13.3. La identificación del modelo en un CFA

En la sección anterior, hemos visto que en el método CFA disponemos de una serie de datos (las varianzas y covarianzas muestrales de las variables observadas) y con ellos hemos de estimar una serie de parámetros (cargas factoriales, varianzas y covarianzas de los factores comunes, y varianzas y covarianzas de los factores específicos o errores). Al igual que ocurre con un sistema de ecuaciones lineales, podemos disponer en principio de más ecuaciones que incógnitas, del mismo número o de mayor número de incógnitas que ecuaciones. Cada una de esas situaciones lleva a distintas posibilidades: el sistema de ecuaciones tiene infinita soluciones, tiene una única solución o no puede estimarse. Pues bien, la

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

identificación del modelo en el CFA hace referencia, precisamente, a la cuestión de si los parámetros del modelo pueden o no ser determinados de forma única.

En palabras de Long (1983a), si se intenta estimar un modelo que no esté identificado, los resultados que se obtendrán serán estimaciones arbitrarias de los parámetros, lo que desembocará en interpretaciones carentes de sentido. En el apéndice A13.1 se demuestra cómo, si no se imponen restricciones a los parámetros que hay que estimar, necesariamente habrá un número infinito de soluciones posibles para los mismos.

¿Qué tipo de restricciones pueden imponerse a los parámetros? Por ejemplo, si una carga factorial λ_{ij} de la matriz se fija a 0, estaremos indicando que el factor j no afecta causalmente a la variable observada x_i. Si fijamos a 0 el elemento ϕ_{ij} de la matriz Φ, estaremos señalando que los factores i y j están incorrelacionados. Si todos los elementos de la matriz Φ fuera de la diagonal se fijan a 0, los factores serán ortogonales (como ocurre en el análisis factorial exploratorio, por ejemplo). Restricciones similares se pueden imponer a los elementos de la matriz.

Long (1983a) señala que existen una serie de condiciones para que el modelo esté identificado: *necesarias* (si no se dan, el modelo no está identificado), *suficientes* (si se dan, el modelo está identificado, pero, si no se dan, no tiene por qué no estarlo) y *necesarias y suficientes* (si se dan, el modelo está identificado y si no se dan, está no identificado). No hay acuerdo entre la literatura acerca de si existen o no las condiciones necesarias y suficientes. Jöreskog y Sorborm (1996) señalan que el análisis de la llamada *matriz de información*, construida a partir de la matriz de varianzas y covarianzas de los estimadores de los parámetros, puede servir para establecer si el modelo está identificado. Estos autores señalan que "si la matriz de información es definida positiva es *casi seguro* que el modelo está identificado. Por el contrario, si la matriz de información es singular, el modelo no está identificado". Las cursivas son de Long (1983a) y las introduce porque indica que, dado que los programas existentes verifican esta condición, si no hacen advertencias acerca de problemas en esta matriz, estaríamos ante un buen indicador de que el modelo está identificado pero, en su opinión, aun siendo la matriz definida positiva, es posible, aunque improbable, que el modelo no esté identificado. Otros autores, como Hatcher (1994) y Ullman (2001) confían también en las advertencias de los programas como indicadores de no identificación. En general, la mayoría de textos optan por recomendar que se comprueben una serie de condiciones necesarias que suelen demostrarse como lo suficientemente exigentes para garantizar la identificación del modelo. Siguiendo a Hatcher (1994) y Ullman (2001), el investigador debería centrarse en las siguientes tareas:

1. Comparar el número de datos con el número de parámetros que han de estimarse. Los datos son siempre las varianzas-covarianzas muestrales, y hemos visto que existen $q(q+1)/2$. Como el número de parámetros por estimar es $qs + [s(s+1)/2] + [q(q+1)/2]$, el modelo estará sin identificar si no se imponen, al menos, $qs + [s(s+1)/2]$ restricciones. Decimos "al

Estadísticos e-Books & Papers

menos" porque solo si hay más datos que parámetros, el modelo está **sobreidentificado** (caso particular de identificación), lo que hace que, al existir grados de libertad, sea posible la aceptación o el rechazo del modelo.

2. **Establecer una escala** para los factores comunes. Esto se consigue fijando la varianza de cada factor común a 1 o fijando a 1 una de las cargas de las variables observadas para cada factor. Si esto no se hace, se produce el denominado problema de indeterminación entre la varianza y las cargas factoriales, es decir, es imposible distinguir entre los casos en los que un factor tiene una varianza grande y las cargas son pequeñas y el caso en que las varianzas son pequeñas y las cargas altas. Veremos en el tema siguiente que saber que una carga grande viene ocasionada por correlaciones elevadas entre los indicadores es importante.

3. Asegurar la **identificabilidad de la parte del modelo que contiene la relación entre las variables observadas y los factores**. Para ello debe analizarse el número de factores y el número de variables observadas que cargan sobre cada factor. Si solo hay *un factor*, el modelo puede estar identificado si el factor tiene al menos *tres variables* con cargas no nulas sobre él. Si hay *dos o más factores*, examínese el número de variables observadas de cada factor. Si cada factor tiene *tres o más indicadores* que cargan sobre él, el modelo puede estar identificado si los errores asociados con los indicadores no están correlacionados entre sí, cada variable carga solo sobre un factor y los factores pueden covariar entre ellos. Si solo hay *dos indicadores por factor*, el modelo puede estar indentificado si los errores asociados con cada indicador no están correlacionados, cada indicador carga solo sobre un factor y ninguna de las covarianzas entre los factores es igual a cero. Vemos que, por lo expuesto, es una buena opción **fijar siempre a 0 los coeficientes de correlación entre los términos de error,** sin que sea un problema, como veremos, liberar posteriormente alguno de ellos si es necesario para favorecer el ajuste del modelo.

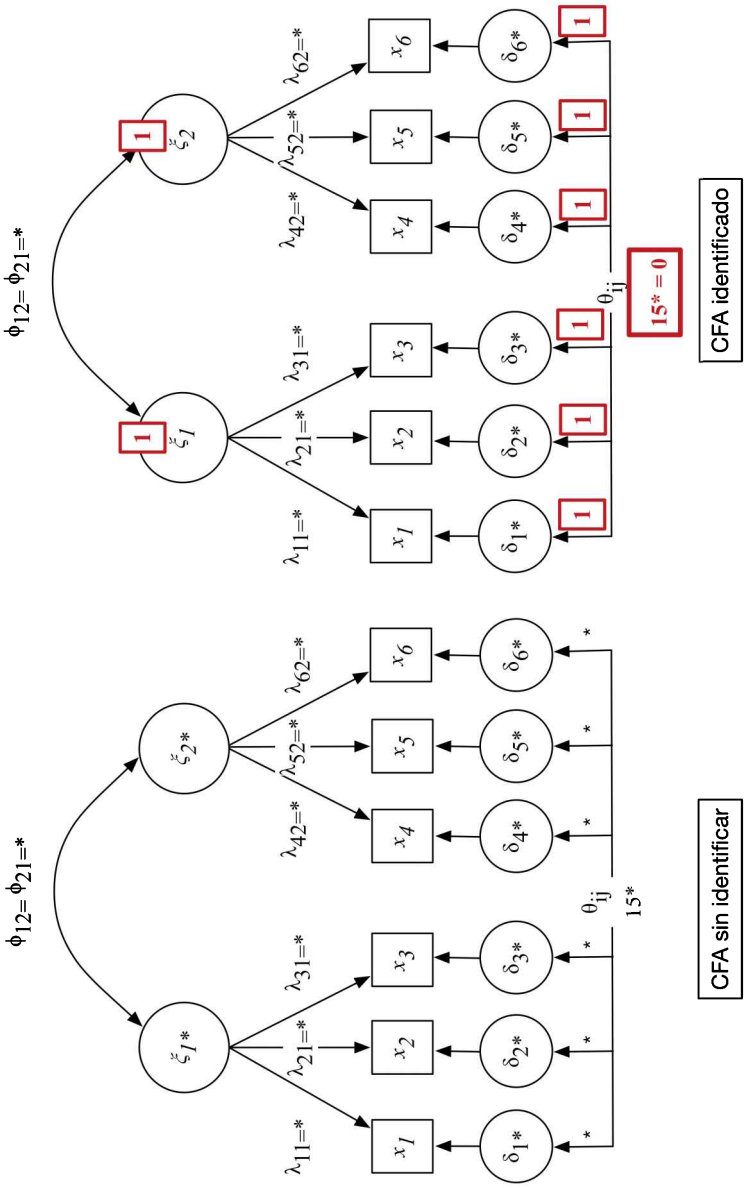
4. También **fijaremos siempre a 1 los coeficientes de regresión que unen los errores con los indicadores**, al igual que hacemos en una regresión lineal, donde esta se estima como $y = \beta_0 + \beta_1 X + \epsilon$ y no $y = \beta_0 + \beta_1 X + \beta_2 \epsilon$, fundamentalmente porque tendríamos —en el caso de la regresión— grados de libertad negativos y ese parámetro no aporta nada a la interpretación.

En la figura 13.6 ilustramos los pasos de identificación. Vamos a comprobar que esta se ha logrado. En primer lugar, hemos de recordar qué parámetros son los que componen la matriz de varianzas y covarianzas teórica en nuestro ejemplo (los hemos marcados con un * en el panel a) de la figura 13.6. Inicialmente en cualquier CFA hay que estimar:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 13.6.: CFA antes y después de la identificación



Estadísticos e-Books & Papers

- Las **correlaciones entre todos los factores independientes** (en un CFA todos son independientes, veremos el sentido de plantearlo así en el tema siguiente, cuando las relaciones estructurales conviertan en dependientes a algunos factores).
- Las **varianzas de todos los factores independientes** (mismo comentario anterior, todos lo son en un CFA pero no en un SEM).
- Todas las **cargas factoriales**.
- Las **varianzas de los errores**.
- Las **covarianzas entre los términos de error**. Aunque ya hemos señalado que el primer paso de la identificación es fijarlos a 0.
- Los **coeficientes de regresión entre los términos de error y los indicadores**, aunque también hemos señalado que siempre se fijan a 1 en la identificación.

Aplicado a nuestro caso:

- 1 correlación ϕ_{12} entre los factores ξ_1 y ξ_2.
- 2 varianzas de los factores ξ_1 y ξ_2, esto es, ϕ_{11} y ϕ_{22}.
- 6 cargas factoriales λ_{ij} que unen los factores ξ_1 y ξ_2 con los 6 indicadores x_i.
- 6 coeficientes de regresión entre los términos de error y los indicadores.
- 6 varianzas de los términos de error δ_i, que hemos denotado como $\theta_{11} \cdots \theta_{66}$.
- 15 covarianzas entre los términos de error δ_i, que ya hemos apuntado fijaremos a 0 inicialmente en el proceso de identificación.

¿Qué decisiones de identificación hemos tomado?

1. Hemos **establecido la escala de medida** de cada factor fijando a 1 la varianza del mismo. Es verdad que podíamos haber fijado como alternativa a 1 una carga por cada factor. Pero ya veremos en el próximo tema que uno de los criterios para la validez convergente es constatar que todas las cargas son significativas. Cuando un parámetro se fija a un valor, este no se estima y, por ello, si fijamos las cargas, no podríamos establecer su significatividad. Dado que ningún criterio de fiabilidad o validez afecta a las varianzas de los factores, es preferible la alternativa elegida.
2. Aplicamos las **condiciones para asegurar la identificabilidad** habituales:
   a) Hemos fijado a cero las 15 correlaciones entre los términos de error.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

b) Hemos fijado a 1 los 6 coeficientes de regresión entre los términos de error y los indicadores.

c) Sí que se ha permitido que las covarianzas entre los factores sean no nulas (la única existente, ϕ_{12}, aparece con un * para ser estimada).

3. Vamos a comprobar que **contamos con grados de libertad**, es decir, tenemos más datos (varianzas y covarianzas muestrales) que parámetros para estimar y el modelo está **sobreidentificado**. Recordemos que disponemos para estimar el modelo señalado de $6(6+1)/2 = 21$ datos, que se corresponden con las varianzas y covarianzas de las variables observadas. Tenemos que estimar, en principio, $6 \times 2 + (23/2) + (67/2) = 36$ parámetros. Estos parámetros son 6 coeficientes de regresión (cargas factoriales), 6 coeficientes de regresión entre los indicadores y los términos de error, la varianza de los 2 factores comunes, la covarianza entre ellos, 6 coeficientes de regresión entre las variables observadas y los factores específicos, las 6 varianzas de los factores específicos y las 15 covarianzas entre esos factores específicos. Si no hacemos nada, el modelo estaría infraidentificado, pero como hemos fijado 2 varianzas + 6 coeficientes de regresión + 15 covarianzas = 23 parámetros, solo quedan por estimar $36 - 23 = 13$ parámetros. Por lo tanto contamos con $21 - 13 = 8$ grados de libertad.

Antes de entrar en el proceso de estimación, veremos cómo hay que definir la sintaxis en `lavaan` para que esta estimación pueda tener lugar. El hecho de que hayamos tomado los datos del trabajo de Sharma (1996), que los proporciona mediante la matriz de correlaciones y desviaciones típicas de los indicadores y no como datos individuales (*raw data*), nos va a permitir también abordar la introducción de datos en este formato, lo que no es habitual, dado que normalmente el investigador ha recogido sus propios datos y los tiene como *raw data*. Esto abre la posibilidad al investigador de tomar datos publicados en libros o artículos y reproducir los análisis, en la medida en que es habitual que siempre se ofrezcan las matrices de covarianzas o las de correlaciones.

Comenzamos explicando la introducción de los datos que aparecen en el cuadro 13.1 y que, como se aprecia, vienen dados como una matriz de correlaciones, las desviaciones típicas de los indicadores y se informa que están obtenidos de 275 casos.

```
x <- c(1.000,
        .493,1.000,
        .401,.314,1.000,
        .278,.347,.147,1.000,
        .317,.318,.183,.587,1.000,
        .284,.327,.179,.463,.453,1.000)

#Convertimos el vector x en la matriz datos.cor
```


Estadísticos e-Books & Papers

```
datos.cor<-lav_matrix_lower2full(x)

#Etiquetamos a las variables de la matriz

colnames(datos.cor) <- rownames(datos.cor) <-
   c("L","FSF", "H","M","FSC","Q")

#Introducimos las desviaciones tipicas SD y las etiquetamos

datos.sd <- c(1.090, 0.590, 0.980, 1.100, 0.410, 1.110)
names(datos.sd) <-
   c("L","FSF", "H","M","FSC","Q")

#Convertimos las correlaciones y desviaciones tipicas
#en varianzas y covarianzas

datos.cov<-cor2cov(datos.cor,datos.sd)
```

Como vemos, lo primero que se hace es generar un vector `x` que tiene todas las correlaciones en línea. Aunque en la sintaxis parece una matriz triangular inferior, lo mostramos así para facilitar su lectura, pero es simplemente un vector con las correlaciones consecutivas. Pero el vector se ha de transformar en matriz, eso lo hacemos mediante la función **`lav_matrix_lower2full {lavaan}`**, que coge el vector `x` y lo transforma en una matriz, ahora sí, triangular inferior que llamamos **`datos.cor`** y luego etiquetamos para que filas y columnas tengan los nombres de las variables mediante la función **`colnames{base}`**. El problema es que el CFA se estima ajustando las matrices de varianzas y covarianzas, no de correlaciones, por lo que es necesario pasar de una a otra. Pero solo se pueden obtener las covarianzas a partir de las correlaciones si se conocen las desviaciones típicas, como es el caso. Las introducimos como un vector al que llamamos **`datos.sd`**. Finalmente utilizamos la función **`cor2cov{lavaan}`** para convertir la matriz de correlaciones en matriz de varianzas y covarianzas dadas las desviaciones típicas. Ya tenemos los datos preparados para introducir la parte de la sintaxis que define el modelo, que sería la siguiente:

```
modelo.cfa <- '

# Modelo de medida

IV=~L+FSF+H
IQ=~M+FSC+Q

#Varianzas de los factores
```

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

```
IV~~1*IV
IQ~~1*IQ

#Covarianzas

IV~~IQ

#Varianzas de los t<U+00E9>rminos de error

L~~L
FSF~~FSF
H~~H
M~~M
FSC~~FSC
Q~~Q
'
#Estimacion del modelo

fit <- lavaan(modelo.cfa, sample.cov=datos.cov,
sample.nobs=275,
std.lv=TRUE,mimic="eqs",
estimator="ML", verbose=TRUE, warn=TRUE)
```

Como vemos, la sintaxis es muy intuitiva. El investigador decide el nombre que le quiere dar a los factores, mientras que los indicadores han de conservar el nombre de la base de datos. El factor, a la izquierda del =~ viene definido por la suma de los indicadores a la derecha del signo (p.ej. `IV =~ L+FSF+H`). Obsérvese que no se añaden los términos de error, el paquete los asume por defecto.

Las varianzas de los factores, se señalan separando el nombre del factor de sí mismo por el signo ~~ (p.ej. `IV~~IV`). Las varianzas de los errores siguen la misma terminología, con la diferencia de que los elementos separados por el ~~ son ahora indicadores, no factores (p.ej. `FSF~~FSF`). También el signo ~~ separa las covarianzas (e.g. `IV~~IQ`), pero no puede haber confusión porque lo que separa el signo son ahora factores distintos (en las varianzas a los dos lados está el mismo factor, como es lógico).

El lector ve que todo el proceso de identificación se refleja en la sintaxis (están las covarianzas, no hay covarianzas entre los errores y, al no aparecer el término de error en las ecuaciones, se asume que su coeficiente de regresión está fijado a 1). Sin embargo, ¿dónde está la escala determinada fijando a 1 la varianza de los factores? Se ha señalado del siguiente modo, aunque avanzamos que hay un mecanismo más eficiente para hacerlo cuando solicitemos al programa que estime el modelo:

Estadísticos e-Books & Papers

```
IV~~1*IV
IQ~~1*IQ
```

Identificado el CFA estamos en disposición de proceder a su estimación. Previamente señalemos algunos **indicadores que nos pueden hacer sospechar que el modelo está incorrectamente identificado** (Hair *et al.*, 2014a):

1. Errores estándar muy grandes para la estimación de algunos coeficientes.
2. Incapacidad del programa para invertir la matriz que, como veremos inmediatamente, es necesario en el proceso de estimación (se suele indicar señalando que no es definida positiva).
3. Estimaciones teóricamente imposibles como varianzas negativas y valores estandarizados de cargas y correlaciones fuera del rango $[-1, +1]$.
4. Resultados muy diferentes cuando en el proceso de iteración en la estimación se dan valores iniciales distintos a algunos parámetros (lo que es una opción en algunos paquetes estadísticos).

13.4. Estimación del análisis factorial confirmatorio

Creemos que ya hemos presentado la intuición del proceso de estimación de un CFA: dar valores a los parámetros que contiene la matriz de varianzas y covarianzas teórica Σ con el fin de que esta se parezca al máximo a la matriz de varianzas y covarianzas muestral $\mathbf{S}$. El modelo tendrá un buen ajuste cuando la diferencia entre la matriz estimada $\hat{\Sigma}$ y la matriz muestral $\mathbf{S}$ sea reducida. Obviamente el procedimiento de ir dando valores a los parámetros de Σ e ir analizando la diferencia entre las dos matrices no es eficiente y de alguna manera se ha de transformar el procedimiento generando una función susceptible de ser optimizada, pero esta función, al final, siempre contará, de algún modo, con la diferencia $\mathbf{S} - \hat{\Sigma}$ en su interior. Veámoslo de una manera algo más formal.

Presentaremos a continuación algunos de los métodos de estimación disponibles. Profundizar en todos ellos va más allá del alcance de este libro y recomendamos recurrir a Bentler (1995) para ello. Sin embargo, se ofrecerán los fundamentos básicos de cada uno de ellos y sobre todo la intuición del más utilizado, el **método de máxima verosimilitud**.

Como hemos señalado, el investigador parte de una matriz de varianzas y covarianzas muestral $\mathbf{S}$. Mientras que la matriz de varianzas y covarianzas poblacional condicionada al modelo (13.3) está relacionada con los parámetros poblacionales por la conocida expresión (13.8):

$$\Sigma = \Lambda\Phi\Lambda' + \Theta$$

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Estimar el modelo supone encontrar valores, a partir de los datos muestrales, para las matrices anteriores (que denotamos con "^") que cumplan las restricciones impuestas en el proceso de identificación y que hagan que la matriz de varianzas y covarianzas estimada mediante la expresión siguiente sea lo más parecida posible a $\mathbf{S}$:

$$\hat{\Sigma} = \hat{\Lambda}\hat{\Phi}\hat{\Lambda}' + \hat{\Theta} \tag{13.11}$$

Long (1983b) ilustra el proceso de estimación como sigue. Inicialmente existirán infinitas matrices estimadas de Σ, Φ y Θ que satisfagan la expresión anterior (13.11), pero habrá que rechazar todas aquellas soluciones que no cumplan las restricciones que se han impuesto en la identificación del modelo. Llamemos genéricamente Σ^*, Φ^* y Θ^* a las matrices que sí cumplen las restricciones. Esas matrices permiten obtener una estimación de la matriz de varianzas y covarianzas poblacional Σ^* mediante (13.11). Si esta última matriz está próxima a $\mathbf{S}$, entonces las estimaciones de los parámetros contenidas en Σ^*, Φ^* y Θ^* serían razonables en el sentido de ser consistentes con los datos de $\mathbf{S}$.

Necesitamos una función, a la que denominaremos una *función de ajuste*, que nos indique en qué medida "Σ^* está próxima a $\mathbf{S}$". Long (1983b) denota a estas funciones de ajuste con la expresión $F(\mathbf{S};\Sigma^*)$ y están definidas para todas las matrices que cumplen las restricciones marcadas en la identificación del modelo. Si entre dos matrices que cumplen esta condición se verifica que $F(\mathbf{S};\Sigma_1^*) < F(\mathbf{S};\Sigma_2^*)$, entonces concluiremos que Σ_1^* está más "próxima" a $\mathbf{S}$ que Σ_2^*. Consecuentemente, aquellos valores de Σ^*, Φ^* y Θ^* que minimizan el valor de $F(\mathbf{S};\Sigma^*)$ serán las estimaciones de los parámetros poblacionales finales $\hat{\Sigma}$, $\hat{\Phi}$ y $\hat{\Theta}$. Los procedimientos de estimación que vamos a describir a continuación son los siguientes: mínimos cuadrados no ponderados, mínimos cuadrados generalizados, máxima verosimilitud, estimación por la teoría de la distribución elíptica y estimación con libre distribución asintótica.

13.4.1. Estimación por mínimos cuadrados no ponderados

La estimación por mínimos cuadrados no ponderados ULS (*Unweighted Least Squares)* toma como estimadores a los valores que minimizan la siguiente función de ajuste:

$$F_{ULS}\left(\mathbf{S};\Sigma^*\right) = \frac{1}{2}tr\left[\left(\mathbf{S}-\Sigma^*\right)^2\right] \tag{13.12}$$

donde por tr indicamos la traza de la matriz resultante de la operación subsiguiente, esto es, la suma de los elementos de su diagonal. Long (1983b) y Ullman (2001) indican que este método tiene dos limitaciones que hacen que no sea muy utilizado. En primer lugar, no existen contrastes estadísticos asociados a este tipo de estimación y, en segundo lugar, los estimadores dependen de la escala de medida de las variables observadas, esto es, no se alcanzaría el mismo mínimo si el nivel de renta, por ejemplo, estuviera medido en pesetas que si lo estuviera en euros.

Estadísticos e-Books & Papers

Este método tiene, sin embargo, algunas ventajas. Así, no es necesario asumir ningún tipo de distribución teórica de las variables observadas frente a la hipótesis de normalidad multivariante que asumen otros métodos de estimación. Por ello, si la violación de esta hipótesis fuera muy evidente, algunos autores recomiendan recurrir a la estimación por este método, pero tomando como datos de partida la matriz de varianzas y covarianzas estandarizada —o matriz de correlaciones— para corregir el problema de la dependencia de las unidades de medida.

13.4.2. Estimación por mínimos cuadrados generalizados

La estimación por mínimos cuadrados generalizados GLS *(Generalized Least Squares)* se basa en ponderar la matriz cuya traza se calcula en (13.12) mediante la inversa de matriz de varianzas y covarianzas muestral, esto es:

$$F_{GLS}\left(\mathbf{S};\Sigma^{*}\right)=\frac{1}{2}tr\left[\left(\mathbf{S}-\Sigma^{*}\right)\mathbf{S}^{-1}\right]^{2} \tag{13.13}$$

13.4.3. Estimación por máxima verosimilitud

La estimación por máxima verosimilitud ML *(Maximum Likelihood)* implica minimizar la siguiente función de ajuste:

$$F_{ML}\left(\mathbf{S};\Sigma^{*}\right)=tr\left(\mathbf{S}\Sigma^{*-1}\right)+\left[\log\left|\Sigma^{*}\right|-\log\left|\mathbf{S}\right|\right]-q \tag{13.14}$$

donde toda la notación es conocida, pues recordemos que q es el número de variables observadas y el hecho de denotar con el símbolo $|\ |$ al determinante de la matriz de referencia. Como señala Long (1983b), cuanto más se aproximen las matrices $\mathbf{S}$ y Σ^{*}, más se aproximará el producto $\mathbf{S}\Sigma^{*-1}$ a la matriz identidad. Como la matriz identidad tiene rango $q\times q$, entonces, dado que la traza de esa matriz identidad es la suma de los q unos de la diagonal (o sea, q), el primer término de (13.14) se aproximará a q cuando las matrices estén próximas, compensándose con el término $-q$ del final de la expresión (13.14). Por otra parte, la diferencia de los logaritmos de los determinantes de $\mathbf{S}$ y Σ^{*} tenderá a 0, dado que, cuando las matrices estén próximas, también lo estarán sus determinantes. De esta forma, cuando las matrices sean iguales, la función de ajuste será cero.

13.4.4. Estimación por la teoría de la distribución elíptica

La estimación EDT *(Elliptical Distribution Theory)* se basa en la distribución de probabilidad de este nombre. La distribución normal multivariante es un caso particular de esta familia con parámetro de curtosis[2] igual a cero. En este caso, la función a minimizar adopta la forma:

[2]El coeficiente de curtosis de una distribución es igual al coeficiente estandarizado de cuarto orden menos tres.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

$$F_{EDT}\left(\mathbf{S};\Sigma^{*}\right)=\frac{1}{2}\left(\kappa+1\right)^{-1}tr\left[\left(\mathbf{S}-\Sigma^{*}\right)\mathbf{W}^{-1}\right]^{2}-\delta tr\left[\left(\mathbf{S}-\Sigma^{*}\right)\mathbf{W}^{-1}\right]^{2} \tag{13.15}$$

siendo κ y δ funciones de curtosis y $\mathbf{W}$ cualquier estimador consistente de Σ.

13.4.5. Estimación con libre distribución asintótica

La estimación ADF *(Asymptotically Distribution Free)* minimiza una función definida mediante la siguiente expresión:

$$F_{ADF}\left(\mathbf{S};\Sigma^{*}\right)=\left[\mathbf{s}-\sigma\left(\Theta\right)\right]'\mathbf{W}^{-1}\left[\mathbf{s}-\sigma\left(\Theta\right)\right] \tag{13.16}$$

donde $\mathbf{s}$ es el vector de datos, es decir, la matriz de varianzas y covarianzas muestrales pero escrita en forma de un solo vector; σ es la matriz de varianzas y covarianzas estimada, de nuevo puesta en forma de vector y donde con el término (Θ) se ha querido indicar que se deriva de los parámetros del modelo (coeficientes de regresión, varianzas y covarianzas). $\mathbf{W}$ es una matriz que pondera las diferencias cuadráticas entre las matrices de varianzas y covarianzas muestrales y estimadas. En este caso, cada elemento de esa matriz se obtiene:

$$w_{ijkl}=\sigma_{ijkl}-\sigma_{ij}\sigma_{kl}$$

siendo σ_{ijkl} momentos de cuarto orden y σ_{ij} y σ_{kl} las covarianzas.

13.4.6. Comparación de los distintos procedimientos de estimación

Resumimos a continuación los resultados del trabajo de Hu *et al.* (1992), que analizó mediante simulación de Montecarlo cómo se comportaban los distintos procedimientos de estimación ante distintos tamaños muestrales, violación de las hipótesis de normalidad y de independencia entre los términos de error y los factores comunes.

Estos autores encontraron que, en caso de que fuera razonable asumir la normalidad, el método ML funcionaba mejor cuando el tamaño muestral era superior a 500 casos, mientras que para tamaños inferiores a esa cifra tenía un mejor comportamiento el método EDT. Finalmente, el método ADF solo ofrecía buenos resultados con muestras superiores a 2500 casos.

Cuando el supuesto de normalidad se violaba, los métodos de ML y GLS solo daban buenos resultados con muestras superiores a 2500 casos, aunque el GLS funcionaba algo mejor que el ML en muestras inferiores. Pese a no adoptar el supuesto de normalidad, el método ADF tampoco daba buenos resultados con muestras inferiores a 2500 casos.

Cuando se produce una violación del supuesto de independencia entre los términos de error y los factores comunes, los métodos de ML y GLS funcionan

Estadísticos e-Books & Papers

muy mal, y también el ADF salvo que la muestra fuera superior a 2500 casos. En cambio, el EDT funcionaba significativamente mejor que los demás.

A la luz de lo expuesto, Ullman (2001) recomienda:

- Los métodos de ML y GLS son la mejor opción con pequeñas muestras siempre que sea plausible la asunción de normalidad e independencia.
- En el caso en que ambos supuestos no parezcan razonables, se recomienda recurrir a la estimación ML denominada "escalada" o "robusta". Una descripción de este procedimiento se encuentra en Satorra y Bentler (1988) y es una opción de estimación en la mayoría de paquetes estadísticos. Su lógica es bastante intuitiva. ¿Para qué es necesaria la normalidad?, no para estimar los parámetros, sino para que los errores estándar sean fiables y el estadístico t que utilizamos para calcular la significatividad de los parámetros sea confiable. Pues entonces lo que hacen estos autores es corregir el cálculo de los estadísticos para que sean robustos ante problemas de ausencia de normalidad en lugar de cambiar la forma de estimar los parámetros.

Veamos, a modo de ilustración, el resultado de estimar mediante máxima verosimilitud el modelo del caso 13.1 que estamos siguiendo como ejemplo y poder familiarizarnos así también con la presentación de la información que nos dan las salidas de **lavaan**.

El primer paso es indicarle mediante la sintaxis que estime el modelo que hemos definido con anterioridad. Basta añadir a la sintaxis anterior que hemos mostrado la siguiente instrucción:

```
#Estimacion del modelo

fit <- lavaan(modelo.cfa, sample.cov=datos.cov, sample.nobs=275,
std.lv=TRUE, mimic="eqs",
estimator="ML", verbose=TRUE, warn=TRUE)

#Peticion de elementos en la salida

summary (fit, fit.measures=TRUE,
standardized=TRUE, rsquare=TRUE)
resid(fit, type="cor")
```

Esta sintaxis, como vemos, tiene dos partes. En primer lugar, se crea un objeto de R, que hemos llamado **fit**, que va a recoger toda la información de la estimación. La instrucción **lavaan()** pide que se estime el modelo y su estructura es la siguiente: primero se le indica qué modelo es ese —el que definimos con anterioridad y que llamamos **modelo.cfa**—, cuáles son los datos

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

que ha de utilizar —la matriz `datos.cov`, solo que, al no ser *raw data*, le hemos de advertir que le damos una matriz de covarianzas, por eso señalamos `sample.cov=datos.cov`, cuando, si fueran *raw data*, le indicaríamos simplemente `data=fichero de datos`—. Como es una matriz y el programa no puede saber cuántos datos la han generado, se lo indicamos —`sample.nobs=275`, lo que es innecesario con *raw data*—, le señalamos (es una opción personal porque era el *software* al que estaba acostumbrado) que utilice los algoritmos que permitan replicar los resultados que obtendríamos con EQS `---mimic="eqs"`, opción que es prescindible—. También le pedimos que estime por máxima verosimilitud —`estimator="ML"`, en caso de que quisiéramos estimación robusta lo indicaríamos como `estimator="MLM"`—, `verbose=TRUE` nos ofrece el historial de iteraciones que nos permite ver la rapidez de la convergencia y, finalmente, `warn=TRUE` nos avisa si ha habido algún problema en el proceso de estimación.

Llamo la atención sobre la opción `std.lv=TRUE`. He señalado con anterioridad que la identificación de escala la habíamos realizado fijando a 1 la varianza de cada factor e incorporándolo a la sintaxis del siguiente modo: `IV~~1*IV y IQ~~1*IQ`. He señalado que había otra forma más eficiente de hacerlo, esta forma es precisamente con el argumento `std.lv=TRUE` que le dice a `lavaan` que identifique la escala fijando la escala del factor. Es decir, con este modificador, la sintaxis `IV~~1*IV y IQ~~1*IQ` es innecesaria y podría haberse dejado `IV~~IV y IQ~~IQ`.

La segunda parte de la sintaxis le pide a `lavaan` que la información que contiene el objeto `fit` nos la muestre con algunas personalizaciones: `fit.measures = TRUE` pide que se amplíe el número de indicadores de ajuste que se muestran por defecto y que veremos en una sección posterior, `standardized=TRUE` pide que junto con las estimaciones no estandarizadas de los parámetros ofrezca también las estandarizadas, `rsquare=TRUE` pide que para la ecuación de cada indicador nos indique el porcentaje de varianza que explica el factor del mismo (que no es otra cosa que la carga al cuadrado, como veremos en el tema siguiente para ver su utilidad como indicador de la validez convergente) y `resid(fit, type="cor")` solicita que se muestren las diferencias para cada elemento de la matriz de varianzas y covarianzas entre el valor muestral y el estimado. Su uso como indicador de la calidad del ajuste lo ilustraremos también posteriormente, solo que estandarizados (correlaciones, no covarianzas).

Para familiarizarnos con las salidas de `lavaan` comentaremos las mismas conectándolas con lo visto hasta este momento: identificación y estimación de parámetros. La parte correspondiente a los indicadores que nos mostrarán si el ajuste es razonable lo veremos después de presentar esos indicadores y, por supuesto, dejaremos para el tema siguiente el uso que se ha de hacer de la información del CFA para validar el instrumento de medida.

La primera información importante contiene la matriz con las cargas estimadas que habíamos denotado como $\hat{\Lambda}$ y que aparece en el cuadro 13.4. El cuadro nos muestra la estimación no estandarizada de las cargas (`Estimate`), el error estándar de la estimación y, derivado de ello, el valor del estadístico t, que etiqueta como `z-value` ($t = \hat{\lambda}/SE$), a continuación ofrece la significatividad de

Estadísticos e-Books & Papers

Cuadro 13.4.: Estimación de las cargas factoriales

```
Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  IV =~
    L                 0.797    0.073   10.882    0.000    0.797    0.731
    FSF               0.406    0.039   10.309    0.000    0.406    0.689
    H                 0.481    0.066    7.327    0.000    0.481    0.491
  IQ =~
    M                 0.835    0.067   12.488    0.000    0.835    0.759
    FSC               0.312    0.025   12.501    0.000    0.312    0.760
    Q                 0.682    0.069    9.921    0.000    0.682    0.615
```

ese parámetro o `P(>|z|)` y el valor estandarizado de la carga `Std.all`.

Con esta información, la matriz $\hat{\Lambda}$ sería la siguiente:

$$\hat{\Lambda} = \begin{bmatrix} 0{,}797 & 0 \\ 0{,}406 & 0 \\ 0{,}481 & 0 \\ 0 & 0{,}835 \\ 0 & 0{,}312 \\ 0 & 0{,}682 \end{bmatrix}$$

Recordemos que también se ha estimado la matriz $\hat{\Phi}$ que contiene las varianzas de los factores —aunque se fijaron a 1 en el proceso de identificación— y la covarianza —correlación al estar fijada a 1 la varianza— entre los dos factores. El cuadro 13.5 nos ofrece la salida de `lavaan` con esta información. La estructura de presentación es la misma y no la repetiremos, solo llamamos la atención sobre el hecho de que, al estar fijadas a 1 las varianzas de IV e IQ, esto es lo que se refleja en su estimación no estandarizada y, por estar fijada, no va acompañada del valor del estadístico t ni la significatividad. También llamamos la atención sobre que, al estar fijada a 1 la varianza, el valor de la covarianza entre los factores (estimación no estandarizada), 0,583, coincide con la estandarizada (correlación). Los elementos etiquetados como .L .FSF .H .M .FSC y .Q son las varianzas de los términos de error de los indicadores. Obsérvese que no hay ninguna estimación de las correlaciones entre ellos debido a la identificación.

Por lo tanto, las matrices $\hat{\Phi}$ y la matriz $\hat{\Theta}$ quedarán como sigue:

$$\hat{\Phi} = \begin{bmatrix} 1{,}000 & 0{,}583 \\ 0{,}583 & 1{,}000 \end{bmatrix} \quad \hat{\Theta} = \begin{bmatrix} 0{,}553 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0{,}183 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0{,}729 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0{,}512 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0{,}071 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0{,}767 \end{bmatrix}$$

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 13.5.: Estimación de las varianzas y covarianzas de los factores

```
Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  IV ~~
    IQ                0.583    0.064    9.120    0.000    0.583    0.583

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
    IV                1.000                               1.000    1.000
    IQ                1.000                               1.000    1.000
   .L                 0.553    0.088    6.264    0.000    0.553    0.465
   .FSF               0.183    0.025    7.287    0.000    0.183    0.526
   .H                 0.729    0.071   10.283    0.000    0.729    0.759
   .M                 0.512    0.075    6.828    0.000    0.512    0.423
   .FSC               0.071    0.010    6.810    0.000    0.071    0.422
   .Q                 0.767    0.079    9.655    0.000    0.767    0.622
```

Cuadro 13.6.: Matriz de varianzas covarianzas estimada

```
$cov
    L     FSF   H     M     FSC   Q
L   1.188
FSF 0.324 0.348
H   0.384 0.196 0.960
M   0.388 0.198 0.234 1.210
FSC 0.145 0.074 0.087 0.260 0.168
Q   0.317 0.162 0.191 0.570 0.213 1.232
```

Bastaría realizar las operaciones de cálculo matricial que refleja la ecuación (13.11) para obtener la matriz de varianzas y covarianzas estimada $\hat{\Sigma}$, que también nos ofrece `lavaan`, dejando al lector la comprobación de que se corresponde con las operaciones matriciales señaladas (cuadro 13.6).

13.5. Bondad de ajuste del modelo estimado

Antes de pasar a interpretar los resultados del análisis factorial confirmatorio que se ha efectuado, es necesario determinar hasta qué punto el modelo asumido se ajusta a los datos muestrales. Si detectáramos problemas de ajuste, sería necesario plantear algún tipo de reespecificación del mismo hasta que se lograra un mejor ajuste o, siendo más estrictos, en la medida en que estamos ante un planteamiento confirmatorio, descartar nuestro modelo teórico. Analizaremos, a continuación, una serie de criterios que se calculan en la mayor parte de programas que abordan este tema. Como ya avanzamos, los estadísticos elaborados con esta finalidad son muchos más de los que aquí se muestran. La selección efectuada recoge, desde nuestro punto de vista, los más utilizados y, al finalizar la sección, se apuntarán los resultados de algunas investigaciones que analizan el desempeño relativo de cada uno de ellos.

Estadísticos e-Books & Papers

Cuadro 13.7.: Cálculo de la matriz de covarianzas residual

```
Matriz de covarianzas muestral (S)

    L     FSF   H     M     FSC   Q
L   1.188
FSF 0.317 0.348
H   0.428 0.182 0.960
M   0.333 0.225 0.158 1.210
FSC 0.142 0.077 0.074 0.265 0.168
Q   0.344 0.214 0.195 0.565 0.206 1.232

Matriz de covarianzas estimada (Sigma hat)

    L     FSF   H     M     FSC   Q
L   1.188
FSF 0.324 0.348
H   0.384 0.196 0.960
M   0.388 0.198 0.234 1.210
FSC 0.145 0.074 0.087 0.260 0.168
Q   0.317 0.162 0.191 0.570 0.213 1.232

Matriz de covarianzas residual (S-Sigma hat)

    L      FSF    H      M      FSC    Q
L    0.000
FSF -0.007  0.000
H    0.045 -0.014  0.000
M   -0.055  0.027 -0.076  0.000
FSC -0.003  0.003 -0.014  0.004  0.000
Q    0.027  0.053  0.003 -0.005 -0.006  0.000
```

13.5.1. Matriz residual de covarianzas

Como indicábamos al presentar los distintos métodos de estimación del CFA, el objetivo básico de los mismos es que la matriz de covarianzas poblacional estimada se parezca lo más posible a la muestral **S**. En otros términos, puede expresarse lo anterior diciendo que la diferencia entre ambas matrices, a la que llamamos matriz residual de covarianzas, esté lo más cercana posible a una matriz nula **0**. Los valores de esta matriz deberían ser pequeños y estar homogéneamente distribuidos. Como señala Byrne (2006), residuos grandes asociados a algunos parámetros podrían indicar que han sido mal especificados, y ello afectaría negativamente al ajuste global del modelo. En nuestro ejemplo, la matriz de covarianzas muestral se introdujo directamente como dato, la matriz estimada se ha mostrado en el cuadro 13.6, por lo que la matriz residual es la diferencia entre ambas, como se muestra en el cuadro 13.7.

Lo habitual, a efectos de interpretación y para facilitarla, es que se le pida

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 13.8.: Matriz de correlaciones residual

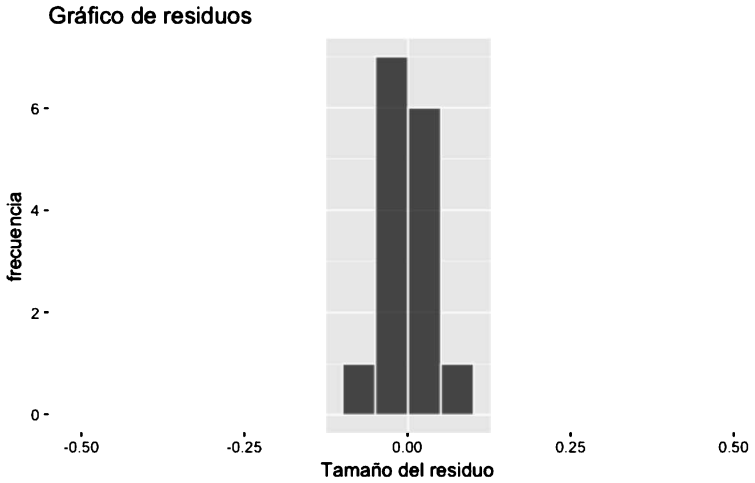
```
     L      FSF    H      M      FSC    Q
L     0.000
FSF  -0.011  0.000
H     0.042 -0.024  0.000
M    -0.046  0.042 -0.070  0.000
FSC  -0.007  0.013 -0.035  0.010  0.000
Q     0.022  0.080  0.003 -0.004 -0.014  0.000
```

al programa que nos ofrezca la matriz de covarianzas residual estandarizada, esto es, la matriz de correlaciones residual, en la medida en que nos resulta más fácil interpretar los residuos de una correlación al ser esta un parámetro acotado en un valor absoluto de 1. Esta solicitud la realizábamos en la última sintaxis mostrada mediante el modificador `resid(fit, type="cor")` y el cuadro 13.8 muestra el resultado. Por definición, la diagonal siempre será nula en la medida en que las varianzas se han ajustado de manera perfecta, por lo que los indicadores que se basen en los residuos se centrarán en el triángulo inferior. Una primera observación nos debería permitir detectar residuos grandes —digamos superiores a 0,08— como fuente de problemas de ajuste, pero profundizaremos posteriormente en este tema al presentar los indicadores concretos. Tanta importancia tienen los residuos que algunos programas como `EQS` generan un histograma con los mismos de tal forma que pueda apreciarse gráficamente su tamaño. Como en el eje horizontal se muestra el tamaño centrado en cero, residuos en los extremos izquierdo y derecho denotan residuos especialmente grandes. `lavaan` no ofrece directamente el gráfico, pero el lenguaje de programación de `R` permite obtenerlo muy fácilmente. No entraremos en el detalle de la sintaxis, pero referimos al lector a consultar la documentación del paquete `ggplot2`, aunque lo que se solicita es muy intuitivo. La figura 13.7 nos muestra el histrograma. La barra central señala que la mayoría son residuos alrededor de cero y que los residuos grandes (hacia la izquierda o derecha), son pocos.

```
a<-resid(fit,type="cor")$cor
b<-c(a[lower.tri(a)])
qplot(b,
      geom="histogram",
      binwidth = 0.01,
      main = "Grafico de residuos",
      xlab = "Tamano del residuo",
      ylab = "frecuencia",
      breaks=seq(-0.5,0.5,by=0.05),
      fill=I("black"),
```


Estadísticos e-Books & Papers

Figura 13.7.: Histograma de residuos




```
col=I("white"),
alpha=I(.7),
xlim=c(-0.5,0.5))
```

13.5.2. Estadístico χ^2

Es el único estadístico para evaluar el ajuste del modelo, el resto de indicadores que comentaremos no siguen distribuciones conocidas y, por lo tanto, no pueden usarse con planteamientos inferenciales para contrastar hipótesis nulas tal como sí ocurre con el estadístico χ^2. Recordemos que la expresión de la función de máxima verosimilitud (13.14) que utilizábamos para estimar el modelo era, básicamente, una diferencia entre la matriz de varianzas y covarianzas estimada $\hat{\Sigma}$ y la muestral $\mathbf{S}$. Cuando el ajuste es perfecto, la función de máxima verosimilitud será nula porque nula será la diferencia entre las matrices. Pues bien, el estadístico χ^2 se define como la función de máxima verosimilitud multiplicada por el tamaño muestral menos uno:

$$\chi^2 = (N-1)\, F_{ML} \tag{13.17}$$

Este estadístico se distribuye como una χ^2 con tantos grados de libertad como diferencia hay entre el número de varianzas y covarianzas muestrales distintas $q(q+1)/2$ y el número de parámetros a estimar. La hipótesis nula es que la matriz de varianzas y covarianzas muestral y la teórica son iguales (lo que implicaría ajuste perfecto):

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

$$H_0 : \hat{\Sigma}^* = \mathbf{S}$$

Muchos son los inconvenientes que tradicionalmente se asocian a la χ^2 que hace que nunca se reporte sola sino acompañada por los indicadores que veremos a continuación. Entre estos inconvenientes o limitaciones cabe destacar:

1. En muchas circunstancias —tamaños muestrales reducidos o ausencia de normalidad— el estadístico no se distribuye como una χ^2 comprometiendo el resultado del contraste de la hipótesis nula (Brown, 2006). Es especialmente sensible a la curtosis excesiva (Marsh *et al.*, 1988).

2. Basta observar la expresión (13.17) para ver que incluso con valores de la función de máxima verosimilitud muy pequeños, es decir, con matrices muestrales y de varianzas y covarianzas prácticamente iguales, al estar multiplicado por el tamaño muestral cuando éste crece, también crece el valor de la χ^2 incrementando la probabilidad de rechazar la hipótesis nula pese a haber obtenido matrices prácticamente idénticas (Hair *et al.*, 2014a).

3. Se basa en una hipótesis nula excesivamente rigurosa $\hat{\Sigma}^* = \mathbf{S}$, el ajuste puede ser razonable sin necesidad de que las matrices sean necesariamente idénticas (Brown, 2006).

En un intento seminal por resolver los problemas que se han manifestado para la χ^2, Wheaton *et al.* (1977) introducen la **ratio** χ^2/df**,** que ha obtenido una gran difusión en la investigación aplicada intentando que corrija la tendencia a rechazar la hipótesis nula de la χ^2, **recomendándose relaciones 3:1 o inferiores** (en nuestro ejemplo la ratio, que `lavaan` no proporciona directamente pero es fácil de obtener con la información del cuadro 13.10, es 1,105). Sin embargo, el propio Wheaton (1987) desaconseja su uso porque es igual de sensible que la χ^2 al problema del tamaño muestral, el estándar de comparación (3:1) varía según el autor que lo proponga y no ha superado los estándares mínimos en los escrutinios mediante simulación de Montecarlo (Hu y Bentler, 1998, 1999).

El cuadro 13.9 nos ofrece la salida de `lavaan` donde aparece el estadístico χ^2. Varias cosas deben destacarse. En primer lugar, el lector puede comprobar como el número de grados de libertad (8) coincide con el cálculo manual que hicimos en el apartado 13.3 al identificar al modelo. En segundo lugar que el nivel de significatividad ($p = 0{,}356$) no nos permite rechazar la hipótesis nula de igualdad entre las matrices y, por lo tanto, confirma el buen ajuste del modelo. Conviene destacar este hecho porque a muchos investigadores noveles les confunde el hecho de que lo que quepa esperar son significatividades superiores al nivel crítico y no inferiores a él.

Ante estas limitaciones del estadístico χ^2 se han desarrollado multitud de indicadores de ajuste que han de ser comparados con valores de referencia en

Estadísticos e-Books & Papers

Cuadro 13.9.: Estadístico χ^2

```
Number of observations                          275

Estimator                                        ML
Minimum Function Test Statistic               8.842
Degrees of freedom                                8
P-value (Chi-square)                          0.356
```

función de los cuales se considera que el ajuste es razonable o no. Aquí solo vamos presentar una selección de ellos, concretamente aquellos para los que distintos trabajos con simulación de Montecarlo (Browne y Cudek, 1993; Hu y Bentler, 1998; Marsh *et al.*, 1988; Tanaka, 1987) han constatado un desempeño superior. Seguiremos la tipología de Brown (2006) que los clasifica en indicadores de ajuste absoluto, corregidos por parsimonia, y de ajuste comparativo o incremental, aunque, como señala el autor, esta clasificación no es perfecta en la medida en que algunos de ellos presentan propiedades que permitirían clasificarlos en varios de estos tipos.

13.5.3. *Standardized Root Mean Residual* (SRMR)

El SRMR es un indicador de **ajuste absoluto**, en la medida en que evalúa la plausibilidad de que las matrices de varianzas y covarianzas muestral y estimada sean la misma. En este sentido, el estadístico χ^2 sería, también un indicador de ajuste absoluto. Su construcción es bastante intuitiva. Si la matriz de varianzas y covarianzas residual (diferencia entre la muestral y la estimada) recoge lo alejadas que están las dos matrices que serían idénticas en el caso de ajuste perfecto, basta con calcular una media de los residuos para tener un buen indicador del promedio del desajuste. Si antes de calcular la media, elevamos al cuadrado las diferencias para evitar la compensación de signos positivos y negativos y luego calculamos la raíz cuadrada de la media para reajustar la escala, eso es el SRMR. Cuanto más pequeño sea el valor del SRMR, mejor será el ajuste. Hu y Bentler (1999) proponen que **valores inferiores a 0,08 denotan un buen ajuste**.

En el cuadro 13.8 tenemos la matriz de correlaciones residual (el término "estandarizado" del indicador hace referencia a que se calcula sobre la matriz residual de correlaciones). Si elevamos al cuadrado todos los residuos, calculamos la media de estos cuadrados y extraemos la raíz cuadrada, el resultado que dejamos al lector debería ser 0,031 que coincide con el valor ofrecido por `lavaan` en el cuadro 13.9.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 13.10.: Indicadores de ajuste del modelo

```
Model test baseline model:

  Minimum Function Test Statistic              392.818
  Degrees of freedom                                15
  P-value                                        0.000

User model versus baseline model:

  Comparative Fit Index (CFI)                    0.998
  Tucker-Lewis Index (TLI)                       0.996

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)              -1831.324
  Loglikelihood unrestricted model (H1)      -1826.887

  Number of free parameters                         13
  Akaike (AIC)                                3688.648
  Bayesian (BIC)                              3735.618
  Sample-size adjusted Bayesian (BIC)         3694.398

Root Mean Square Error of Approximation:

  RMSEA                                          0.020
  90 Percent Confidence Interval          0.000  0.075
  P-value RMSEA <= 0.05                          0.758

Standardized Root Mean Square Residual:

  SRMR                                           0.031
```


Estadísticos e-Books & Papers

13.5.4. *Root Mean Square Error of Approximation* (RMSEA)

El RMSEA entraría dentro del grupo de los denominados **indicadores corregidos por parsimonia.** Los indicadores de esta categoría son parecidos en algunos casos al χ^2 y al SRMR pero incorporan en su cálculo una penalización por poca parsimonia. Intentemos explicar este concepto. Imaginemos dos modelos con los mismos indicadores y con la misma χ^2. El ajuste de ambos es el mismo, sin embargo imaginemos que uno de ellos es mucho más complejo, incorpora muchas más relaciones que estimar. Al ser el número de indicadores el mismo, que haya más parámetros que estimar implica que tendrá menos grados de libertad disponibles. Los grados de libertad se convierten así en un indicador de parsimonia, menos grados de libertad con los mismos datos son modelos más complejos. Si aplicamos el principio de la navaja de Ockham o *lex parsimoniae*, que dice que, en igualdad de condiciones, ante dos explicaciones alternativas de una misma realidad, siempre será preferible la más sencilla, nos quedaríamos con el modelo con más grados de libertad. Pues bien, los indicadores que vamos a ver introducen los grados de libertad como factor de corrección, haciendo que, ante una misma χ^2, el indicador señale un mejor ajuste cuando el número de grados de libertad sea superior.

El RMSEA propuesto por Steiger y Lind (1980) corrige la χ^2 dividiendo por el tamaño muestral y también por el número de grados de libertad (*df*) como sigue:

$$RMSEA = \sqrt{\frac{\chi^2 - df}{df\,(N-1)}} \tag{13.18}$$

Véase la lógica. Si la χ^2, al estar multiplicada por el tamaño muestral, se hace grande cuando este es grande, pues dividamos por él. Además fíjese que los grados de libertad (más grados de libertad, más parsimonia) hacen caer el valor de la χ^2 (mejor ajuste) de dos formas: restando en el numerador y dividiendo en el denominador. El *benchmark* que Browne y Cudek (1993) ofrecen para este indicador es **buen ajuste (RMSEA<0,05), ajuste aceptable (0,05<RMSEA<0,08), ajuste pobre si RMSEA>0,08**. Una ventaja de este indicador es que se puede construir a su alrededor un intervalo de confianza, lo que hace ganar en precisión a la hora de presentar los resultados. También Browne y Cudek (1993) han desarrollado un test estadístico para evaluar la cercanía (*approximation*) del modelo a los datos utilizando como referencia el nivel de 0,05 para el RMSEA. Así en el cuadro 13.10 puede comprobarse como no solo el RMSEA es inferior a 0,05 (es 0,02) sino que el mencionado test no puede descartar la hipótesis nula de que el RMSEA es inferior a 0,05 (p=0,758).

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

13.5.5. *Tucker-Lewis Index* (TLI)

El TLI, propuesto por Tucker y Lewis (1973), entra dentro de los denominados índices de ajuste comparativo o incremental. Su lógica reside en que evalúa el ajuste del modelo mediante la χ^2 comparándola con qué χ^2 se obtendría en un modelo naíf en el que cada indicador formará un único factor y todas las covarianzas entre ellos serán nulas. Si los indicadores realmente se agrupan en factores, el ajuste de este modelo base o ingenuo debería ser muy malo. El TLI lo que evalúa es cuánto mejor es el modelo propuesto respecto a ese modelo ingenuo. Quizás la mejor forma de entenderlo sea empezar por presentar, aunque no lo ofrece `lavaan`, el indicador NFI de Bentler y Bonett (1980), que ilustra bien esta idea. Si llamamos χ^2_B al valor del estadístico para el modelo ingenuo y llamamos χ^2_M al valor para nuestro modelo, el NFI adopta la expresión:

$$NFI = \frac{\chi^2_B - \chi^2_M}{\chi^2_B}$$

Si nuestro modelo tiene un ajuste perfecto, entonces $\chi^2_M = 0$, y el NFI sería la ratio χ^2_B/χ^2_B y $NFI = 1$. Por su parte, si nuestro modelo fuera tan malo como el ingenuo, entonces $\chi^2_B = \chi^2_M$ y el numerador de la expresión que calcula el NFI sería 0 y $NFI = 0$. De una manera muy sencilla hemos conseguido un indicador que toma el valor 0 en una situación de mal ajuste y vale 1 en una situación de buen ajuste. El problema del NFI, entre otros, es que vemos que no corrige por parsimonia, por eso Tucker y Lewis (1973) proponen el TLI, que se define de este modo (en algunos programas, como EQS, este indicador aparece como Bentler-Bonett Non-Normed Fit Index o NNFI):

$$TLI = \frac{\chi^2_B - \frac{df_B}{df_M}\chi^2_M}{\chi^2_B - df_B} \tag{13.19}$$

Obsérvese que el TLI tiene la estructura base del NFI pero incorpora un factor de corrección por parsimonia a través de los grados de libertad. Tiene, sin embargo, un problema. Imaginemos que nuestro modelo tiene un ajuste perfecto, por lo que $\chi^2_M = 0$. Entonces, la expresión (13.19) está formada por un numerador χ^2_B que es mayor que el denominador, puesto que a χ^2_B se le restan los grados de libertad df_B. Es decir, en caso de ajuste perfecto, el TLI puede tomar valores superiores a la unidad, de ahí el término *Non Normed Fit Index* con que se refieren a él en algunos programas. Un modelo tendrá un **buen ajuste cuando su TLI >0,90** (Schumacker y Lomax, 1996). `lavaan` ofrece los valores de la χ^2 del modelo base (cuadro 13.10) bajo la etiqueta *model test baseline model.* Es inmediato con esta información derivar el resultado del TLI que aparece en ese cuadro:

$$TLI = \frac{382{,}818 - \frac{15}{8}8{,}842}{382{,}818 - 15} = 0{,}996$$

Estadísticos e-Books & Papers

13.5.6. *Comparative Fit Index* (CFI)

Propuesto por Bentler (1990) lo que hace es mantener el principio de los indicadores incrementales como el NFI, incorporar la corrección por parsimonia como el TLI pero evitar que pueda tomar valores superiores a la unidad. Se obtiene del siguiente modo:

$$CFI = 1 - \frac{\chi^2_M - df_M}{\chi^2_B - df_B} \tag{13.20}$$

Es uno de los indicadores más usados porque es bastante robusto ante crecimientos de la complejidad del modelo (Hair *et al.*, 2014a). Los valores de referencia para este indicador, de acuerdo con Hu y Bentler (1999) serían: **[0,90-0,95] aceptable, >0,95 bueno**. El cálculo a partir de los datos del cuadro 13.10 también es inmediato:

$$CFI = 1 - \frac{8{,}842 - 8}{392{,}818 - 15} = 0{,}998$$

13.6. Interpretación del modelo

Hasta este momento nos hemos centrado en analizar la razonabilidad del modelo en términos globales (su ajuste). Ahora vamos a examinar si los estimadores de los parámetros son también razonables en dos sentidos: (i) ¿toman valores adecuados teóricamente?, y (ii) ¿son significativos?

La mayor parte de la información necesaria para esta fase ya se ha mostrado en los cuadros 13.4 y 13.5 y a ellos referiremos nuestros comentarios.

En primer lugar, vamos a analizar si los valores que toman los parámetros estimados son o no compatibles con el modelo estadístico. Para que exista tal compatibilidad las respuestas a las siguientes preguntas deben ser en todos los casos negativas:

- ¿Existen correlaciones superiores a la unidad?
- ¿Existen cargas factoriales estandarizadas fuera del intervalo –1,+1?
- ¿Son los errores estándar anormalmente grandes o pequeños?
- ¿Hay estimaciones negativas de las varianzas?

Si hubiera respuestas no negativas, y aunque el ajuste global del modelo fuera óptimo, estaríamos ante un indicador claro (Long, 1983b) de que esta incompatibilidad puede haberse originado por uno o más de los siguientes motivos:

- El modelo está mal especificado.
- Los datos no respaldan la hipótesis de normalidad multivariante de las variables observadas.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 13.11.: Indicadores de ajuste del modelo

```
R-Square:
                    Estimate
    L                  0.535
    FSF                0.474
    H                  0.241
    M                  0.577
    FSC                0.578
    Q                  0.378
```

- La muestra es demasiado pequeña.
- El modelo está demasiado cerca de no estar identificado, lo que hace la estimación de algunos parámetros difícil o inestable.
- Los valores perdidos de algunas variables observadas han provocado que cada elemento de la matriz de covarianzas muestral esté calculado sobre una muestra diferente.

Si se revisan los cuadros 13.4 y 13.5 se puede comprobar que, en el modelo del ejemplo, no se presenta ninguna de las incompatibilidades señaladas.

La segunda cuestión que debemos examinar es la significatividad estadística de cada parámetro individual. Centraremos la explicación en los coeficientes de regresión entre variables observadas y factores comunes, aunque lo expuesto es válido para el resto de parámetros (varianzas y covarianzas).

Si tomamos, por ejemplo, la primera línea del cuadro 13.4, comprobamos que nos informa de que la inteligencia verbal (IV) está influyendo de manera significativa ($\lambda_{estandarizado} = 0{,}731; t = 10{,}882; p < 0{,}01$) en el primer indicador, la nota en lengua (L). Como ya apuntamos en su momento, en primer lugar ofrece la estimación no estandarizada de la carga bajo el nombre de `Estimate`, que sirve para calcular el valor t (`z-value`) mediante la expresión $t = \lambda_{no\,estandarizado}/SE$, donde el error estándar aparece con la etiqueta `Std.Err`.

La parte de la varianza del indicador explicada por el factor es el cuadrado de la carga estandarizada, esto es, $\lambda^2 = 0{,}731^2 = 0{,}535$, por lo que la parte no explicada o varianza del término de error sería $1 - 0{,}534 = 0{,}466$. Esta información, la varianza explicada por el factor, es la misma que proporciona la R^2 de cada indicador que sí que solicitamos en su momento en la sintaxis y que mostramos en el cuadro 13.11.

No profundizamos más en la interpretación de los resultados porque, como ya señalamos al principio del tema, la información que proporciona el CFA sirve fundamentalmente para evaluar el instrumento de medida y, para ello, además de la información que hemos mostrado en las salidas, hace falta una serie de criterios que aplicar a esa información, y ese será el objetivo del tema siguiente.

Estadísticos e-Books & Papers

13.7. Reespecificación del modelo

Como señala Ullman (2001), existen básicamente dos motivos para reespecificar un modelo (esto es, eliminar o introducir relaciones entre las variables que los conforman): (i) mejorar su ajuste o (ii) contrastar alguna hipótesis teórica. Existen, sin embargo, muchos problemas que pueden generarse como consecuencia de una reespecificación poco meditada. Como veremos a continuación, existe un instrumento analítico proporcionado por los programas, los **índices de modificación**, que nos indica qué relaciones causales pueden añadirse y qué mejoras en el ajuste obtendríamos con cada una de esta modificaciones. Si el investigador cae en la tentación de ir incorporando o eliminando relaciones sin más, hasta lograr un ajuste razonable y no tiene en cuenta si estas modificaciones están o no soportadas por el marco teórico que sustenta su investigación, puede acabar provocando que el modelo al que se llega no sea en absoluto generalizable porque se ha capitalizado la asociación causal de los datos que emanan del error muestral y que no se obtendrían en réplicas con nuevas muestras independientes (McCallum *et al.*, 1982).

En este mismo sentido, Pedhazur (1982) y Sorbom (1989) afirman que es científicamente incorrecto modificar un modelo simplemente porque mejore su ajuste, ya que el cambio debe ser teóricamente interpretable y el investigador debe ser capaz de justificar cuál es el motivo para añadir una relación causal determinada.

Todo lo expuesto lleva a Hatcher (1994) a plantear las siguientes recomendaciones para la modificación de un modelo, aunque la mayoría se basan en el trabajo de McCallum *et al.* (1982):

1. Utilizar muestras grandes. Los modelos basados en menos de 100 o 150 casos llevan a modelos finales poco estables si las modificaciones se basan en los datos y no en la teoría.

2. Hacer pocas modificaciones. Es posible que las primeras modificaciones puedan estar derivadas de un modelo que refleje las relaciones poblacionales; las siguientes, probablemente, reflejarán relaciones específicas de la muestra.

3. Realizar solo aquellos cambios que puedan ser interpretados desde una perspectiva teórica o tengan soporte en trabajos precedentes. En todo caso, se deben detallar todos los cambios realizados sobre el modelo inicial en el informe del trabajo final.

4. Seguir un procedimiento paralelo de especificación. Siempre que sea posible, el investigador debería trabajar con dos muestras independientes. Si las dos muestras desembocan en las mismas modificaciones del modelo, se podrá tener una mayor confianza en la estabilidad del mismo.

5. Comparar modelos alternativos desde el principio. Más que proponer un modelo e ir modificándolo, puede ser conveniente en algunas ocasiones

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

plantear modelos alternativos y determinar con cuál se obtiene un mejor ajuste.

6. Finalmente, describir detalladamente las limitaciones de su estudio. Como indica Hatcher (1994), la mayoría de los trabajos que se publican están basados en una única muestra y sobre los que se efectúan sucesivas modificaciones basadas en los datos hasta lograr un ajuste razonable. Si se sigue este enfoque, sería recomendable que el trabajo advirtiera al lector de todas estas circunstancias.

Una vez planteadas estas precauciones, veamos a continuación los instrumentos de los que se dispone para reespecificar un modelo.

La información que la mayoría de programas proporcionan para reespecificar el modelo son los llamados **índices de modificación [MI]** (aunque algunos programas como EQS los denominan multiplicadores de Lagrange). Los índices de modificación se calculan para cada parámetro que no se ha estimado, dicho de manera algo más estricta, que se ha restringido fijándolo a cero. Por ejemplo, en nuestro modelo, podríamos haber hecho que el indicador FSF también fuera un indicador de la inteligencia cuantitativa (IQ), pero no lo hemos planteado, es decir, esa carga está fijada a cero. El índice de modificación refleja cuál sería el decrecimiento de la χ^2 que se produciría si ese parámetro se liberara, es decir, se estimara libremente. Recordemos que caídas en la χ^2 implican mejoras en el ajuste del modelo. Estamos hablando de la comparación entre la χ^2 del modelo que hemos estimado y la χ^2 resultante de estimar ese parámetro adicional, es decir, una prueba de diferencia de dos χ^2 con un grado de libertad. Habrá, pues, un índice de modificación para cada parámetro no estimado en el modelo.

En general un modelo que tenga un buen ajuste debería producir índices de modificación pequeños. Como ya hemos señalado, cada MI es una diferencia de χ^2 con un grado de libertad. En tablas para $p < 0{,}05$, el valor crítico de este estadístico es 3,84, por lo que se suele redondear señalando que valores superiores a 4 implican índices de modificación significativos (Jaccard y Wan, 1989).

Como señala Brown (2006), los MI, en cuanto que prueba χ^2, tienen limitaciones similares a este estadístico que ya apuntamos: sensibilidad a problemas de normalidad y al tamaño muestral. Por ello no es raro que nos indique que la mejora en la χ^2 es significativa aunque la magnitud del parámetro que se añade —una carga factorial, por ejemplo— sea muy pequeña. Por este motivo, la mayoría de programas, `lavaan` entre ellos, añaden el *expected parameter change* (EPC) o valor esperado del parámetro cuando se estime (como está fijado a cero, el *change* será siempre el valor del parámetro tras la estimación). Estos valores se ofrecen no estandarizados, estandarizados o "completamente" estandarizados (teniendo en cuenta todos los parámetros del modelo).

En la sintaxis que ofrecimos para estimar nuestro CFA, la solicitud de los índices de modificación se realizaba como sigue. En la sintaxis le estamos señalando que nos los ofrezca ordenados de mayor a menor (`sort.=TRUE`) y que

Estadísticos e-Books & Papers

Cuadro 13.12.: Índices de modificación (MI)

```
   lhs op rhs    mi   epc sepc.lv sepc.all sepc.nox
20  IQ =~ FSF 4.759 0.146   0.146    0.247    0.247
23   L ~~   H 4.759 0.156   0.156    0.146    0.146
```

no nos los enseñe todos, solo aquellos que sean significativos de acuerdo con el valor crítico de una χ^2 con un grado de libertad (3.84): `minimum.value=3.84`. El resultado aparece en el cuadro 13.12.

```
#Indices de modificacion

modindices(fit, sort.=TRUE, minimum.value = 3.84)
```

En lo primero que hemos de fijarnos es en el tipo de parámetro que se está liberando. Cuando el signo es $=\sim$ lo que se muestra son cargas factoriales (repásese la sintaxis en la que definimos el modelo para ver que es así), en la primera línea del cuadro 13.12 lo que nos propone con `IQ=~FSF` es que el indicador FSF también sea un indicador de IQ. Si esto se hiciera, la χ^2 tendría una caída de `mi`=4.759 puntos (caída significativa), pues es superior al valor crítico de 3,84), pero esa carga tendría un valor muy bajo (`sepc.all`=0,247) con lo que el interés de introducirla es escaso dado, además, que el ajuste del modelo era bueno. Por leer toda la tabla con `epc` nos está señalando cuál sería el valor del parámetro sin estandarizar, `sepc.lv`, solo estandarizando las variables latentes, `sepc.all`, como hemos indicado, estandarizando todas las variables y `sepc.nox` estandarizándolas todas menos los indicadores independientes. Siempre leeremos `sepc.all`.

Cuando el signo es $\sim\sim$, como en la segunda línea, lo que nos está proponiendo es que añadamos una correlación entre las varianzas de los términos de error, concretamente de los indicadores L y H.

13.8. Un ejemplo completo de CFA

En las secciones anteriores hemos simultaneado la presentación de los resultados del CFA con la teoría que los apoyaba. Con el fin de centrarnos en los resultados en sí mismos, vamos a desarrollar a continuación un caso completo donde no haremos alusión alguna a la teoría que subyace bajo los procesos de identificación y estimación, sino que nos limitaremos a su aplicación para facilitar su seguimiento.

Caso 13.2 Determinantes de la aceptación de la publicidad en el móvil
El caso que vamos a desarrollar era el que presentábamos en la figura 13.1 y que se corresponde con el trabajo de Aldás *et al.* (2013). Remitimos a este trabajo para la justificación teórica de las relaciones pero, básicamente, lo que

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 13.13.: Instrumento de medida del modelo

Factor	Ítem	Enunciado	Fuente
Actitud	att1	Me gusta la PM	Taylor y Todd (1995)
	att2	Creo que la PM es interesante	
	att3	Me parece una buena idea la PM	
	att4	La PM es algo positivo para mí	
Entrete-nimiento	ent1	La PM es entretenida	Merisavo *et al.* (2007)
	ent2	La PM es divertida	
Utilidad	usf1	La PM me permite ahorrar dinero	Tsang *et al.* (2004) Ducoffe (1996)
	usf2	La PM me permite ahorrar tiempo	
	usf3	La PM me reporta beneficios	
Irritación	irr1	La PM es irritante	Ducoffe (1996)
	irr2	La PM es entrometida	
	irr3	La PM es engañosa	
	irr4	La PM es molesta	
Acepta-ción	acc1	No me importa recibir PM	Karjaluoto *et al.* (2008) Bauer *et al.* (2005)
	acc2	Me gusta recibir PM	
	acc3	Leeré toda la PM que reciba	

Fuente: Aldás *et al.* (2013)

plantea es que la aceptación de la publicidad en el móvil viene condicionada porque el usuario perciba una utilidad en la misma y que el entretenimiento que genere su creatividad sea capaz de minorar la irritación que la intrusión pueda causar. Para estimar el modelo estructural que aparece en la figura 13.1 (capítulo 15) es necesario, previamente y como hemos señalado reiteradamente, estimar un CFA para, con la información que proporciona (capítulo 13), evaluar la fiabilidad y validez del instrumento de medida (capítulo 14).

La figura 13.8 ilustra el CFA que se corresponde con el instrumento de medida del modelo representado en la figura 13.1. Sobre la representación llamamos la atención respecto al modo en que hemos ilustrado las covarianzas entre los factores. Para no saturar el dibujo de flechas hemos seguido la convención que se observa, pero esas flechas están representando el total de 15 correlaciones que se dan entre los 5 factores implicados.

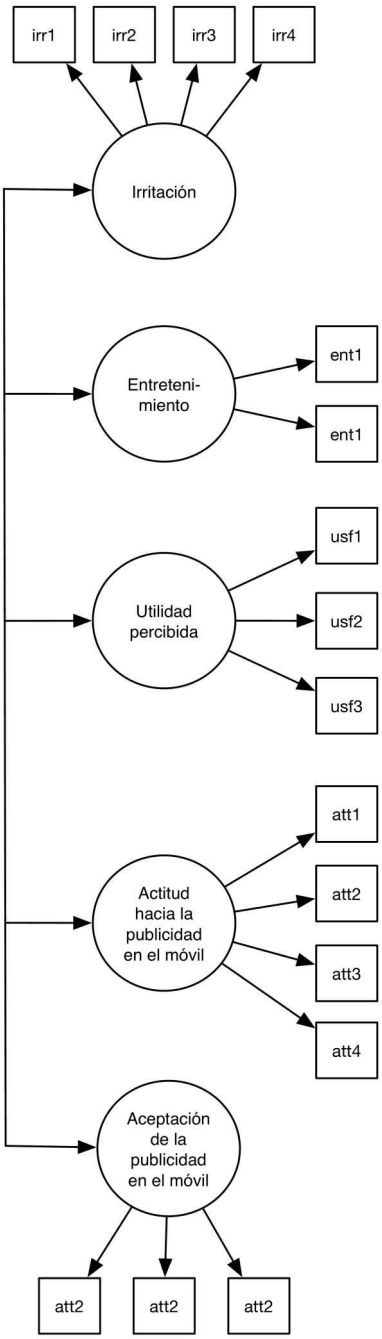
El instrumento de medida que se está empleando se corresponde con las escalas que mostramos en el cuadro 13.13.

Comenzaremos con el proceso de **identificación del modelo**. Para ello:

- Estableceremos la escala fijando a 1 la varianza de los factores.
- Comprobaremos que el modelo está sobreidentificado (número positivo de grados de libertad). Bastará mirar en la salida el número de grados de libertad de la χ^2, pero los contaremos manualmente una vez más con carácter previo.

Estadísticos e-Books & Papers

Figura 13.8.: CFA por estimar en el caso 13.2



Fuente: Aldás *et al.* (2013)

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

- Fijaremos a 1 los coeficientes de regresión de los términos de error con los indicadores. Basta con no indicar nada en la sintaxis, dado que las ecuaciones de medida no incorporan el término de error asumiéndose que es 1.
- Fijaremos a 0 las covarianzas entre los términos de error. Basta con no incorporarlas expresamente a la sintaxis, con lo que se asume que son 0.

La visión gráfica de la identificación la tenemos en la figura 13.9. En ella podemos observar que han de estimar un total de 42 parámetros marcados con un * que se corresponden con:

- 16 cargas factoriales.
- 16 varianzas de los términos de error.
- 10 correlaciones entre los factores.

Se han fijado para la identificación:

- 16 coeficientes de regresión de los errores a 1.
- 136 covarianzas entre los errores a 0.
- 5 varianzas de los factores a 1 (escala).

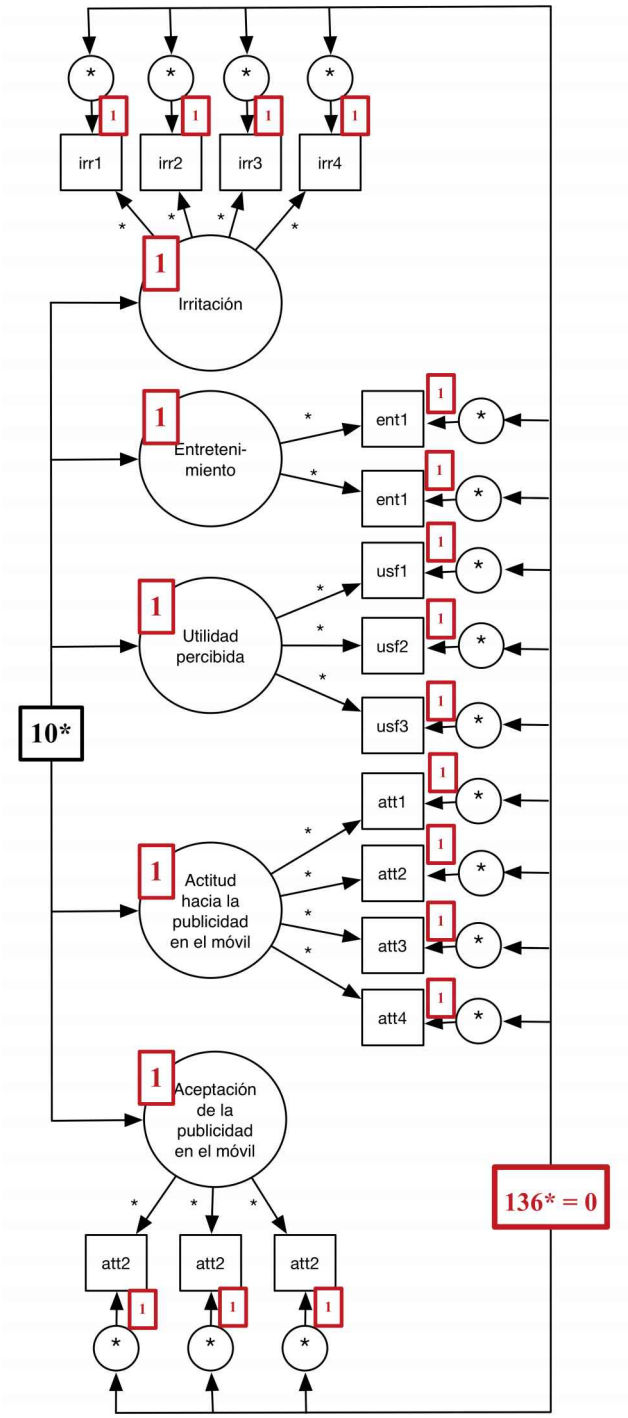
Dado que disponemos de $16(16+1)/2 = 136$ varianzas y covarianzas muestrales distintas, el número de grados de libertad será de $136-42 = 94$ grados de libertad, contando con un modelo sobreidentificado, como podremos comprobar posteriormente en la salida de **lavaan**.

Veamos la sintaxis de **lavaan**. En primer lugar importamos los datos, cargamos las librerías y definimos el modelo:

```
datos<-Datos_13_2_Caso

library (lavaan)
library(semTools)
library(semPlot)
library(ggplot2)

modelo.cfa <- '

# Modelo de medida
actitud          =~ att1+att2+att3+att4
entretenimiento  =~ ent1+ent2
utilidad         =~ usf1+usf2+usf3
```


Estadísticos e-Books & Papers

Figura 13.9.: CFA por estimar en el caso 13.2



Fuente: Aldás *et al.* (2013)

© Ediciones Paraninfo

Estadísticos e-Books & Papers

```
irritacion       =~ irr1+irr2+irr3+irr4
aceptacion       =~ acc1+acc2+acc3

#Varianzas de los factores
actitud~~actitud
entretenimiento~~entretenimiento
utilidad~~utilidad
irritacion~~irritacion
aceptacion~~aceptacion

#Covarianzas
actitud~~entretenimiento
actitud~~utilidad
actitud~~irritacion
actitud~~aceptacion
entretenimiento~~utilidad
entretenimiento~~irritacion
entretenimiento~~aceptacion
utilidad~~irritacion
utilidad~~aceptacion
irritacion~~aceptacion

#Varianzas de los terminos de error
att1~~att1 att2~~att2 att3~~att3 att4~~att4 ent1~~ent1
ent2~~ent2 usf1~~usf1 usf2~~usf2 usf3~~usf3 irr1~~irr1
irr2~~irr2 irr3~~irr3 irr4~~irr4 acc1~~acc1 acc2~~acc2
acc3~~acc3 '
```

Donde todas las expresiones se han presentado con anterioridad, solo destacar que no hemos fijado a 1 la varianza de los factores mediante las expresiones tipo **actitud~~1*actitud** porque lo haremos con el planteamiento más eficiente de **std.lv=TRUE** cuando planteemos la estimación a continuación.

```
#Estimacion del modelo
fit <- lavaan(modelo.cfa, data=datos, std.lv=TRUE, mimic="eqs",
estimator="ML", verbose=TRUE, warn=TRUE)

#Peticion de elementos en la salida
summary (fit, fit.measures=TRUE, standardized=TRUE, rsquare=TRUE)
resid(fit, type="cor")

#Indices de modificacion
modindices(fit, sort.=TRUE, minimum.value = 3.84)
```

Una vez más toda la notación ya se ha presentado con anterioridad, de hecho,

Estadísticos e-Books & Papers

como hemos mantenido los nombres de los objetos de `R (modelo.cfa, fit)`, no nos es necesario modificar nunca esta parte de la sintaxis.

Nos resta únicamente pedirle a `lavaan` que nos ofrezca el histograma de residuos y también el grafo del modelo, que es recomendable pedir para asegurarnos que no hemos cometido errores en la sintaxis y se corresponde con el modelo teórico que queríamos estimar. Simplemente le pedimos que nos los muestre con forma de árbol `(layout=tree)`, el `style="lisrel"` solo afecta de momento a la forma en que se representan en el gráfico los errores, con `what="std"` le pedimos que las flechas con las cargas factoriales tengan un grosor proporcional al tamaño de las cargas y con `whatLabels="std"` que etiquete esas flechas con las cargas estandarizadas.

Basta ejecutar la sintaxis para obtener los resultados de las estimaciones. Nos centraremos, en primer lugar, en el análisis del ajuste que aparece recogido en el cuadro 13.14. Vemos que, pese a que contamos con una χ^2 significativa que rechaza la hipótesis nula de igualdad entre las matrices de varianzas y covarianzas muestrales y estimadas, el resto de indicadores de ajuste apunta hacia un ajuste excelente: el CFI es superior a 0,95 y el TLI a 0,90, es decir, en ambos casos de acuerdo con el *benchmark* establecido para considerarlo de ese modo. El SRMR es 0,045, muy inferior al 0,08 que señalaría el máximo para considerarse razonable y el RMSEA (0,062) está en el intervalo del ajuste aceptable [0,05;0,08]. Nótese que el número de grados de libertad de la χ^2 es de 94, que coincide con la estimación manual que habíamos hecho en el proceso de identificación. También el histograma de los errores de la figura 13.10 muestra que estos están centrados alrededor del cero sin errores en los extremos (valores grandes).

Aunque hasta el tema 14 no sabremos qué debemos hacer con esa información, el cuadro 13.15 nos muestra que todas las cargas son significativas y con valores que se mueven entre 0,6 y 0,8. Veremos en su momento que estos valores nos van a permitir constatar la validez convergente del modelo.

Finalmente, el cuadro 13.16 nos muestra que no hay estimaciones anómalas que nos hagan sospechar de una incorrecta especificación. No hay correlaciones superiores a la unidad ni varianzas negativas. En el cuadro 13.15 ya habíamos visto que ninguna carga estandarizada tomaba tampoco valores superiores a la unidad.

La figura 13.11 se muestra solo a efectos de comprobar que el modelo que hemos introducido con la sintaxis se corresponde con el teórico que buscábamos estimar y que habíamos representado en la figura 13.9.

Con toda la información disponible de un CFA bien ajustado y sin errores de identificación, pasamos en el tema siguiente a ver cómo analizar esta información para comprobar la validez del instrumento de medida y poder, con confianza, estimar el modelo estructural de la figura 13.1.

 © Ediciones Paraninfo

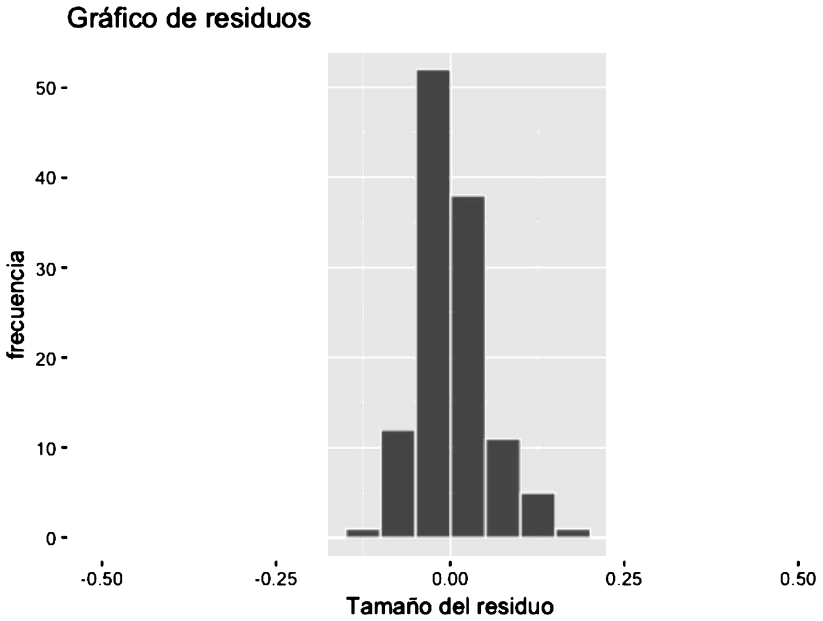
Estadísticos e-Books & Papers

Cuadro 13.14.: Indicadores de ajuste del CFA

```
lavaan (0.5-22) converged normally after  30 iterations

  Number of observations                           353

  Estimator                                         ML
  Minimum Function Test Statistic              219.772
  Degrees of freedom                                94
  P-value (Chi-square)                           0.000

Model test baseline model:

  Minimum Function Test Statistic             2993.948
  Degrees of freedom                               120
  P-value                                        0.000

User model versus baseline model:

  Comparative Fit Index (CFI)                    0.956
  Tucker-Lewis Index (TLI)                       0.944

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)                     NA
  Loglikelihood unrestricted model (H1)             NA

  Number of free parameters                         42
  Akaike (AIC)                                      NA
  Bayesian (BIC)                                    NA

Root Mean Square Error of Approximation:

  RMSEA                                          0.062
  90 Percent Confidence Interval          0.051  0.072
  P-value RMSEA <= 0.05                          0.036

Standardized Root Mean Square Residual:

  SRMR                                           0.045
```


Estadísticos e-Books & Papers

Figura 13.10.: Histograma de errores



Cuadro 13.15.: Estimación de las cargas factoriales

```
Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  actitud =~
    att1              0.786    0.044   17.993    0.000    0.786    0.810
    att2              0.828    0.041   20.351    0.000    0.828    0.877
    att3              0.935    0.049   19.018    0.000    0.935    0.840
    att4              0.868    0.043   20.145    0.000    0.868    0.872
  entretenimiento =~
    ent1              0.813    0.050   16.422    0.000    0.813    0.850
    ent2              0.801    0.046   17.344    0.000    0.801    0.895
  utilidad =~
    usf1              0.954    0.054   17.633    0.000    0.954    0.835
    usf2              0.805    0.044   18.319    0.000    0.805    0.860
    usf3              0.706    0.056   12.563    0.000    0.706    0.639
  irritacion =~
    irr1              1.003    0.072   13.926    0.000    1.003    0.698
    irr2              0.972    0.069   13.998    0.000    0.972    0.701
    irr3              1.030    0.062   16.619    0.000    1.030    0.798
    irr4              1.033    0.062   16.538    0.000    1.033    0.795
  aceptacion =~
    acc1              0.818    0.068   12.034    0.000    0.818    0.636
    acc2              0.794    0.056   14.177    0.000    0.794    0.731
    acc3              0.836    0.051   16.313    0.000    0.836    0.823
```

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 13.16.: Estimación de las varianzas de factores, errores y covarianzas entre factores

```
Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  actitud ~~
    entretenimient    0.427    0.050    8.482    0.000    0.427    0.427
    utilidad          0.402    0.052    7.770    0.000    0.402    0.402
    irritacion       -0.194    0.059   -3.313    0.001   -0.194   -0.194
    aceptacion        0.498    0.050   10.025    0.000    0.498    0.498
  entretenimiento ~~
    utilidad          0.496    0.049   10.068    0.000    0.496    0.496
    irritacion       -0.180    0.061   -2.970    0.003   -0.180   -0.180
    aceptacion        0.363    0.057    6.336    0.000    0.363    0.363
  utilidad ~~
    irritacion       -0.327    0.057   -5.714    0.000   -0.327   -0.327
    aceptacion        0.480    0.053    9.084    0.000    0.480    0.480
  irritacion ~~
    aceptacion       -0.362    0.058   -6.215    0.000   -0.362   -0.362

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
    actitud           1.000                               1.000    1.000
    entretenimient    1.000                               1.000    1.000
    utilidad          1.000                               1.000    1.000
    irritacion        1.000                               1.000    1.000
    aceptacion        1.000                               1.000    1.000
   .att1              0.323    0.030   10.928    0.000    0.323    0.343
   .att2              0.206    0.023    9.124    0.000    0.206    0.231
   .att3              0.364    0.035   10.306    0.000    0.364    0.294
   .att4              0.238    0.025    9.340    0.000    0.238    0.240
   .ent1              0.254    0.050    5.112    0.000    0.254    0.277
   .ent2              0.159    0.046    3.463    0.001    0.159    0.199
   .usf1              0.394    0.052    7.512    0.000    0.394    0.302
   .usf2              0.229    0.035    6.514    0.000    0.229    0.261
   .usf3              0.720    0.061   11.762    0.000    0.720    0.591
   .irr1              1.057    0.097   10.909    0.000    1.057    0.512
   .irr2              0.977    0.090   10.870    0.000    0.977    0.508
   .irr3              0.605    0.069    8.827    0.000    0.605    0.363
   .irr4              0.620    0.070    8.912    0.000    0.620    0.368
   .acc1              0.982    0.088   11.105    0.000    0.982    0.595
   .acc2              0.551    0.059    9.403    0.000    0.551    0.466
   .acc3              0.333    0.050    6.599    0.000    0.333    0.323
```


Estadísticos e-Books & Papers

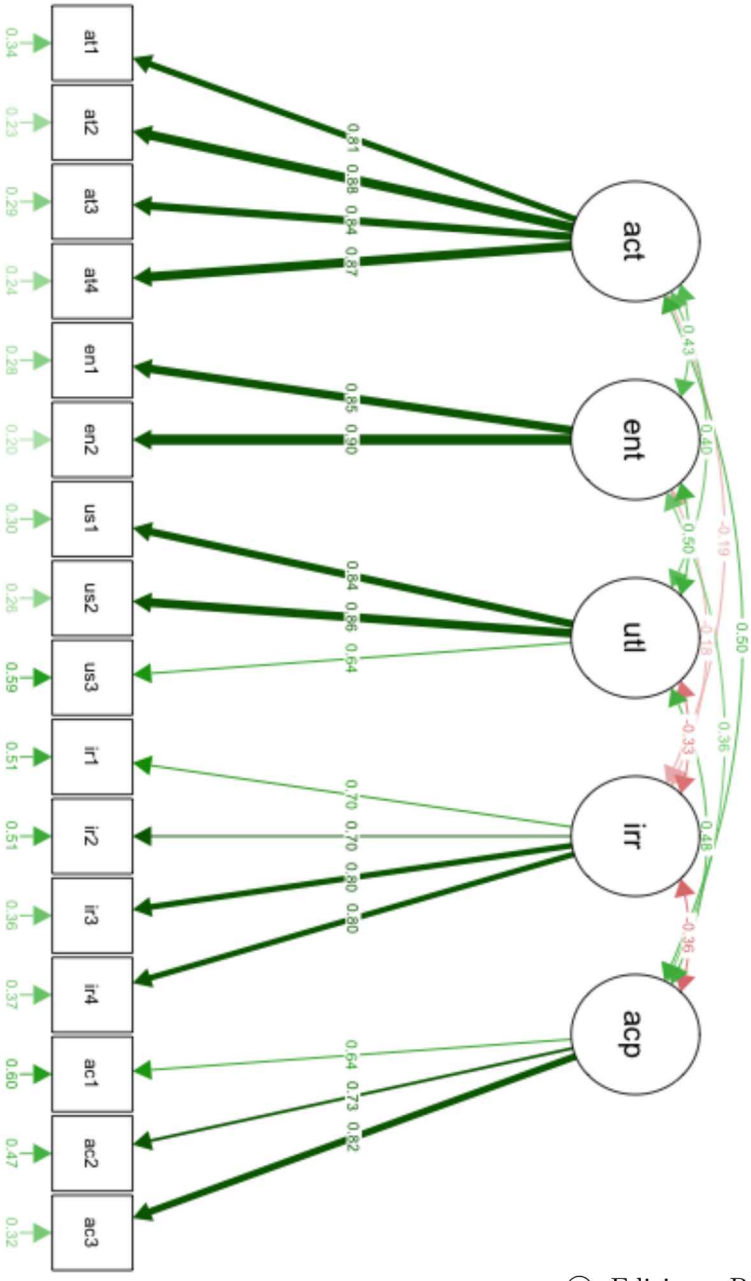


Figura 13.11.: Grafo del CFA estimado por `lavaan`

© Ediciones Paraninfo

Estadísticos e-Books & Papers

13.9. Anexo 13.1

El problema de la identificación

Consideremos un CFA tal y como fue planteado en la ecuación (13.3):

$$\mathbf{x} = \Lambda\xi + \delta$$

Sabemos que la matriz Σ que contiene las varianzas y covarianzas de las variables observadas puede descomponerse tal y como marcaba la expresión (13.8):

$$\Sigma = \Lambda\Phi\Lambda' + \Theta$$

Si no se impone ningún tipo de restricción a los parámetros de Λ, Θ y Φ en el momento en que haya un conjunto de parámetros que cumplen (13.8), entonces habrá un infinito número de ellos. Veámoslo. Sea M cualquier matriz $s \times s$ no singular, por tanto, invertible. Si definimos:

$$\ddot{\Lambda} = \Lambda\mathbf{M^{-1}};\ \ddot{\xi} = \mathbf{M}\xi$$

y entonces:

$$\ddot{\Lambda}\ddot{\xi} + \delta = \left(\Lambda\mathbf{M^{-1}}\right)(\mathbf{M}\xi) + \delta$$

$$= \Lambda\left(\mathbf{M^{-1}M}\right)\xi + \delta$$

$$= \Lambda\xi + \delta$$

de tal forma que si $\mathbf{x}$ cumple (13.3) también cumplirá que:

$$\mathbf{x} = \ddot{\Lambda}\ddot{\xi} + \delta$$

La matriz de varianzas y covarianzas de $\ddot{\xi}$ vendrá dada por:

$$\ddot{\Phi} = E\left[\ddot{\xi}\ddot{\xi}'\right] = E\left[(\mathbf{M}\xi)(\mathbf{M}\xi)'\right] = \mathbf{M}E\left[\xi\xi'\right]\mathbf{M}' = \mathbf{M}\Phi\mathbf{M}'$$

Si operamos en (13.8) se cumplirá, entonces, que:

$$\ddot{\Lambda}\ddot{\Phi}\ddot{\Lambda}' + \Phi = \left(\Lambda\mathbf{M^{-1}}\right)(\mathbf{M}\Phi\mathbf{M}')\left(\mathbf{M}'^{-1}\Lambda'\right) + \Theta$$

$$= \Lambda\left(\mathbf{MM^{-1}}\right)\Phi\left(\mathbf{M'M'^{-1}}\right)\Lambda' + \Theta$$

$$= \Lambda\Phi\Lambda' + \Theta = \Sigma$$

Por lo tanto, si Σ cumple (13.8), también se cumple que:

Estadísticos e-Books & Papers

$$\Sigma = \ddot{\Lambda}\ddot{\Phi}\ddot{\Lambda}' + \Theta$$

Dado que las matrices marcadas con ¨ solo serían iguales a las originales en el caso en que $\mathbf{M} = \mathbf{I}$, existen infinitas matrices $\mathbf{M}$ invertibles que dan lugar a infinitas soluciones del modelo. En consecuencia, este modelo se definiría como no identificado.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

14. Modelos de ecuaciones estructurales: validación del instrumento de medida

14.1. Introducción

Como planteábamos en el tema anterior, la estimación de un modelo de ecuaciones estructurales es un proceso que comienza validando el instrumento de medida utilizado antes de estimar las relaciones estructurales que soportan las hipótesis del investigador. Para validar el instrumento de medida son necesarios dos pasos: (1) estimar un análisis factorial confirmatorio CFA, como vimos en el tema 13, y (2) aplicar a los resultados obtenidos del CFA una serie de criterios que nos han de permitir determinar si ese instrumento de medida tiene la "calidad" suficiente como para poder estimar el modelo estructural, puesto que de nada sirve establecer que el factor A influye significativamente en el B si no sabemos qué significan A y B porque confiamos en la calidad de los indicadores que hemos utilizado para medirlos. En este tema veremos que ese análisis de la "calidad" del instrumento de medida se traducirá en evaluar su fiabilidad y su validez. El objetivo de este tema es definir esos conceptos y los procedimientos para su constatación.

14.2. La medición en ciencias sociales

Medir es una de las principales tareas de la ciencia. Adquirimos conocimientos acerca de determinadas características de los consumidores o de las empresas observándolos, pero para dotar de sentido a estas características es necesario cuantificarlas.

Podríamos definir la **medida** como "un conjunto de reglas que permiten asignar números a los objetos observados de tal forma que representen de manera adecuada la cantidad de un determinado atributo que poseen" (Nunnally y Bernstein, 1994). En algunos casos esas reglas son muy obvias, como cuando se trata de medir la altura de un individuo con una cinta métrica. Pero, desgraciadamente, la obviedad no es común en las ciencias sociales. Así, las reglas para determinar el grado de timidez, de etnocentrismo o de inteligencia de un consumidor, o el grado de orientación al mercado de una empresa no son tan intuitivas.

Estadísticos e-Books & Papers

Para medir esas variables es necesario crear (o utilizar si ya existe) una **escala de medida**. Una escala de medida es un conjunto de ítems, frases o preguntas que permiten medir el nivel que alcanza un atributo determinado (etnocentrismo, orientación al mercado) no directamente observable en un objeto (un consumidor, una empresa).

¿Qué **ventajas** aporta a la investigación la utilización de escalas estandarizadas de medida? Podemos resumirlas en las siguientes:

1. *Objetividad.* Un principio científico básico es que cualquier afirmación efectuada por un investigador ha de poder ser verificada independientemente por cualquier otro repitiendo el experimento en las mismas circunstancias. Aunque este principio es de limitada aplicación en ciencias sociales, no deja de ser cierto que, si, ni siquiera se dispone de un instrumento para medir el atributo, difícilmente se podrán comparar los resultados.

2. *Cuantificación.* Los resultados numéricos de medidas estandarizadas tienen la ventaja de que permiten la utilización de técnicas estadísticas avanzadas.

3. *Comunicación.* Las medidas objetivas permiten la fácil comunicación de los resultados de las investigaciones entre los distintos científicos.

4. *Economía.* Aunque desarrollar escalas adecuadamente es una tarea costosa, una vez creadas suponen un gran ahorro de tiempo. Así, por ejemplo, un investigador preparado puede juzgar bastante acertadamente el grado de etnocentrismo de un consumidor manteniendo con él una entrevista en profundidad. Valorar esta misma variable en 100 consumidores con 100 entrevistas le resultará, sin embargo, muchísimo más costoso que administrarles un cuestionario con las preguntas que conforman la escala CETSCALE (Shimp y Sharma, 1987).

Desgraciadamente, pese a sus evidentes ventajas, no todas las escalas se desarrollan con el suficiente cuidado. Una escala no es un conjunto de preguntas unidas sin más, sino que debe reunir una serie de propiedades adecuadas conocidas como propiedades psicométricas. Estas son básicamente dos: fiabilidad y validez. En principio seguiremos los pasos de una investigación habitual: tomamos las escalas desarrolladas por otros y solo hemos de comprobar determinadas propiedades en su aplicación. Sin embargo, aunque es menos habitual, incluso raro, que una investigación se centre en el desarrollo de una nueva escala, terminaremos el capítulo planteando un esquema en ocho pasos que sirva como guía para obtener escalas adecuadas.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

14.3. Análisis de la fiabilidad del instrumento de medida

Por **fiabilidad** se entiende la propiedad de que aplicaciones repetidas de un mismo instrumento de medida deben dar resultados consistentes. Sin embargo rara vez es posible la administración repetida en una misma investigación —por limitaciones de tiempo y económicas— por lo que solemos evaluar la **consistencia interna** del instrumento de medida. Entendemos por consistencia interna el grado en que los ítems que conforman una escala están correlacionados entre sí. Si comparten una causa común, la variable latente que están midiendo, esta correlación debería ser alta. Si diseñamos las preguntas de un examen para medir el conocimiento de una asignatura, aquellas personas con un conocimiento alto contestarán las preguntas más o menos igual —bien— y las notas de cada pregunta estarán correlacionadas y también lo estarán si la persona tiene un conocimiento bajo porque también contestará más o menos igual a todas —mal—. La correlación será un indicador de que hemos diseñado bien el examen, porque la causa común —el conocimiento de la asignatura— provoca un comportamiento solidario de los ítems.

Es muy importante señalar que la fiabilidad de una escala indica solamente que los distintos ítems que la componen, al estar muy correlacionados entre sí, están midiendo la misma variable latente. Pero que una escala sea fiable no quiere decir que la variable latente que está midiendo sea la que tiene que medir, es decir, que sea válida. Por ejemplo, supongamos que en un examen de investigación de mercados hemos diseñado un test de 50 preguntas y que, por error, el día del examen repartimos otro test que hemos diseñado para el examen de física cuántica. Este test será probablemente fiable, en cuanto que todas las preguntas obtendrán puntuaciones muy correlacionadas entre sí, dado que están midiendo lo mismo: el (des)conocimiento de los alumnos de la asignatura de física cuántica. Lo que ocurre es que no son los conocimientos de física cuántica los que queríamos medir, sino los de investigación de mercados. La escala diseñada es fiable, pero no válida.

Sin embargo, una escala no podrá ser válida si no es fiable, de tal forma que la fiabilidad se convierte en una condición necesaria, aunque no suficiente, de la validez. Existen diversos procedimientos para medir la fiabilidad de una escala tal y como la hemos definido, es decir, en términos de consistencia interna. En este tema, nos centraremos en el coeficiente α de Cronbach (Cronbach, 1951) por varias razones. En primer lugar porque es la medida de fiabilidad más utilizada y, en segundo lugar, porque su conexión con la definición de fiabilidad que acabamos de proporcionar es tan directa que nos ayudará a reforzar el concepto.

14.3.1. Coeficiente α de Cronbach

Supongamos que tenemos una escala formada por k ítems $(X_1, X_2, .., X_k)$ que hemos construido para medir la variable latente Y y de la cual queremos calcu-

Estadísticos e-Books & Papers

lar su fiabilidad. Para ello, construiremos y analizaremos su matriz de varianzas y covarianzas **C**[1]:

$$C = \begin{bmatrix} \sigma_1^2 & & & & \\ \sigma_{12}^2 & \sigma_2^2 & & & \\ \sigma_{13}^2 & \sigma_{23}^2 & \sigma_3^2 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \sigma_{1k}^2 & \sigma_{2k}^2 & \sigma_{3k}^2 & \cdots & \sigma_k^2 \end{bmatrix}$$

Una matriz de varianzas y covarianzas es una generalización de una matriz de correlaciones, solo que sus elementos no están estandarizados. Los elementos σ_i^2 de la diagonal expresan la varianza de cada ítem i de la escala. Los elementos σ_{ij} de fuera de la diagonal indican la covarianza entre dos ítems i y j de la escala.

Nosotros hemos construido la escala para medir el constructo Y (p. ej. orientación al mercado) y por ello asumiremos que Y está bien representado por la suma de los k ítems de la escala.

Las matrices de varianzas y covarianzas tienen un conjunto de propiedades muy útiles. Si sumamos todos los elementos de la matriz **C** tenemos la **varianza total** de la escala Y. Pues bien, el coeficiente α de Cronbach, se define como la proporción de la varianza total de la escala que es atribuible a la variable latente Y. Cuanto mayor sea este valor, querrá decir que Y está mejor representado por la escala porque está causando (explicando) la mayor parte de la varianza de esta (**varianza común**).

La parte de la varianza total que no explica la variable latente es la causada por los errores de medida de cada ítem y se denomina **varianza específica o residual.** Hay que hacer notar que cada término de error provoca solo varianza en cada ítem por separado y esos errores no están correlacionados unos con otros. En resumen, cada ítem (y por suma, el conjunto de la escala) solo varía como función de:

1. La fuente de variación que supone la variable latente (varianza común).
2. La fuente de variación que provoca el error (varianza específica).

Veamos cómo nos ayuda la matriz de varianzas y covarianzas a recoger esta información. Como hemos dicho, la varianza total de la escala (σ_y^2) es la suma de todos los elementos de esa matriz. Asimismo, la suma de los elementos de la diagonal nos proporcionará la suma de la varianza de los ítems individuales que no viene explicada por el factor o varianza residual ($\Sigma\sigma_i^2$). Las covarianzas recogen variación entre pares de ítems, y como a los ítems solo los puede hacer covariar la causa común, así se define ésta, como la suma de las covarianzas. Como la varianza total es la suma de la común y la específica, es obvio que la común será la resta de la total menos la específica. La definición más intuitiva

[1] Al ser simétrica representamos únicamente el triángulo inferior.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

para el α de Cronbach es la siguiente: es la parte de la varianza total que es común, es decir, cuanto más explique el factor de la varianza total, mayor será la fiabilidad, pero como la varianza común es la total menos la residual $\sigma_y^2 - \Sigma\sigma_i^2$, entonces la ratio sobre la total puede ponerse:

$$\alpha = \frac{\text{Varianza común}}{\text{Varianza total}} = \frac{\sigma_y^2 - \Sigma\sigma_i^2}{\sigma_y^2} = 1 - \frac{\Sigma\sigma_i^2}{\sigma_y^2}$$

En la expresión anterior, sin embargo, hace falta introducir una última corrección para tener la expresión del α de Cronbach. El número total de elementos de la matriz de varianzas y covarianzas es k^2. El número de elementos de la matriz que son específicos (la diagonal) es k, mientras que los elementos comunes (fuera de la diagonal) son $k^2 - k$. Tenemos, en la expresión anterior, una fracción con un numerador basado en k valores y un denominador basado en k^2 valores. Para ajustar los cálculos de tal forma que la ratio exprese las magnitudes relativas, más que el número de términos que hay en numerador y denominador, corregiremos la expresión anterior para contrarrestar el efecto de la diferencia por $k^2/(k^2 - k)$ o, lo que es lo mismo, por $k/(k-1)$, de forma que ahora α estará acotado entre 0 y 1:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_y^2}\right) \tag{14.1}$$

Para resumir, la fiabilidad medida por la expresión anterior nos indica qué proporción de la varianza total está provocada por la variable latente, corrigiendo este valor por el número de casos que intervienen en los cálculos.

Si se desea trabajar en términos de correlaciones entre los ítems de la escala, en lugar de en términos de varianzas-covarianzas, la expresión anterior puede adaptarse de una manera sencilla. La suma de las varianzas individuales de cada ítem $\Sigma\sigma_i^2$ puede ponerse como el producto entre el número de ítems (k) y la media de las varianzas de los mismos v (por ejemplo, 10 ítems que suman 50 y 10 veces su media, que es 5, dan el mismo valor). Por lo que respecta al denominador, si llamamos c a la media de las covarianzas, como hay k varianzas y $k^2 - k$ covarianzas, podremos poner la expresión anterior del α de Cronbach como sigue:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{kv}{kv + (k^2 - k)c}\right)$$

sustituyendo el 1 por su equivalente $[kv + (k^2 - k)c]/[kv + (k^2 - k)c]$ y operando, se llega a:

$$\alpha = \frac{kc}{v + (k-1)\,c}$$

Como estamos buscando una expresión que utilice correlaciones en lugar de varianzas y covarianzas, si estandarizamos las variables implicadas en la

Estadísticos e-Books & Papers

expresión anterior, la media de las covarianzas c se convierte en la media de las correlaciones (ρ) y la media de las varianzas es igual a 1, por lo que el α de Cronbach se puede calcular como sigue:

$$\alpha = \frac{k\rho}{1 + (k-1)\rho} \tag{14.2}$$

donde ρ es, como ya hemos indicado, la media de los coeficientes de correlación entre todos los ítems que conforman la escala. A esta fórmula se la conoce como la **fórmula de Spearman-Brown** (Crocker y Algina, 1986). Esta fórmula es, en mi opinión, tremendamente ilustrativa del concepto de fiabilidad. Al fin y al cabo nos dice que el α de Cronbach no es sino un promedio de las correlaciones entre los indicadores. Si los indicadores se comportan de manera solidaria por tener una causa común —el factor que están midiendo— estamos ante una escala fiable.

¿Cuáles son los valores del α de Cronbach por debajo de los cuales no se puede considerar como fiable a una escala? Siguiendo a Nunnally y Bernstein (1994), podremos afirmar que este nivel depende de para qué vaya a utilizarse la escala. En etapas preliminares de desarrollo de una escala, un nivel de 0,7 puede ser suficiente y, tras las depuraciones oportunas de la escala, este valor no debe bajar nunca de 0,8. Si en función de los valores de la escala se van a tomar decisiones que afecten a los individuos (asignar alumnos a clases distintas según los resultados de un test de inteligencia, por ejemplo), el α no podrá ser inferior a 0,9.

Caso 14.1. Una escala para la conciencia social

Hatcher (1994) y O'Rourke y Hatcher (2013) ofrecen un buen ejemplo de funcionamiento del α de Cronbach y su capacidad para detectar ítems con comportamientos anómalos. Pretenden medir la conciencia social de las personas como formada por dos conceptos distintos: la propensión a ayudar a las personas que tenemos a nuestro alrededor (compañeros de trabajo, amigos y familiares) y la ayuda a los que no están cerca de nosotros a través del respaldo económico. Proponen para ello la escala que recoge el cuadro 14.1 donde pide a las personas que indiquen con qué frecuencia han realizado la acción señalada durante los últimos 6 meses, siendo 1 = nunca, 2 = casi nunca, 3 = pocas veces, hasta 7 = muy frecuentemente.

El investigador pretende analizar la fiabilidad de las dos subescalas y para ello va a calcular el α de Cronbach. Es evidente que los tres primeros ítems forman la escala de ayuda a los conocidos, mientras que los tres últimos forman la de las donaciones económicas, sin embargo comete un error al solicitar el α de Cronbach e incorpora el ítem B1 a la escala de ayuda a los conocidos que no le corresponde. Usaremos para el caso la función `alpha {psych}`.

```
myvars <- c("A1", "A2", "A3", "B1")
subescala1 <- datos[myvars]
alpha(subescala1)
```

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 14.1.: Escalas para medir dos dimensiones de la conciencia social

Dimensión	Ítem	Enunciado	Escala
Ayuda a conocidos	A1	Abandono lo que estoy haciendo para ayudar a un compañero de trabajo	1 2 3 4 5 6 7
	A2	Abandono lo que estoy haciendo para ayudar a un pariente	1 2 3 4 5 6 7
	A3	Abandono lo que estoy haciendo para ayudar a un amigo	1 2 3 4 5 6 7
Donaciones económicas	B1	Doy dinero para caridad en la iglesia	1 2 3 4 5 6 7
	B2	Doy dinero para caridad no relacionada con la religión	1 2 3 4 5 6 7
	B3	Doy dinero a la gente que pide en la calle	1 2 3 4 5 6 7

Cuadro 14.2.: Estadísticos descriptivos para la subescala "ayuda a los conocidos"

```
Item statistics
    n raw.r std.r  r.cor r.drop mean  sd
A1 50  0.74  0.79  0.799  0.462  5.2 1.4
A2 50  0.67  0.73  0.568  0.433  5.4 1.1
A3 50  0.74  0.78  0.756  0.501  5.5 1.2
B1 50  0.48  0.35 -0.022 -0.037  3.6 1.8
```

Antes de analizar el α de Cronbach, basta ver los estadísticos descriptivos (cuadro 14.2) para darse cuenta de que algo va mal. No solo la media del ítem B1 —el que no debería estar ahí— es claramente inferior (3,6) a las medias del resto de ítems que están alrededor de 5, sino que la correlación de este ítem con respecto a la escala sin ese ítem es negativa (`r.drop` = –0,037).

Si nos fijamos en el α de Cronbach, la salida del cuadro 14.3 nos ofrece tanto la versión estándar del mismo `raw_alpha` que reflejábamos en la expresión (14.1), como la basada en correlaciones que se correspondía con la fórmula de Spearman-Brown (14.2) o `std.alpha`. Cuando las unidades de medida de los ítems son las mismas, como es el caso, los resultados son similares pero cuando son distintas debemos fijarnos en la basada en correlaciones. En ambos casos vemos que los niveles no son aceptables.

Para ver como el α de Cronbach nos puede ayudar a mejorar la fiabilidad de la escala basta fijarnos en el cuadro 14.4, que nos informa de cuál sería el α de Cronbach si eliminámos cada uno de los ítems que forman la escala. Partimos de un α de Cronbach de 0,49 como veíamos en el cuadro anterior. Cuando eliminamos un ítem que sí corresponde a la escala, ese α de Cronbach empeora (por ejemplo, baja a 0,33 si eliminamos A1). Pero cuando llegamos al ítem que no debería estar ahí, el B1, eliminarlo hace que el α de Cronbach suba a 0,77, que es un valor razonable para el conjunto de la escala que, en

Estadísticos e-Books & Papers

Cuadro 14.3.: α de Cronbach para la subescala "ayuda a los conocidos"

```
 raw_alpha std.alpha G6(smc) average_r S/N  ase mean   sd
      0.49      0.58    0.63      0.25  1.4 0.13  4.9 0.88

 lower alpha upper     95% confidence boundaries
  0.24 0.49 0.74
```

Cuadro 14.4.: α de Cronbach para la subescala "ayuda a los conocidos" cuando se elimina cada uno de los ítems

```
Reliability if an item is dropped:
   raw_alpha std.alpha G6(smc) average_r  S/N alpha se
A1      0.24      0.33    0.30      0.14 0.48    0.188
A2      0.32      0.42    0.53      0.19 0.73    0.182
A3      0.24      0.34    0.37      0.15 0.52    0.195
B1      0.78      0.77    0.73      0.53 3.41    0.053
```

ese caso, sí estaría formada por los tres ítems que le corresponden. Esta es la utilidad del α de Cronbach, ayudar a detectar ítems poco correlacionados con el resto de elementos de la escala y, por ello, con pocas probabilidades de estar causados por la misma variable latente que se supone que han de medir.

Calculado ahora de manera correcta a los dos factores implicados, vemos que en los dos casos las escalas superan el nivel mínimo de 0,70.

```
myvars1 <- c("A1", "A2", "A3")
subescala.conocidos <- datos[myvars1]
myvars2 <- c("B1", "B2", "B3")
subescala.donaciones <- datos[myvars2]
alpha(subescala.conocidos,check.keys = FALSE)
alpha(subescala.donaciones,check.keys = FALSE)
```

14.3.2. Fiabilidad compuesta

El α de Cronbach tiene algunas propiedades no deseables, como su dependencia del número de ítems que forman la escala. Si no hay correlación entre los términos de error, el α de Cronbach infraestimará la fiabilidad de la escala (Raykov, 1997, 2001). Sin embargo, para nuestro caso, su principal limitación procede del hecho de que no tiene en cuenta el conjunto del instrumento de medida. Cuando queremos estimar un modelo estructural y efectuamos un CFA lo que nos interesa no es el análisis de cada escala por separado, sino del conjunto del instrumento de medida. Pero, como hemos visto en la subsección anterior, el

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 14.5.: α de Cronbach para las dos subescalas correctamente calculadas

```
Call: alpha(x = subescala.conocidos, check.keys = FALSE)

  raw_alpha std.alpha G6(smc) average_r S/N   ase mean sd
      0.78      0.77    0.73      0.53  3.4 0.053  5.4  1

lower alpha upper     95% confidence boundaries
 0.67 0.78 0.88


Reliability analysis

Call: alpha(x = subescala.donaciones, check.keys = FALSE)

  raw_alpha std.alpha G6(smc) average_r S/N   ase mean  sd
       0.8       0.8    0.74      0.57    4 0.048  3.7 1.4

 lower alpha upper     95% confidence boundaries
  0.7 0.8 0.89
```

cálculo del α de Cronbach lo hemos realizado con la matriz de datos individual y escogiendo las variables de la escala que analizamos, sin que el resto de variables juegue ningún papel.

Para solventar esta limitación se han desarrollado indicadores de fiabilidad que sí que tienen en cuenta el conjunto del instrumento de medida (Raykov, 2004) dado que se calculan a partir de las cargas factoriales de la estimación del CFA. Al investigador le sorprende que un indicador que se calcula para cada uno de los factores del modelo tenga en cuenta el conjunto del instrumento, pero basta recordar cuando explicábamos en el tema anterior el proceso de estimación del CFA que cada parámetro —y las cargas no son una excepción— se estima en el ajuste simultáneo de todos los elementos de las matrices de varianzas y covarianzas muestral y teórica, por lo que cada estimación de un parámetro influye y está influida por la del resto de parámetros.

Fornell y Larcker (1981) proponen el cálculo de la fiabilidad compuesta (CR) bajo la siguiente expresión:

$$CR_i = \frac{\left(\sum_i \lambda_{ij}\right)^2}{\left(\sum_i \lambda_{ij}\right)^2 + \sum_j var\left(\varepsilon_{ij}\right)} \tag{14.3}$$

donde toda la notación es conocida del tema anterior, pero fundamentalmente λ recoge la estimación de las cargas estandarizadas, mientras que $var\left(\varepsilon_{ij}\right)$ es la varianza del término de error de cada indicador. Dado que estimamos fijando a 1 la varianza del factor, la varianza del término de error se puede calcular como 1 (varianza del factor) menos la varianza que el factor explica del indicador λ^2, esto es:

$$var\left(\varepsilon_{ij}\right) = 1 - \lambda_{ij}^2 \tag{14.4}$$

Estadísticos e-Books & Papers

La lógica de la CR es idéntica a la del α de Cronbach. La fiabilidad de una escala sería la parte de la varianza del total que se debe al efecto del factor común, la carga estandarizada λ es la correlación entre el factor y el indicador, mientras que su cuadrado λ^2 es la varianza del indicador explicada por el factor. La expresión (14.3) tiene en el numerador la suma de las cargas que luego eleva al cuadrado —varianza explicada por el factor— y en el denominador ese mismo valor más la varianza de los errores —varianza total— luego la ratio es idéntica conceptualmente al α de Cronbach solo que calculada con las cargas estandarizadas que implican todo el instrumento de medida.

Ilustraremos su cálculo para el caso 14.1 para lo cual, lógicamente, hay que calcular el CFA del mismo, tal y como abordamos en el capítulo anterior. En esta ejemplificación, lógicamente, no realizaremos la asignación incorrecta del indicador B1 al factor 1. La sintaxis de `lavaan` equivalente a la del capítulo 13 y que por ello no necesita de especial aclaración sería:

```
modelo.cfa <-
'
# Modelo de medida
conocidos   =~ A1+A2+A3
donaciones  =~ B1+B2+B3

#Varianzas de los factores
conocidos~~conocidos
donaciones~~donaciones

#Covarianzas
conocidos~~donaciones

#Varianzas de los terminos de error
A1~~A1
A2~~A2
A3~~A3
B1~~B1
B2~~B2
B3~~B3
'
#Estimacion del modelo
fit <- lavaan(modelo.cfa, data=datos, std.lv=TRUE, mimic="eqs",
estimator="ML", verbose=TRUE, warn=TRUE)
```

El cuadro 14.6 nos ofrece la salida con las cargas estandarizadas que, como ya hemos visto, es lo único necesario para el cálculo manual de la CR. Con ello efectuamos los cálculos que se ilustran en el cuadro 14.7, lo que nos permite obtener los CR de los dos factores:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 14.6.: Cargas estandarizadas del CFA

```
Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  conocidos =~
    A1                1.344    0.211    6.355    0.000    1.344    0.963
    A2                0.569    0.161    3.534    0.000    0.569    0.514
    A3                0.901    0.180    5.004    0.000    0.901    0.741
  donaciones =~
    B1                1.697    0.253    6.713    0.000    1.697    0.947
    B2                1.095    0.237    4.620    0.000    1.095    0.656
    B3                1.045    0.221    4.735    0.000    1.045    0.672
```

Cuadro 14.7.: Cálculo de la fiabilidad compuesta CR

Factor	Item	λ	λ^2	$1-\lambda^2$	$\sum\lambda$	$\sum\lambda^2$	$\sum(1-\lambda^2)$
Ayuda a conocidos	A1	0.963	0,927	0,073	2,218	1,741	1,259
	A2	0,514	0,264	0,736			
	A3	0,741	0,549	0,451			
Donaciones económicas	B1	0,947	0,897	0,103	2,275	1,779	1,221
	B2	0,656	0,430	0,570			
	B3	0,672	0,452	0,548			

$$CR_{conocidos} = \frac{\left(\sum_i \lambda_{ij}\right)^2}{\left(\sum_i \lambda_{ij}\right)^2 + \sum_j var\left(\varepsilon_{ij}\right)} = \frac{(2{,}218)^2}{(2{,}218)^2 + 1{,}259} = 0{,}796$$

$$CR_{donaciones} = \frac{(2{,}275)^2}{(2{,}275)^2 + 1{,}221} = 0{,}809$$

El trabajar con `R` nos permite no tener que realizar estos cálculos de manera manual, y sí programarlos para añadirlos a la sintaxis y que se calculen automáticamente. Aunque existen muchas aproximaciones para hacerlo, presentamos la que solemos aplicar nosotros.

Si se revisa la sintaxis del CFA, veremos que la información de la estimación se ha guardado en un objeto de `R` llamado `fit`. La matriz `lambda` de ese objeto tiene las cargas factoriales, por lo que extraemos el vector de cargas del primer factor, que llamaremos `l_f1`, y las del segundo factor, que denominaremos `l_f2`. El investigador solo debe saber que los indicadores del primer factor son las tres primeras columnas de la matriz `[1:3,1]` y los del segundo factor las tres últimas `[4:6,2]`. También que sean las cargas estandarizadas «`std`».

```
l_f1<-lavInspect(fit,"std")$lambda[1:3,1]
l_f2<-lavInspect(fit,"std")$lambda[4:6,2]
```


Estadísticos e-Books & Papers

Cuadro 14.8.: Fiabilidad compuesta CR

```
cr_conocidos cr_donaciones
   0.7961761     0.8090004
```

Para calcular el CR nos hacen falta las varianzas de los errores, que denominaremos `v_f1` y `v_f2` y que están en la diagonal de la matriz que en `fit lavaan` ha denominado `theta`.

```
v_f1<-diag(lavInspect(fit,"std")$theta)[1:3]
v_f2<-diag(lavInspect(fit,"std")$theta)[4:6]
```

Solo falta calcular el CR aplicando la expresión (14.3):

```
cr_f1<-sum(l_f1)^2/(sum(l_f1)^2+sum(v_f1))
cr_f2<-sum(l_f2)^2/(sum(l_f2)^2+sum(v_f2))
```

Si se quiere etiquetar la salida, ponemos los dos CR en un vector, lo etiquetamos y lo imprimimos. El resultado aparece en el cuadro 14.8 y vemos que coincide con los cálculos que habíamos efectuado a mano.

```
cr<-c(cr_f1,cr_f2)
names(cr)=c("cr_conocidos", "cr_donaciones")
print(cr)
```

Los valores exigibles para CR son los mismos que para el α de Cronbach. Se considera que en un planteamiento exploratorio son aceptables entre 0,60-0,70, en etapas avanzadas de la investigación entre 0,70-0,90 se pueden considerar adecuados (Nunnally y Bernstein, 1994). No se consideran deseables valores superiores a 0,90 porque implicaría que todos los indicadores están midiendo exactamente el mismo fenómeno de manera redundante, por ejemplo cuando incorporamos ítems que tienen enunciados que son prácticamente idénticos. Esta práctica de incremento artificial de la fiabilidad no se recomienda porque, como veremos, afecta a la validez de contenido (Rossiter, 2002) y acrecienta la correlación entre los términos de error (Drolet y Morrison, 2001; Hayduk y Littvay, 2012)

14.4. Análisis de la validez del instrumento de medida

¿Cuándo es válida una escala? Pueden darse muchas definiciones, pero, al final, probablemente la más sencilla sea la más ajustada: cuando lo que está midiendo realmente es la variable latente que se supone que tiene que medir. Esta

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

definición puede sofisticarse mucho más, pero esa es la esencia de la validez. Sarabia y Sánchez (1999) reproducen la definición de Bohrnstedt (1976) que apunta en esta misma línea: "validez es el grado en que un instrumento mide el concepto bajo estudio". La validez, sin embargo, es un concepto poliédrico y tiene diversas dimensiones que deben explicarse y analizarse por separado.

14.4.1. Validez de contenido

La validez de contenido representa el grado en que los indicadores de un instrumento de medida son relevantes y representativos de la variable latente que quieren medir (Haynes *et al.*, 1995). La representatividad hace referencia a en qué medida todos los matices, todas las dimensiones del concepto han sido recogidas por ítems de la escala, es decir, es una medida de la coherencia del dominio teórico del concepto que se ha de reflejar en los enunciados elegidos para los ítems, los formatos de respuesta y las instrucciones de contestación (Haynes *et al.*, 1999, 1995; Netemeyer *et al.*, 2002; Robinson *et al.*, 1991).

La validez de contenido, por tanto, parece evidente que es más un factor a cuidar en el proceso de desarrollo de la escala que en su aplicación. Es necesario generar un conjunto de ítems conectados con el concepto y que deberían recoger todas las áreas, todos los matices del mismo (Clark y Watson, 1995). Probablemente algunos de esos ítems se eliminarán posteriormente en el proceso de aplicación de la escala siendo necesario evaluar si estas eliminaciones ponen en peligro o no la validez de contenido.

Este proceso de validación será necesariamente cualitativo, con jueces que deberán recibir la definición y valorar en qué medida cada ítem está o no conectado con el concepto hasta que se llegue a un conjunto donde el acuerdo sea sólido.

En nuestras investigaciones, salvo que su objetivo sea precisamente la creación de una escala, lo habitual es asumir que las escalas que hemos tomado de la literatura, en la medida en que han sido evaluadas por pares y publicadas, disfrutan de validez de contenido. El único cuidado que habrá que tener es no tomar ciertas decisiones en las etapas posteriores de validación, por ejemplo eliminar un excesivo número de ítems, que pongan en peligro la validez de contenido de partida.

14.4.2. Validez convergente

La validez convergente existe cuando distintas medidas (indicadores) de un mismo concepto (factor) están correlacionadas de manera intensa entre sí, "convergiendo" o "compartiendo" una proporción elevada de la varianza, puesto que la varianza que no compartan por la causa común que ha de hacerles covariar —el factor— será varianza residual, no explicada por el factor.

Hay dos aproximaciones para buscar evidencias de validez convergente. La primera es que las **cargas factoriales que unen el factor con el indicador sean altas y significativas**. Recordemos que la carga factorial es la correlación

Estadísticos e-Books & Papers

entre el factor y el indicador. Si un indicador no está fuertemente correlacionado con el factor que lo genera, no parece un indicador válido. Pensemos en el caso extremo de que esa carga factorial sea nula. Si el indicador no guarda ninguna relación con el factor, difícilmente puede reflejar su contenido. El nivel mínimo que se suele elegir para el valor de esas cargas es 0,707 que se suele redondear a 0,70. Este número no es un número mágico ni arbitrario, la única propiedad que tiene es que su cuadrado es 0,50. Recordemos que el cuadrado de la carga —la comunalidad— es la varianza del indicador explicada por el factor, por lo que al exigir ese 0,70 estamos exigiéndole que al menos la mitad de la varianza del indicador se deba al factor y no sea residual.

A veces este criterio se relaja, por ejemplo Bagozzi y Yi (1988) plantean un valor mínimo de 0,60. Más que discutir sobre el valor exacto, el investigador ha de ser consciente de que la eliminación automática de cualquier indicador con una carga inferior a 0,70 puede descapitalizar la validez de contenido. Si no comprobamos esta validez porque hemos tomado prestada la escala de un trabajo ya publicado, pero en el proceso de validación eliminamos un número excesivo de indicadores, la asunción de que la validez se mantiene puede ser cuestionada. Aunque en el contexto de PLS-SEM y no de los modelos basados en covarianzas que estamos analizando, Hair *et al.* (2014b) hacen una propuesta que pretende equilibrar ambas facetas de la validez. Si la carga es inferior a 0,40 debe eliminarse, si está comprendida entre 0,40-0,70, hay que preguntarse si la fiabilidad compuesta CR y otro indicador que analizaremos inmediatamente, la varianza extraída promedio o AVE, están por debajo de sus límites. Si lo están y eliminar el indicador los hace pasar el límite aceptable, se eliminan, pero, si eliminándolos, esto no ocurre, se mantienen. La lógica es no añadir un problema de validez de contenido a un problema de validez convergente o de fiabilidad que no se va a resolver con esa eliminación.

Las cargas factoriales estandarizadas y su significatividad ya se obtuvieron para el caso mediante la CFA y se mostraban en el cuadro 14.6. Vemos que todas las cargas estandarizadas (`std.all`) son significativas. Todas son superiores a 0,40 pero no todas a 0,70. En este momento no nos planteamos eliminar ninguna porque no tenemos un problema de fiabilidad (α de Cronbach y CR). Deberemos ver el indicador de validez convergente que nos resta, el AVE, para ver si está comprometido y eliminando algún indicador mejora. Si no está comprometido, no eliminaríamos ningún indicador.

Como señalamos el segundo indicador para evaluar la validez convergente es la **varianza extraída promedio (AVE)**. Es un indicador cuya lógica es muy directa. Se define del siguiente modo:

$$AVE_i = \frac{\sum_j \lambda_{ij}^2}{\sum_j \lambda_{ij}^2 + \sum_j var\left(\varepsilon_{ij}\right)} = \frac{\sum_j \lambda_{ij}^2}{k} \tag{14.5}$$

donde la notación es la misma que para la CR y k es el número de indicadores del factor. Dado que λ^2 es la varianza del indicador j explicada por la variable latente i, entonces la AVE es la media de las varianzas de los indicadores que

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 14.9.: Varianza extraída promedio (AVE)

```
ave_conocidos ave_donaciones
    0.5802198      0.5927624
```

explica el factor que están midiendo. Dicho de otra manera, calculamos la media de lo que el factor explica de cada indicador. ¿Qué nivel mínimo habrá que exigirle? Pues que al menos el factor explique, en promedio, el 50 % de la varianza de los indicadores. Valores inferiores implicarían que la varianza residual es superior a la varianza explicada por el factor.

Aplicado a nuestro caso, toda la información parea el cálculo ya estaba en el cuadro 14.7. Aplicando la expresión (14.5) a esos datos, las varianzas extraídas promedio serían:

$$AVE_{conocidos} = \frac{\sum_j \lambda_{ij}^2}{\sum_j \lambda_{ij}^2 + \sum_j var\left(\varepsilon_{ij}\right)} = \frac{1{,}741}{1{,}741 + 1{,}259} = 0{,}5802$$

$$AVE_{donaciones} = \frac{1{,}779}{1{,}779 + 1{,}221} = 0{,}5928$$

por lo que comprobamos que las cargas inferiores a 0,7 tampoco estaban provocando que el AVE cayera por debajo de su nivel crítico de 0,5 confirmando la lógica de mantener esos indicadores.

De nuevo, al trabajar con `R` no es necesario realizar el cálculo manual pues la obtención del AVE es inmediata. Al calcular el CR ya extrajimos de la matriz **`lambda`** las cargas estandarizadas y de la matriz **`theta`** las varianzas residuales, por lo que solo es necesario realizar el cálculo que recoge la expresión (14.5).

```
ave_f1<-(sum(l_f1^2))/((sum(l_f1^2))+sum(v_f1))
ave_f2<-(sum(l_f2^2))/((sum(l_f2^2))+sum(v_f2))
```

Si se quiere etiquetar la salida, como con la CR, ponemos los dos AVE en un vector, lo etiquetamos y lo imprimimos. El resultado aparece en el cuadro 14.9 y vemos que, una vez más, coincide con los cálculos que habíamos efectuado a mano.

```
ave<-c(ave_f1,ave_f2)
names(ave)=c("ave_conocidos", "ave_donaciones")
print(ave)
```


Estadísticos e-Books & Papers

Cuadro 14.10.: Estimación de la correlación entre los factores

```
Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  conocidos ~~
    donaciones       -0.062    0.154   -0.401    0.689   -0.062   -0.062
```

14.4.3. Validez discriminante

Es obvio que dos variables latentes pueden estar correlacionadas entre sí. De hecho, si no lo están, cuando incorporemos el modelo estructural y planteemos una hipótesis que señale que esas variables latentes están unidas por una relación estructural, difícilmente podrá ser significativa si en el CFA no había un cierto nivel de correlación.

El problema es cuando esa correlación supera un determinado umbral. Si para medir el factor "ayuda a los conocidos" hemos usado distintos indicadores de los que hemos utilizado para medir las "donaciones económicas", si la correlación —en un caso extremo— entre ambas fuera 1, sería un indicador de que las escalas elegidas no han tenido capacidad para separar el contenido distinto de los dos factores. Dicho de otro modo, los factores pueden estar correlacionados, pero no tanto que se ponga en duda la capacidad discriminante de las escalas utilizadas para medirlos.

Hay distintos procedimientos para evaluar la validez discriminante. Veremos los tres más habituales: el test del intervalo de confianza, el criterio de Fornell y Larcker (1981), y la ratio HTMT. Todos ellos son sencillos de comprender si el concepto de validez discriminante ha quedado claro.

A. Test del intervalo de confianza

En la estimación del CFA se ha estimado la covarianza entre todos los factores —dos en nuestro caso, por lo tanto una única covarianza— covarianza que, al estar calculada entre dos factores cuya varianza se ha fijado a 1 para establecer la escala de medida —véase el capítulo 13— coincide con la correlación entre factores. El cuadro 14.10 ofrece el resultado de esta estimación que se ha obtenido con la sintaxis ya presentada con anterioridad. Que la covarianza coincide con la correlación se observa inmediatamente al comprobar que la estimación de la covarianza (`Estimate`) coincide con su estimación estandarizada (`Std.all`).

Anderson y Gerbing (1988) proponen la siguiente prueba para evaluar la validez discriminante. Dado que la estimación de la correlación es una estimación puntual que tiene un error estándar (SE), proponen construir un intervalo de confianza alrededor de la estimación de $\pm 2SE$. Si ese intervalo de confianza contiene al 1, querrá decir que la correlación entre los dos factores sea 1 es tan probable como que tome el valor de la estimación puntual, poniendo de manifiesto que las escalas no han sido capaces de separar los dos conceptos. Si llamamos ρ_{cd} a la correlación entre los dos factores, el intervalo de confianza

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

con los datos del cuadro 14.10 será:

$$\rho_{cd} \pm 2SE = -0{,}062 \pm 2 \times 0{,}154 \longrightarrow [-0,370; 0{,}246]$$

intervalo que contiene al 0 —no es estadísticamente significativo como se observaba en el cuadro 14.10— pero no contiene al 1, lo que no evidencia problemas de validez discriminante.

B. Criterio de Fornell y Larcker (1981)

Cuando tenemos dos factores nos encontraremos con una correlación y dos AVE. Recordemos de nuevo el concepto de varianza extraída promedio (AVE). El AVE de un factor es la parte de la varianza de los indicadores que está explicada por el factor. Pero ¿qué es la correlación?, o, mejor aún, qué es el ¿cuadrado de la correlación? Si una carga es la correlación entre factor e indicador, la carga al cuadrado es la varianza del indicador explicada por el factor, entonces el cuadrado de la correlación entre los factores será la parte de la varianza de los indicadores del factor 1 que está explicada por los indicadores del factor 2 a través de la correlación entre los factores.

Parece lógico que si los indicadores del factor 2 explican más de los indicadores del factor 1 (cuadrado de la correlación entre los factores) que el propio factor 1 (AVE del factor 1) tenemos un problema de validez discriminante. Lo mismo ocurrirá con los indicadores del factor 1 respecto a los del 2. Fornell y Larcker (1981) operativizan este criterio señalando que no habrá problemas de validez discriminante cuando los AVE de los factores implicados sean mayores que el cuadrado de la correlación entre esos factores, es decir:

$$\begin{array}{c} AVE_i > \rho_{ij}^2 \\ AVE_j > \rho_{ij}^2 \end{array} \tag{14.6}$$

En nuestro caso tenemos dos factores cuya correlación, tal como se observa en el cuadro 14.10, es $\rho = -0{,}062 \longrightarrow \rho^2 = 0{,}004$. Los AVE los tenemos calculados en el cuadro 14.9. Y vemos que es fácil comprobar que no hay ningún problema de validez discriminante en la medida en que:

$$\begin{array}{c} AVE_{conocidos} = 0{,}580 > \rho_{cd}^2 = 0{,}004 \\ AVE_{donaciones} = 0{,}593 > \rho_{cd}^2 = 0{,}004 \end{array}$$

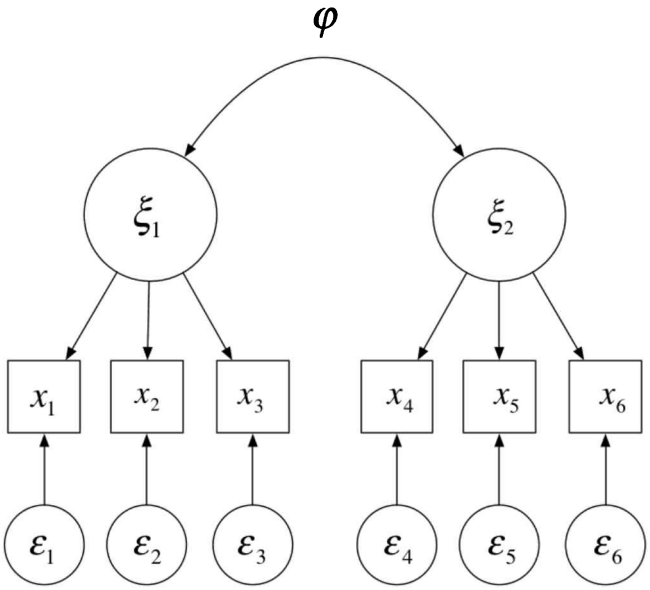
Naturalmente, cuando tengamos más de dos factores tendremos más correlaciones y habrá que comprobar que el criterio aplica a todas ellas.

C. Criterio de la ratio HTMT

Recientemente, Henseler *et al.* (2014) han propuesto un procedimiento complementario para evaluar la validez discriminante que, en el fondo, está basado en la misma lógica. Bien cierto es que está propuesto para los modelos basados en varianzas que veremos en el tema 16 y por ello lo repetiremos en ese tema,

Estadísticos e-Books & Papers

Figura 14.1.: Modelo de dos factores para ilustrar la el criterio de la ratio HTMT



Fuente: Henseler *et al.* (2014)

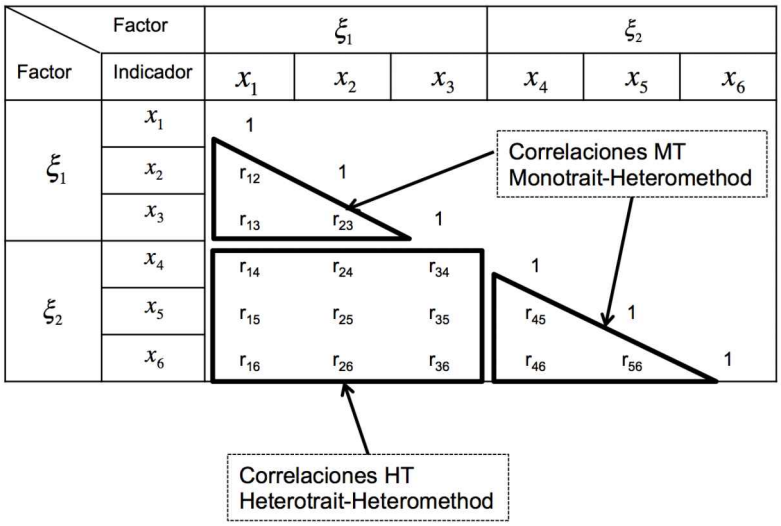
sin embargo la lógica es trasladabe. Si partimos de un ejemplo de dos factores como el de la figura 14.1, en el que los factores, ξ_1 y ξ_2 son medidos a partir de tres indicadores cada uno de ellos, podemos construir la matriz de correlaciones entre esos indicadores que está recogida en la figura 14.2. Esta matriz recoge dos tipos de correlaciones, por un lado, las correlaciones de los indicadores de un mismo factor, que Henseler *et al.* (2014) denominan *monotrait-heteromethod correlations*, cuya media denominaremos MT y, por otro lado, las correlaciones entre los ítems de un factor con los del otro factor, que denominan *heterotrait-heteromethod* y cuya media denominaremos HT. De una manera simplificada la ratio $HTMT$ sería la ratio HT/MT. Parece lógico que si la media de las correlaciones entre los indicadores de dos constructos distintos es más grande que la que tienen los indicadores del constructo al que corresponden, entonces tendremos un problema de validez discriminante (en ese caso $HT/MT > 1$). Esto será poco habitual, lo normal es que las medias de las correlaciones entre los indicadores de los factores a los que van asociados sean más altas, la cuestión es cuán parecidas pueden ser a las asociadas a factores distintos. El criterio que se propone es el de Gold *et al.* (2001) quienes plantean que la ratio $HT/MT < 0{,}90$ para cada par de factores, de no ser así podemos estar ante un problema de validez discriminante.

El criterio está implementado por la funcion `htmt{semTools}` y su petición es sencilla como anexo a la sintaxis que hemos planteado con anterioridad para

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 14.2.: Matriz de correlaciones para ilustrar la ratio HTMT



Fuente: Henseler *et al.* (2014)

Cuadro 14.11.: Criterio de la ratio HTMT

```
> htmt(datos,modelo.cfa)
           concds doncns
conocidos  1.000
donaciones 0.169  1.000
```

la estimación del CFA:

```
htmt(datos,modelo.cfa)
```

El cuadro 14.11 nos muestra que la ratio obtenida es de 0,169 muy lejos del valor crítico de 0,90 lo que confirma la inexistencia de problemas de validez discriminante en el caso que estamos resolviendo.

Es importante señalar que los problemas de validez discriminante son de muy difícil solución y es importante que el investigador los anticipe cuando está eligiendo las escalas que va a utilizar en su modelo asegurándose de que sus enunciados difieren significativamente entre sí. Si una vez realizado el trabajo de campo, el problema aflora, lo único que el investigador puede hacer es analizar las correlaciones cruzadas entre los ítems de los factores que tienen el problema y eliminar uno de los dos más correlacionados, que muy probablemente tendrán cargas significativas y altas con el problema de afectación a la validez convergente y probablemente a la de contenido.

Estadísticos e-Books & Papers

14.4.4. Validez nomológica

El último paso en el análisis de la validez de las escalas utilizadas es establecer su validez nomológica. Podemos afirmar la existencia de validez nomológica cuando los valores de un constructo elaborado con las escalas validadas están relacionados con los de otro constructo apoyando empíricamente relaciones teóricas. Es decir, dado que no medimos por medir, sino para evaluar la plausibilidad de relaciones estructurales entre los factores medidos, es necesario que esas relaciones afloren al estimar el modelo estructural.

Este paso, dada la definición, ya no puede efectuarse sobre el modelo de medida, es necesario efectuarlo sobre el modelo estructural que incorpora esas relaciones. No lo aplicaremos al caso actual, pero sí al caso final que presentaremos en la sección 14.6 cuando estimemos su correspondiente modelo estructural en el tema 15.

14.5. Un ejemplo completo de evaluación del instrumento de medida

Caso 13.2 Determinantes de la aceptación de la publicidad en el móvil
El caso que vamos a desarrollar era el que presentábamos en el capítulo anterior en la figura 13.1 y que se corresponde con el trabajo de Aldás *et al.* (2013). Remitimos de nuevo a este trabajo para la justificación teórica de las relaciones pero, básicamente, lo que plantea es que la aceptación de la publicidad en el móvil viene condicionada porque el usuario perciba una utilidad en la misma y que el entretenimiento que genere su creatividad sea capaz de minorar la irritación que la intrusión pueda causar. Para estimar el modelo estructural que aparece en la figura 13.1 (capítulo 15) es necesario, previamente y como hemos señalado reiteradamente, estimar un CFA para, con la información que proporciona (capítulo 13), evaluar la fiabilidad y validez del instrumento de medida con esa información, que es lo que hemos de realizar en este tema.

La figura 14.3 ilustra el CFA que se corresponde con el instrumento de medida del modelo representado en la figura 13.1. Por comodidad de lectura repetiremos en este capítulo los cuadros del capítulo 13 que tienen la información necesaria para la evaluación de la fiabilidad y validez de dicho instrumento de medida. El caso terminará con una propuesta de cuadros que sinteticen la información para su publicación en un artículo.

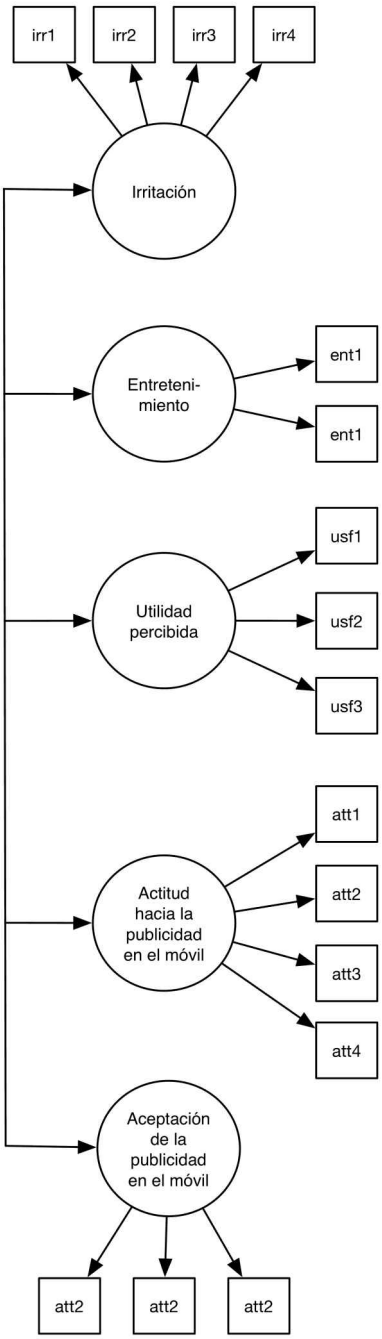
La sintaxis de `lavaan` necesaria para la estimación del CFA se presentó en el tema 13. De la ejecución de esa sintaxis se obtienen las cargas que mostramos en el cuadro 14.12.

A lo largo del tema hemos realizado una presentación "didáctica" de la comprobación de la fiabilidad y la validez. En el caso que nos ocupa haremos una presentación operativa. No comenzaremos analizando la fiabilidad, sino la validez convergente, porque si se eliminan aquellas cargas que no superen el nivel mínimo, esto mejorará la fiabilidad simple y compuesta, por lo que no vale la

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 14.3.: CFA por estimar en el caso 13.2



Fuente: Aldás *et al.* (2013)

Estadísticos e-Books & Papers

Cuadro 14.12.: Estimación de las cargas factoriales

```
Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  actitud =~
    att1              0.786    0.044   17.993    0.000    0.786    0.810
    att2              0.828    0.041   20.351    0.000    0.828    0.877
    att3              0.935    0.049   19.018    0.000    0.935    0.840
    att4              0.868    0.043   20.145    0.000    0.868    0.872
  entretenimiento =~
    ent1              0.813    0.050   16.422    0.000    0.813    0.850
    ent2              0.801    0.046   17.344    0.000    0.801    0.895
  utilidad =~
    usf1              0.954    0.054   17.633    0.000    0.954    0.835
    usf2              0.805    0.044   18.319    0.000    0.805    0.860
    usf3              0.706    0.056   12.563    0.000    0.706    0.639
  irritacion =~
    irr1              1.003    0.072   13.926    0.000    1.003    0.698
    irr2              0.972    0.069   13.998    0.000    0.972    0.701
    irr3              1.030    0.062   16.619    0.000    1.030    0.798
    irr4              1.033    0.062   16.538    0.000    1.033    0.795
  aceptacion =~
    acc1              0.818    0.068   12.034    0.000    0.818    0.636
    acc2              0.794    0.056   14.177    0.000    0.794    0.731
    acc3              0.836    0.051   16.313    0.000    0.836    0.823
```

pena calcularlas hasta que no se tengan las cargas definitivas. Por tanto, comenzamos evaluando la **validez convergente** del modelo de medida. Del cuadro 14.12 vemos que solo hay tres cargas inferiores a 0,7, usf3, irr1 y acc1. En los tres casos están tan cercanas a ese valor que no vamos a plantearnos eliminar ninguna de ellas a no ser que, cuando calculemos la CR y el AVE, sus niveles estén por debajo de los niveles críticos. Para el cálculo de CR y AVE, generalizamos la propuesta que hemos realizado de extraer los datos directamente del objeto `fit` que recoge los resultados de la estimación. Para dar más alternativas al investigador vamos, también, a calcular por este procedimiento de extracción de información el α de Cronbach mediante la fórmula de Spearman-Brown que está en la ecuación (14.2). Comenzando por este último calculamos las correlaciones entre los indicadores de cada factor sacando la matriz de covarianzas de `lavaan`:

```
matriz_cov_muestral<-lavInspect(fit,"sampstat")$cov
```

Esta matriz se convierte en una matriz de correlaciones:

```
matriz_cor_muestral<-cov2cor(matriz_cov_muestral)
```

El investigador solo ha de recordar cuántos indicadores son y qué lugar ocupan en la matriz de correlaciones, por ejemplo, que el primero tiene cuatro indicadores [1:4], el segundo, dos [5:6], etc. Creamos un vector que tiene las

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

correlaciones entre los indicadores de cada factor por separado extrayéndolos de la matriz de correlaciones que acabamos de calcular:

```
cor_f1<-matriz_cor_muestral[1:4,1:4]
cor_f2<-matriz_cor_muestral[5:6,5:6]
cor_f3<-matriz_cor_muestral[7:9,7:9]
cor_f4<-matriz_cor_muestral[10:13,10:13]
cor_f5<-matriz_cor_muestral[14:16,14:16]
```

Ahora solo hay que calcular las medias de las correlaciones como las medias de los elementos de esos vectores:

```
mean_cor_f1<-mean(cor_f1[lower.tri(cor_f1)])
mean_cor_f2<-mean(cor_f2[lower.tri(cor_f2)])
mean_cor_f3<-mean(cor_f3[lower.tri(cor_f3)])
mean_cor_f4<-mean(cor_f4[lower.tri(cor_f4)])
mean_cor_f5<-mean(cor_f5[lower.tri(cor_f5)])
```

Y ahora ya podemos aplicar la expresión (14.2) para calcular el α de Cronbach. El investigador ha de recordar siempre cuántos indicadores tiene cada factor (k) porque forman parte de la expresión:

```
ca_f1<-4*mean_cor_f1/(1+(4-1)*mean_cor_f1)
ca_f2<-2*mean_cor_f2/(1+(2-1)*mean_cor_f2)
ca_f3<-3*mean_cor_f3/(1+(3-1)*mean_cor_f3)
ca_f4<-4*mean_cor_f4/(1+(4-1)*mean_cor_f4)
ca_f5<-3*mean_cor_f5/(1+(3-1)*mean_cor_f5)
```

Para el cálculo del CR y del AVE, generalizamos el procedimiento ya presentado en el caso anterior. Extraemos de **lavaan** cargas y varianzas residuales de las matrices **lambda** y **theta**:

```
l_f1<-lavInspect(fit,"std")$lambda[1:4,1]
l_f2<-lavInspect(fit,"std")$lambda[5:6,2]
l_f3<-lavInspect(fit,"std")$lambda[7:9,3]
l_f4<-lavInspect(fit,"std")$lambda[10:13,4]
l_f5<-lavInspect(fit,"std")$lambda[14:16,5]

v_f1<-diag(lavInspect(fit,"std")$theta)[1:4]
v_f2<-diag(lavInspect(fit,"std")$theta)[5:6]
v_f3<-diag(lavInspect(fit,"std")$theta)[7:9]
v_f4<-diag(lavInspect(fit,"std")$theta)[10:13]
v_f5<-diag(lavInspect(fit,"std")$theta)[14:16]
```

A continuación, aplicamos las expresiones (14.3) para CR y (14.5) para el AVE.

Estadísticos e-Books & Papers

Cuadro 14.13.: Cálculo de α de Cronbach, AVE y CR

```
> print(ca)
   ca_att    ca_ent    ca_usf    ca_irr    ca_acc
0.9118552 0.8642338 0.8137834 0.8368166 0.7724637
> print(cr)
   cr_att    cr_ent    cr_usf    cr_irr    cr_acc
0.9124915 0.8648050 0.8252200 0.8364241 0.7760058
> print(ave)
  ave_att   ave_ent   ave_usf   ave_irr   ave_acc
0.7229493 0.7619316 0.6152507 0.5621166 0.5386106
```

```
cr_f1<-sum(l_f1)^2/(sum(l_f1)^2+sum(v_f1))
cr_f2<-sum(l_f2)^2/(sum(l_f2)^2+sum(v_f2))
cr_f3<-sum(l_f3)^2/(sum(l_f3)^2+sum(v_f3))
cr_f4<-sum(l_f4)^2/(sum(l_f4)^2+sum(v_f4))
cr_f5<-sum(l_f5)^2/(sum(l_f5)^2+sum(v_f5))

ave_f1<-(sum(l_f1^2))/((sum(l_f1^2))+sum(v_f1))
ave_f2<-(sum(l_f2^2))/((sum(l_f2^2))+sum(v_f2))
ave_f3<-(sum(l_f3^2))/((sum(l_f3^2))+sum(v_f3))
ave_f4<-(sum(l_f4^2))/((sum(l_f4^2))+sum(v_f4))
ave_f5<-(sum(l_f5^2))/((sum(l_f5^2))+sum(v_f5))
```

Por estética pueden etiquetarse los indicadores y pedir su incorporación a la salida. Los resultados se muestran en el cuadro 14.13. Observamos que el α de Cronbach y el CR son superiores a 0,7 mientras que todos los AVE son superiores a 0,5 no constatándose problema alguno de fiabilidad y tampoco de validez convergente en la medida en que aquellas cargas algo más pequeñas no han afectado a estos indicadores.

```
ca<-c(ca_f1,ca_f2,ca_f3,ca_f4,ca_f5)
names(ca)=c("ca_att", "ca_ent", "ca_usf", "ca_irr", "ca_acc")
cr<-c(cr_f1,cr_f2,cr_f3,cr_f4,cr_f5)
names(cr)=c("cr_att", "cr_ent", "cr_usf", "cr_irr", "cr_acc")
ave<-c(ave_f1,ave_f2,ave_f3,ave_f4,ave_f5)
names(ave)=c("ave_att", "ave_ent", "ave_usf",
"ave_irr", "ave_acc")
print(ca) print(cr) print(ave)
```

Para evaluar la **validez discriminante**, comenzamos con el criterio del intervalo de confianza. La estimación de las correlaciones ya se realizó en el tema anterior y las ofrecemos en el cuadro 14.14 de nuevo:

Con esta información es inmediato calcular los intervalos de confianza del

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 14.14.: Estimación de las correlaciones entre factores

```
Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  actitud ~~
    entretenimient    0.427    0.050    8.482    0.000    0.427    0.427
    utilidad          0.402    0.052    7.770    0.000    0.402    0.402
    irritacion       -0.194    0.059   -3.313    0.001   -0.194   -0.194
    aceptacion        0.498    0.050   10.025    0.000    0.498    0.498
  entretenimiento ~~
    utilidad          0.496    0.049   10.068    0.000    0.496    0.496
    irritacion       -0.180    0.061   -2.970    0.003   -0.180   -0.180
    aceptacion        0.363    0.057    6.336    0.000    0.363    0.363
  utilidad ~~
    irritacion       -0.327    0.057   -5.714    0.000   -0.327   -0.327
    aceptacion        0.480    0.053    9.084    0.000    0.480    0.480
  irritacion ~~
    aceptacion       -0.362    0.058   -6.215    0.000   -0.362   -0.362
```

Cuadro 14.15.: Intervalos de confianza para las correlacione entre los factores

Factores correlacionados		ρ	SE	$\rho - 2SE$	$\rho + 2SE$
Actitud	Entretenimiento	0,427	0,050	0,327	0,527
	Utilidad	0,402	0,052	0,298	0,506
	Irritación	-0,194	0,059	-0,312	-0,076
	Aceptación	0,498	0,050	0,398	0,598
Entretenimiento	Utilidad	0,496	0,049	0,398	0,594
	Irritación	-0,180	0,061	-0,302	-0,058
	Aceptación	0,363	0,057	0,249	0,447
Utilidad	Irritación	-0,327	0,057	-0,441	-0,213
	Aceptación	0,480	0,053	0,374	0,586
Irritación	Aceptación	-0,362	0,058	-0,478	-0,246

cuadro 14.15, donde vemos que ninguno de ellos contiene al 1, apuntando a que no existen problemas de validez discriminante.

El criterio de Fornell y Larcker (1981) implica comparar el cuadrado de las correlaciones que se obtienen directamente del cuadro 14.14 con las varianzas extraídas, que se tienen calculadas en el cuadro 14.13. Por lo tanto se trata más de encontrar una forma de presentar elegantemente los resultados que de realizar cálculo alguno. El cuadro 14.16 ofrece los AVE de los factores en la diagonal tomados del cuadro 14.13 y en el triángulo inferior los cuadrados de las correlaciones del cuadro 14.14. Debe cumplirse que cada AVE sea siempre mayor que el cuadrado de cualquier correlación en la que ese factor esté implicado. Así, por ejemplo, el AVE de F3 (0,616) ha de ser más grande que las correlaciones al cuadrado de F3 con el resto de factores, es decir, la fila (F1:0,162, F2:0,246) y su columna (F4: 0,003) y (F5: 0,003). Como es el caso, confirmando la ausencia de problemas.

Finalmente, el ratio HTMT se obtiene directamente con la ejecución de la

Estadísticos e-Books & Papers

Cuadro 14.16.: Criterio de Fornell y Larcker (1981)

	F1	F2	F3	F4	F5
F1. Actitud	0.723				
F2, Entretenimiento	0,182	0,762			
F3, Utilidad	0,162	0,246	0,616		
F4, Irritación	0,038	0,032	0,003	0,562	
F5, Aceptación	0,248	0,132	0,003	0,003	0,539

AVE en la diagonal, ρ^2 en el triángulo inferior

Cuadro 14.17.: Ratio HTMT

```
                actitd entrtn utildd irrtcn acptcn
actitud         1.000
entretenimiento 0.434  1.000
utilidad        0.433  0.535  1.000
irritacion      0.203  0.172  0.316  1.000
aceptacion      0.486  0.374  0.525  0.377  1.000
```

sintaxis siguiente y los resultados están en el cuadro 14.17, donde vemos que todos los ratios son inferiores a 0,90 (Gold *et al.*, 2001) reafirmando la ausencia de problemas de validez discriminante,

```
htmt(datos,modelo,cfa)
```

Validado el instrumento de medida estamos en condiciones de incorporar la parte estructural, pero eso lo realizaremos en el tema 15. Solo nos resta ofrecer una propuesta de cuadros que sinteticen la información para ser utilizados en una publicación. La forma más sintética que hemos encontrado es utilizar dos cuadros, el cuadro 14.18, que resume la fiabilidad y validez convergente, y el cuadro 14.19, que resume la validez discriminante, Solo hacemos notar que es más habitual ofrecer la raíz cuadrada del AVE en la diagonal para ofrecer las correlaciones en el triángulo inferior ya que es una información muy relevante para valorar, posteriormente, el modelo estructural.

14.6. Guía para el desarrollo de escalas

La última sección que vamos a presentar propone una serie de pasos que todo investigador debería seguir para desarrollar un instrumento de medida (escala) adecuado de una variable latente.

Paso 1. Determinar claramente qué es lo que se quiere medir

Aunque puede parecer que todo investigador sabe qué es lo que quiere medir, muchas veces se encuentra con que sus ideas son más vagas de lo que creía a

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 14.18.: Fiabilidad, consistencia interna y validez convergente del instrumento de medida.

Factor	Indicador	λ	t	CA	CR	AVE
Actitud	att1	0,810**	17,99	0,91	0,91	0,72
	att2	0,877**	20,35			
	att3	0,840**	19,02			
	att4	0,872**	20,15			
Entretenimiento	ent1	0,850**	16,42	0,86	0,86	0,76
	ent2	0,895**	17,34			
Utilidad	usf1	0,835**	17,63	0,81	0,83	0,62
	usf2	0,860**	18,32			
	usf3	0,639**	12,56			
Irritación	irr1	0,698**	13,93	0,84	0,84	0,56
	irr2	0,701**	13,99			
	irr3	0,798**	13,62			
	irr4	0,795**	16,54			
aceptación	acc1	0,636**	12,03	0,77	0,77	0,54
	acc2	0,731**	14,18			
	acc3	0,823**	16,31			

$^{**}p < 0,01$; CA = α de Cronbach; CR = Fiabilidad Compuesta; AVE = Varianza extraída promedio;
$\chi^2(94) = 219,77^{**}$; CFI=0,956; TLI=0,944; RMSEA (90 %)=0,062 (0,051;0,072)

Estadísticos e-Books & Papers

Cuadro 14.19.: Validez discriminante

	F1	F2	F3	F4	F5
F1, Actitud	**0,850**	0,527	0,506	-0,076	0,598
F2, Entretenimiento	0,427	**0,873**	0,594	-0,058	0,477
F3, Utilidad	0,402	0,496	**0,785**	-0,213	0,586
F4, Irritación	-0,194	-0,180	0,057	**0,750**	-0,246
F5, Aceptación	0,498	0,363	0,053	0,058	**0,734**

$\sqrt{AVE}$ en la diagonal, ρ en el triángulo inferior, límite superior del intervalo de confianza en el triángulo superior

la hora de elaborar preguntas. La teoría es la mejor ayuda para esta fase. Una escala solo será correcta si todas las dimensiones del concepto que se quiere medir son conocidas por el investigador, y solo el conocimiento profundo de la literatura puede garantizar que esta fase se desarrolla correctamente.

Paso 2. Generar un listado de ítems

Una vez que el propósito de la escala está claro, el investigador ha de comenzar a construirla, El primer paso es generar un amplio conjunto de ítems de entre los que deberán salir los que conformen definitivamente la escala.

Cada uno de estos ítems tiene que reflejar la variable subyacente que se pretende medir. Asimismo, es mejor generar varios ítems para cada una de las dimensiones del constructo que recurrir a una sola pregunta, por cuanto que así puede calcularse su fiabilidad y, además, cuando se vayan sumando las respuestas aflorará el contenido común a ellos y se minimizará el efecto de las particularidades irrelevantes de cada uno por compensación. Así "haría casi cualquier cosa por asegurar la felicidad de mis hijos" y "ningún sacrificio es mucho si ayuda a conseguir la felicidad de mis hijos" podría ser una redundancia útil por cuanto expresa una idea similar de distintas maneras.

En cuanto al número de ítems, es imposible dar una cifra de cuál debería ser este en el listado inicial. En todo caso debe recordarse que la fiabilidad de la escala es función de la correlación que exista entre los ítems. Como en esta fase no podemos calcular esas correlaciones, es deseable tener ítems suficientes para efectuar sustituciones que mejoren las correlaciones al desarrollar posteriormente la escala. No es extraño comenzar con un listado de ítems tres o cuatro veces superior a la longitud final que se espera dar a la escala.

Referente a las *características de un buen ítem*, es imposible hacer un listado exhaustivo de qué es lo que hace que un ítem sea bueno o malo. En todo caso, debería tenerse en cuenta lo siguiente:

1. Deben evitarse los ítems excesivamente largos, por cuanto la longitud incrementa la complejidad y disminuye la claridad. Sin embargo tampoco debe sacrificarse la claridad en aras de la brevedad eliminando preposi-

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

ciones y conjunciones. En general un ítem del tipo "Con frecuencia tengo problemas para expresar mis puntos de vista" es mejor que "Debo decir que una de las cosas con las que parece que tengo un problema la mayor parte de las veces es la transmisión de mi punto de vista al resto de las personas".

2. La complejidad sintáctica y de léxico de las frases es también muy importante. Debe evitarse la sucesión de negaciones "No estoy a favor de que las empresas no sigan dando fondos a los grupos antinucleares" es mucho más confuso que "Estoy a favor del apoyo de las empresas a grupos antinucleares".

3. Deben evitarse también los ítems con doble argumentación. Así "estoy en contra de la discriminación racial porque es un crimen contra Dios" es un ejemplo de doble argumentación. Si se está en contra de la discriminación racial por motivos que nada tienen que ver con la religión, ¿qué debe contestarse?

4. Combinación de ítems formulados en positivo y en negativo. Muchos investigadores abogan por combinar ítems que suponen la presencia del constructo con ítems que suponen su ausencia. Por ejemplo, la escala de Ronsenberg (1965) de autoestima RSE incluye frases como "Creo que tengo bastantes buenas cualidades" junto con otras como "Ciertamente, muchas veces me siento inútil". El motivo de esta combinación es evitar el sesgo afirmativo del encuestado, es decir, la tendencia manifiesta de estar de acuerdo con las afirmaciones independientemente de su contenido. La combinación permite detectar a las personas que han contestado con este sesgo, dado que estarán de acuerdo tanto con aquellas variables que indican un alto grado de autoestima como con las que la suponen baja. Desgraciadamente, este tipo de formulaciones tienen un aspecto negativo: provocan confusión en los encuestados. Nuestra experiencia demuestra que no todos los entrevistados prestan atención a la inversión y esto provoca que esos ítems guarden poca correlación con el resto de la escala, sus cargas sean bajas y sea necesario eliminarlos, por lo que abogamos por su no inclusión.

Paso 3. Determinar el formato de medida

Existen muchas formas de hacer preguntas y solo presentaremos las más habituales.

Escala Thurstone. Las escalas Thurstone formulan las preguntas generando una serie de ítems que suponen la presencia en distintos grados del constructo que se pretende medir. Es también habitual que estén diseñadas de tal forma que la diferencia de nivel del constructo entre cada par de afirmaciones sea

Estadísticos e-Books & Papers

la misma. Una hipotética escala Thurstone para medir las aspiraciones de los padres respecto a los logros académicos de sus hijos sería la siguiente:

- Que mi hijo alcance el éxito es lo único que puede compensar mis esfuerzos como padre.
- Ir a una buena universidad y conseguir un buen trabajo es importante, pero no esencial, para la felicidad de mi hijo.
- La felicidad no tiene nada que ver con conseguir objetivos materiales o educacionales.
- Lo que habitualmente se considera un éxito es un obstáculo para la verdadera felicidad.

Como señalan Nunnally y Bernstein (1994) es mucho más fácil explicar una escala Thurstone que construirla, por la dificultad que entraña generar ítems que aporten grados diferenciales de presencia del atributo.

Escala Guttman. En una escala Guttman, los ítems están generados de tal forma que responder afirmativamente a uno de ellos supone responder afirmativamente a todos los anteriores. Si se pregunta: ¿Fuma?, ¿Fuma usted más de 10 cigarrillos?, ¿Fuma usted un paquete diario?... Responder afirmativamente a una de estas preguntas supone que también se posee el nivel inferior del atributo medido.

Aunque tanto la escala Thurstone como la Guttman están formadas por ítems que gradúan la presencia del atributo, en la primera se busca una afirmación que fije el nivel del atributo poseído, mientras que en la segunda se busca el punto de transición entre las respuestas afirmativas y las negativas. La siguiente escala sería una versión Guttman del ejemplo anterior:

- Que mi hijo alcance el éxito es lo único que puede compensar mis esfuerzos como padre.
- Ir a una buena universidad y conseguir un buen trabajo es importante para la felicidad de mi hijo.
- La felicidad es más probable si una persona ha conseguido sus objetivos educacionales y materiales.
- Lo que habitualmente se considera un éxito no es un obstáculo para la verdadera felicidad.

La dificultad de las escalas Guttman está en lograr formular los ítems de tal forma que responder afirmativamente a uno suponga hacerlo también a los anteriores. En el ejemplo anterior se supone que si se contesta afirmativamente a la segunda expresión se debe haber contestado igual a la tercera y cuarta. Sin embargo, si un entrevistado viera el éxito como un fenómeno complejo que

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

puede ser a la vez una ayuda y un obstáculo para la felicidad, podría dar un patrón de respuestas no esperado.

Escalas con ítems del mismo peso. En los dos tipos de escalas mostradas, las dificultades de generación suelen superar sus ventajas y no son muy utilizadas. Lo más habitual es generar ítems que sean detectores equivalentes del fenómeno medido, y no de niveles de este. Una de sus ventajas es que las respuestas pueden recogerse de muchas formas distintas, lo que permite al investigador buscar las más adecuadas para su propósito. Veremos algunos aspectos generales de las ventajas e inconvenientes de distintos formatos.

¿Cuántas categorías de respuesta? La mayor parte de los ítems de una escala consisten en dos partes: una afirmación y una serie de opciones de respuesta. Estas opciones pueden ir desde un gran número de alternativas (valorar de 0 a 100 un acuerdo, por ejemplo) a unas pocas (muy de acuerdo, de acuerdo, indiferente, en moderado desacuerdo, muy en desacuerdo). No debemos olvidar que la misión de una escala es detectar la variabilidad en las respuestas, y fracasará en esta tarea si no puede discriminar entre los distintos niveles de presencia del atributo si las alternativas de respuesta son muy limitadas,.

Una segunda cuestión es si el entrevistado tiene capacidad para discriminar de manera significativa entre los distintos niveles de respuesta, lo que depende claramente de qué se esté midiendo. Si se le pregunta sobre su ideología política, puede tener necesidad de matizar su respuesta y requerirá de una escala de bastantes puntos (digamos 10). Pero, si se le pregunta acerca de su acuerdo o desacuerdo con la afirmación "Fumar perjudica la salud", esta necesidad de matización no será tan acuciante y es razonable graduar la respuesta en menor número de alternativas.

¿Han de ser estas alternativas un número par o impar? De nuevo depende de lo que se pregunte. El número impar supone la existencia de un punto neutral. Si el investigador considera que, dadas las características de la pregunta, el entrevistado puede buscar conscientemente la indefinición, será recomendable un número par de alternativas para forzar que tome partido.

En otras preguntas, sin embargo, el encontrar el punto de indefinición puede ser muy útil. Veamos el ejemplo que plantea De Vellis (1991). En un estudio para determinar cuál de dos riesgos prefiere la gente asumir (aburrimiento frente a peligro), el investigador puede ir variando la peligrosidad de la actividad hasta dar con el punto de inflexión donde el entrevistado duda entre aburrimiento y peligro, Esa actividad sería un indicador de la propensión al riesgo del entrevistado:

Estadísticos e-Books & Papers

Señale su preferencia relativa por las actividades A o B entre las alternativas señaladas a continuación.
Actividad A: Leer un libro de filosofía de la ciencia (ningún peligro).
Actividad B: Viajar en un vuelo comercial (poco peligro).
1 = Claramente prefiero A
2 = Prefiero A
3 = No prefiero una ni otra
4 = Prefiero B
5 = Claramente prefiero B

B: Viajar en una avioneta	1 2 3 4 5
B: Saltar en paracaídas, teniendo paracaídas de reserva	1 2 3 4 5
B: Saltar en paracaídas sin paracaídas de reserva	1 2 3 4 5
B: Saltar de una avión sin paracaídas y llegar a una colchoneta	1 2 3 4 5

Escalas Likert. Las escalas Likert son las más utilizadas en el desarrollo de escalas. El ítem se presenta como una afirmación, seguida por alternativas de respuesta que suponen diversos niveles de acuerdo con ella. A estas escalas se les pueden aplicar las consideraciones anteriores sobre el número par o impar de alternativas y la cantidad de esta. Estas escalas plantean básicamente un problema. Si las afirmaciones son muy neutras, existe una gran tendencia al acuerdo y, además, pueden estar recogiendo, de hecho, más la ausencia de opinión que la opinión. Por ejemplo, no es lo mismo pedir el acuerdo o desacuerdo con afirmaciones como "Los médicos no hacen caso normalmente a lo que les dicen sus pacientes" o "Muchas veces los médicos no prestan demasiada atención a sus pacientes" que hacerlo sobre "De vez en cuando, los médicos pueden olvidar algo de lo que les han dicho sus pacientes". En resumen, un buen ítem Likert debería manifestar la opinión de una manera clara.

Diferencial semántico. En general, el diferencial semántico va asociado a un estímulo (como un grupo de personas, por ejemplo, los vendedores de coches). Una vez identificado el estímulo se presenta una lista de pares de adjetivos que representan los puntos opuestos de un continuo (honrado vs, no honrado) como se muestra en el ejemplo:

Vendedores de coches

Honrados	□□□□□□□	No honrados
Apacibles	□□□□□□□	Ruidosos

Paso 4. Hacer que el listado inicial de ítems sea revisado por expertos

Esta fase tiene mucho que ver con la difícil tarea de lograr la validez de contenido de la escala que no puede constatarse de ningún otro modo. Tras dar al panel de expertos la definición que el investigador ha hecho del constructo que tiene que medir y del conjunto de ítems que componen la escala, solicitará al panel que realice las siguientes tareas:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

1. En primer lugar, los expertos confirmarán o invalidarán la definición que el investigador haya hecho del constructo que quiere medir.

2. A continuación el panel de expertos debería valorar el nivel de relevancia que creen que cada ítem tiene para medir el fenómeno que se pretende. Estas valoraciones permitirán seleccionar la muestra definitiva de ítems eligiendo entre los más significativos.

3. Cuando el constructo tiene distintas dimensiones, los expertos deberán revisar la asignación hecha de afirmaciones a dimensiones.

4. Los expertos también deberán valorar la claridad y precisión de la formulación concreta de cada ítem, apuntando, si es necesario, formas alternativas de redactarlos. Esto hace referencia tanto a problemas de claridad estrictamente literaria como a que puedan, por esta falta de claridad, reflejar factores extraños al constructo que se pretende medir.

5. También se pretende que los expertos determinen si se ha dejado fuera de la escala alguna de las dimensiones del fenómeno.

Paso 5. Considérese la inclusión de ítems de validación

Además de los ítems que pretenden medir la variable subyacente, el investigador debe considerar la incorporación de dos tipos de ítems adicionales:

1. Ítems para detectar defectos o problemas de la escala. A veces los entrevistados pueden no estar respondiendo a las preguntas con la motivación que el investigador presupone. Una de las motivaciones más habituales es presentar una imagen socialmente aceptable, y no su verdadera forma de ser. Suele recomendarse para ciertos tipos de estudios la inclusión de una escala que mida esta tendencia, como la de Strahan y Gerbasi (1972).

2. Ítems para evaluar la validez de constructo de la escala. Si la teoría apunta a que el constructo que se está midiendo está relacionado con otros constructos, la inclusión de escalas de medición de algunos de ellos puede ayudar a analizar la validez de la nueva escala desarrollada si se confirman esas relaciones.

Paso 6. Administrar la escala a una muestra de prueba

La muestra tiene que ser lo suficientemente grande como para eliminar la varianza que aporta un individuo como tal, haciendo que no sea representativa de una población determinada, sino de sí misma. Nunnally y Bernstein (1994) apuntan la cifra de 300 personas como un número adecuado, aunque De Vellis (1991) señala que se han desarrollado escalas con éxito con un número menor.

Estadísticos e-Books & Papers

Paso 7. Evaluar los ítems

Con los resultados de administrar la escala a una muestra de prueba, debe analizarse cómo ha funcionado cada uno de los ítems de tal forma que se pueda seleccionar finalmente a los más adecuados.

1. Análisis inicial del funcionamiento de cada ítem. Cuanto mayores sean las correlaciones entre los ítems, mayor será la fiabilidad de la escala. Por ello serán candidatos a abandonar la escala aquellos ítems que, perteneciendo a una misma dimensión del constructo, guarden poca correlación con los demás integrantes de la misma.

2. Formulación inversa. Si algún ítem guarda una elevada correlación negativa con los demás, deberá formularse gramaticalmente en sentido inverso para mantenerlo en la escala. Si esta forma de enunciarlo se ha hecho intencionadamente, deberá cambiarse la codificación de la respuesta.

3. Analizar la varianza de los ítems. Si todos los invididuos responden igual a un ítem, la varianza será cero. En este caso (o en casos de varianzas muy pequeñas) el ítem no ha sido capaz de discriminar entre los individuos con distintos niveles del constructo que se está midiendo, y debería suprimirse.

4. Analizar la media de los ítems. Es deseable que la media de los ítems esté en torno al centro de la escala de respuesta. Si está en la zona del acuerdo, indicará que el enunciado de la afirmación era demasiado neutro.

5. Cálculo del α de Cronbach. Se aplica todo lo expuesto al presentar el análisis de fiabilidad.

Paso 8. Optimizar la longitud de la escala

Ya se ha indicado que la fiabilidad de una escala viene influenciada por el número de ítems de la misma. Por un lado, escalas cortas reducen los problemas de su administración a los entrevistados, pero las largas son más fiables. Encontrar el punto de equilibrio debe ser el objeto de esta etapa. Si una escala es poco fiable, entonces la brevedad no es una virtud. Dicho de otro modo, se puede intentar eliminar ítems para facilitar la administración, pero nunca permitiendo que los α de Cronbach sean inferiores al 0,7 o 0,8 recomendados.

Sin embargo, como vimos en el ejemplo del análisis de fiabilidad, eliminar malos ítems puede mejorar significativamente la fiabilidad de la escala. En ese mismo ejemplo mostrábamos cómo detectar y eliminar esos ítems. Debe mantenerse, sin embargo, un cierto margen de seguridad con el valor de los alfa, porque no debe olvidarse que la escala se ha administrado a una muestra de prueba y estos valores pueden caer cuando se administren a la muestra definitiva.

Si la muestra de prueba es lo suficientemente grande, debería considerarse su división en dos mitades, de tal forma que una de ellas se utilice para calcular

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

los alfa, evaluar los ítems y eliminar los inadecuados, mientras que la segunda mitad se podrá utilizar para validar estos resultados.

Estadísticos e-Books & Papers

Estadísticos e-Books & Papers

15. Modelos de ecuaciones estructurales: modelos de estructuras de covarianza (CB-SEM)

15.1. Introducción

Al presentar el análisis factorial confirmatorio (CFA) en el tema 13, señalábamos que, además de su interés en sí mismo, ayudaba a realizar la transición entre las técnicas que se habían analizado hasta el momento y el capítulo que nos ocupa ahora, esto es, el dedicado a presentar los modelos de estructuras de covarianza (CB-SEM). Y, efectivamente, gran parte de los contenidos que son necesarios para comprender los CB-SEM ya se han ofrecido en el mencionado capítulo: identificación de los modelos, mecanismos de estimación, indicadores de ajuste, reespecificación de los modelos teóricos, etc.

Esto hace que la lectura del capítulo dedicado al CFA sea un requisito imprescindible para seguir la exposición del presente. Es más, a efectos didácticos, aprovecharemos que la mayoría de aspectos fueron abordados de una manera formalizada en el tema del CFA para no recurrir excesivamente a la formalización de los contenidos y dotar al capítulo dedicado a los CB-SEM de un carácter mucho más aplicado, utilizando la presentación de ejemplos como eje conductor del mismo.

En la comparación de las figuras 13.1 y 13.8 del capítulo trece, ya apuntábamos la diferencia entre un CFA y un MEC: en el CFA no se plantean relaciones teóricas entre las variables latentes, y son solo un paso previo necesario para la validación del modelo de medida, mientras que en un CB-SEM, sin embargo, sí que se considera que un factor común puede influir sobre otro y se apunta la dirección de esa causalidad. Siguiendo a Long (1983b), los CB-SEM intentan explicar las relaciones entre un conjunto de variables observadas basándose en las relaciones de un número más pequeño de variables latentes, caracterizando dichas relaciones mediante las covarianzas de las variables observadas.

Caso 15.1. Modelización del desempeño de una fuerza de ventas

La figura 15.1 nos muestra un ejemplo de CB-SEM que nos permitirá ilustrar esta definición, introducir nuevos términos respecto a lo expuesto en el CFA y presentar la formalización matemática del CB-SEM en el epígrafe siguiente. Dicha figura modeliza la relación propuesta por Bagozzi (1980) entre

Estadísticos e-Books & Papers

el desempeño en el trabajo de una fuerza de ventas y la satisfacción con el mismo, indagando acerca de sus antecedentes. Expliquemos con más detalle el contenido del modelo planteado.

Entre los directores de una fuerza de venta existe el convencimiento de que existe una relación entre el *desempeño*, o eficacia de los vendedores en su trabajo (variable latente η_1), y la *satisfacción* que estos declaran tener con su empleo (variable latente η_2). Sobre lo que no hay acuerdo es sobre si es el desempeño quien provoca la satisfacción o la satisfacción quien influye positivamente sobre el desempeño. Asimismo, la influencia de una variable latente sobre la otra puede no estar causada por la relación causal directa entre ambas, sino deberse a la existencia de unos antecedentes comunes que afectan a los individuos y que, tras revisar la literatura, Bagozzi (1980) sintetiza en: *motivación para el logro* (ξ_1), es decir, valor que cada individuo da a las recompensas que puede recibir por hacer bien su trabajo, la *autoestima* (ξ_2), es decir, el buen concepto que uno tiene de sí mismo y que se considera que puede influir sobre el buen desempeño de un trabajo, y la *inteligencia verbal* (ξ_3), esto es, la habilidad cognitiva para percibir fielmente los contenidos de las conversaciones, las instrucciones escritas y otra formas de comunicación, lo que hará que se pueda reaccionar adecuadamente ante las objeciones de los clientes y organizar correctamente las tareas de ventas.

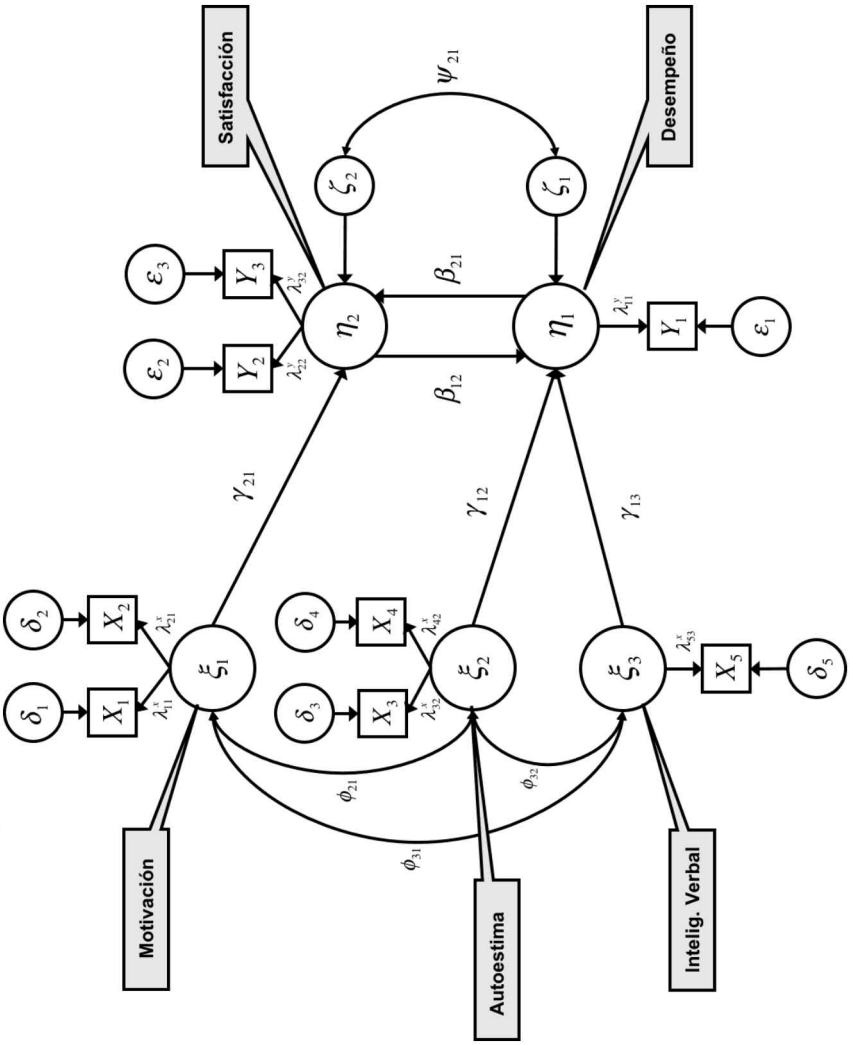
Para facilitar el desarrollo matemático posterior de la presentación del modelo seguiremos la notación habitualmente empleada en los textos donde se presenta esta herramienta, que es la de Jöreskog (1970), recogida en la figura 15.1. Todo modelo de estructuras de covarianzas se descompone en dos componentes. El primero de ellos es el *componente estructural*, que incluye las relaciones entre los factores latentes. La convención de notación es la siguiente, el factor latente que actúa como variable independiente (de él solo salen relaciones directas, es decir, flechas de una punta) se denota con la letra griega ξ. Los factores latentes dependientes (reciben flechas de una sola punta) se denotan por la letra griega η. Los parámetros que representan los coeficientes de regresión entre los factores latentes se denotan con la letra γ si la relación que representa es entre un factor independiente sobre uno dependiente y mediante la letra β si es entre dos dependientes. Las covarianzas entre los factores latentes independientes se recogen mediante la letra ϕ y están ocasionadas por predictores comunes de los factores independientes no contemplados en el modelo. Por su parte, los factores latentes dependientes no se espera que estén predichos perfectamente por los independientes, por lo que se les asocia un término de error estructural (ζ). Estos términos pueden covariar entre ellos (ψ), indicando que los factores dependientes asociados con ellos comparten una variación común no explicada por las relaciones que se expresan en el modelo.

Como veíamos en el CFA, cada variable latente ha de medirse de algún modo, es decir, se han de encontrar variables observadas que midan ese factor, ya sea dependiente o independiente, esto es, cada factor latente se modeliza como un factor común que subyace bajo una serie de variables observadas. Esta parte del CB-SEM se denomina *modelo de medida*. A las variables observadas que

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 15.1.: Modelo de desempeño de la fuerza de ventas



Fuente: Bagozzi (1980)

Estadísticos e-Books & Papers

miden un factor dependiente se las denota como Y, y a las que miden un factor independiente, como X. Las cargas factoriales de las variables observadas sobre el factor se designan como λ_x si el factor es independiente y como λ_y si es dependiente.

El investigador ha de reconocer que sus medidas son imperfectas y tratará de modelizar esa imperfección. Por ello, los CB-SEM, como ocurría con los CFA, incorporan los denominados *errores de medida*, que denominábamos en el CFA como factores únicos. Los asociados a las variables observadas (Y) de un factor dependiente se denotan con la letra ε, y los asociados a variables observadas (X) de un factor independiente, con δ.

Presentados los términos y la notación específica de los modelos de estructura de covarianzas, dedicaremos el siguiente epígrafe a la formalización matemática del problema que resuelve un CB-SEM. El epígrafe puede saltarse sin solución de continuidad, dado que posteriormente aplicaremos a un nuevo caso el procedimiento de identificación y estimación de una manera mucho más intuitiva. Pero, dado que en la mayoría de textos se presenta el tema de este modo, queremos facilitar el seguimientos de otros textos desarrollándolo de manera aplicada al caso que estamos utilizando.

15.2. Formalización matemática del CB-SEM

En su forma más general, los modelos de estructuras de covarianzas se pueden representar a partir de tres sistemas de ecuaciones. En primer lugar, el **componente estructural** recoge la relación causal entre las variables latentes:

$$\eta = \mathbf{B}\eta + \mathbf{\Gamma}\xi + \zeta \tag{15.1}$$

donde, como se sintetiza en el cuadro 14.1, η es un vector $r \times 1$ que contiene los r factores dependientes; $\mathbf{B}$ es una matriz $r \times r$ que contiene los coeficientes que relacionan entre sí a los factores dependientes; $\mathbf{\Gamma}$ es una matriz $r \times s$ con los coeficientes que relacionan los r factores dependientes con los s independientes; ξ es un vector $s \times 1$ con los factores independientes; ζ es un vector $r \times 1$ que contiene los errores asociados con los factores dependientes y que, como se ha señalado, indican que estos factores no están totalmente determinados por las ecuaciones estructurales.

En la literatura econométrica, la ecuación (15.1), tras restar $\mathbf{B}\eta$ a ambos miembros, se suele escribir como:

$$\eta - \mathbf{B}\eta = \mathbf{\Gamma}\xi + \zeta$$

y definir

$$\ddot{\mathbf{B}} = \mathbf{I} - \mathbf{B}$$

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 15.1.: Resumen del componente estructural

Matriz	Dimensión	Media	Covarianza	Dimensión	Descripción
η	$r \times 1$	$\mathbf{0}$	$Cov(\eta) = E\left(\eta\eta'\right)$	$r \times r$	Factores dependientes
ξ	$s \times 1$	$\mathbf{0}$	$\mathbf{\Phi} = E\left(\xi\xi'\right)$	$s \times s$	Factores independientes
ζ	$r \times 1$	$\mathbf{0}$	$\mathbf{\Psi} = E(\zeta\zeta')$	$r \times r$	Errores en las ecuaciones
$\mathbf{B}$	$r \times r$	_	_	_	Efectos directos $\eta \longrightarrow \eta$
$\dot{\mathbf{B}}$	$r \times r$	_	_	_	$\mathbf{I} - \mathbf{B}$
$\mathbf{\Gamma}$	$r \times s$	_	_	_	Efectos directos $\xi \longrightarrow \eta$

Fuente: Long (1983b)

© Ediciones Paraninfo

Estadísticos e-Books & Papers

se puede expresar de la siguiente forma:

$$\ddot{\mathbf{B}}\eta = \mathbf{\Gamma}\xi + \zeta \tag{15.2}$$

Long (1983b) señala que, aunque la expresión (15.2) es más adecuada para deducir una serie de resultados, (15.1) es más cómoda de interpretar, dado que un valor positivo en $\mathbf{B}$ indica una relación positiva entre los factores dependientes, mientras que un valor positivo de $\ddot{\mathbf{B}}$ indica una relación negativa.

¿Cuáles son las hipótesis que se postulan en un modelo CB-SEM? Pueden sintetizarse en las siguientes:

1. Las variables están medidas en desviaciones respecto a su media, lo que implica que:
$$E(\eta) = E(\xi) = E(\zeta) = \mathbf{0}$$
Esto no afecta a la generalizabilidad del modelo, dado que a los parámetros estructurales contenidos en $\mathbf{B}$ y $\mathbf{\Gamma}$ no les afecta esta hipótesis.

2. Los términos de error de los factores dependientes y los factores independientes están incorrelacionados entre sí, es decir:
$$E(\xi\zeta') = E(\zeta\xi') = \mathbf{0}$$

3. También se asume que $\mathbf{B}$ es no singular (existe su inversa), lo que implica que (15.2) se pueda resolver para el vector de factores dependientes, es decir, para η.

Para evitar confusiones en la notación, veamos su aplicación al ejemplo que venimos utilizando. El componente estructural del modelo, que es el que recogemos en la ecuación (15.1), venía ilustrado en la figura 15.1. Así, las ecuaciones que resumen las relaciones entre los factores y que se desprenden directamente del gráfico son las siguientes:

$$\eta_1 = \gamma_{12}\xi_2 + \gamma_{13}\xi_3 + \beta_{12}\eta_2 + \zeta_1$$

$$\eta_2 = \gamma_{21}\xi_1 + \beta_{21}\eta_1 + \zeta_2$$

Si expresamos matricialmente las ecuaciones anteriores, tenemos que:

$$\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} 0 & \beta_{12} \\ \beta_{21} & 0 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} 0 & \gamma_{12} & \gamma_{13} \\ \gamma_{21} & 0 & 0 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{bmatrix} + \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix}$$

La ecuación anterior recoge exactamente las mismas matrices que se recogían en (15.1). Además de las ecuaciones anteriores, sería necesario añadir para completar el contenido del cuadro 15.1 las matrices que recogen las matrices de

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

varianzas y covarianzas de los factores independientes ($\mathbf{\Phi}$) y entre los términos de error de los factores dependientes ($\mathbf{\Psi}$), esto es:

$$\mathbf{\Psi} = \begin{bmatrix} \psi_{\mathbf{11}} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{bmatrix} \Phi = \begin{bmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ \phi_{21} & \phi_{22} & \phi_{23} \\ \phi_{31} & \phi_{32} & \phi_{33} \end{bmatrix}$$

Los otros dos sistemas de ecuaciones que representan un CB-SEM son las que se ocupan de reflejar el **instrumento de medida**, es decir, la parte del CB-SEM que establece los indicadores de cada factor. Estos sistemas de ecuaciones son los siguientes:

$$\mathbf{x} = \mathbf{\Lambda_x}\xi + \delta \tag{15.3}$$

$$\mathbf{y} = \mathbf{\Lambda_y}\eta + \varepsilon \tag{15.4}$$

Parte de los elementos implicados ya se han presentado con anterioridad. Los nuevos elementos de las ecuaciones anteriores son el vector $\mathbf{x}$ de dimensión $q \times 1$, que contiene las variables observadas que sirven para medir los factores independientes (q es el número de esas variables observadas); $\mathbf{y}$ es el vector de variables observadas (hay p de ellas, por ello su dimensión es $p \times 1$) que miden los factores dependientes. Por su parte, las matrices que contienen los coeficientes de regresión o cargas factoriales entre esas variables y sus respectivos factores vienen denotadas por Λ_x para los factores independientes y Λ_y para los dependientes. Sus dimensiones son, respectivamente, $q \times s$ y $p \times r$, siendo s y r el número de factores de cada tipo.

Las hipótesis que se postulan sobre las variables de las ecuaciones anteriores son las siguientes:

1. Las variables están medidas en desviaciones respecto a su media, lo que implica que:
$$E(\mathbf{x}) = E(\delta) = E(\varepsilon) = \mathbf{0}$$

2. Los factores y los términos de error de las variables observadas están incorrelacionados entre sí, es decir:
$$E(\xi\delta') = E(\delta\xi') = E(\eta\varepsilon') = E(\varepsilon\eta') = \mathbf{0}$$

3. Los términos de error de las variables observadas están incorrelacionados entre sí.
$$E(\delta\varepsilon') = E(\varepsilon\delta') = \mathbf{0}$$

Toda esta información se sintetiza en el cuadro 15.2

Los cuadros 15.1 y 15.2 resumen la notación y las hipótesis que subyacen bajo un modelo de ecuaciones estructurales. El proceso de estimación consiste, como

Estadísticos e-Books & Papers

Cuadro 15.2.: Resumen del instrumento de medida

Matriz	Dimensión	Media	Covarianza	Diemensión	Descripción
η	$r \times 1$	$\mathbf{0}$	$Cov(\eta) = E(\eta\eta')$	$r \times r$	Factores dependientes
ξ	$s \times 1$	$\mathbf{0}$	$\mathbf{\Phi} = E(\xi\xi')$	$s \times s$	Factores independientes
$\mathbf{x}$	$q \times 1$	$\mathbf{0}$	$\sum_{xx} = E(\mathbf{xx}')$	$q \times q$	Var. observadas indep.
$\mathbf{y}$	$p \times 1$	$\mathbf{0}$	$\sum_{yy} = E(\mathbf{yy}')$	$p \times p$	Var. observadas depend.
$\mathbf{\Lambda_x}$	$q \times s$	_	_	_	Cargas de $\mathbf{x}$ sobre ξ
$\mathbf{\Lambda_y}$	$p \times r$	_	_	_	Cargas de $\mathbf{y}$ sobre η
δ	$q \times 1$	$\mathbf{0}$	$\mathbf{\Theta}_\delta = E(\delta\delta')$	$q \times q$	Errores de $\mathbf{x}$
ε	$p \times 1$	$\mathbf{0}$	$\mathbf{\Theta}_\varepsilon = E(\varepsilon\varepsilon')$	$p \times p$	Errores de $\mathbf{y}$

Fuente: Long (1983b)

© Ediciones Paraninfo
Estadísticos e-Books & Papers

en el CFA, en estimar los parámetros que la matriz de varianzas y covarianzas —definida más adelante—, que se deriva de las ecuaciones (15.1), (15.3) y (15.4) y que inmediatamente deduciremos, sea lo más parecida posible a la matriz de varianzas y covarianzas muestral $\mathbf{S}$.

Antes de examinar esta deducción, identifiquemos el componente de medida del CB-SEM en el ejemplo que venimos siguiendo. Basta observar la figura 15.1 para obtener de manera inmediata las siguientes ecuaciones:

$$\left.\begin{array}{l} x_1 = \lambda_{11}^x \xi_1 + \delta_1 \\ x_2 = \lambda_{21}^x \xi_1 + \delta_2 \\ x_3 = \lambda_{32}^x \xi_2 + \delta_3 \\ x_4 = \lambda_{42}^x \xi_2 + \delta_4 \\ x_5 = \lambda_{53}^x \xi_3 + \delta_5 \end{array}\right\} \quad \left.\begin{array}{l} y_1 = \lambda_{11}^y \eta_1 + \varepsilon_1 \\ y_2 = \lambda_{22}^y \eta_1 + \varepsilon_2 \\ y_3 = \lambda_{32}^y \eta_1 + \varepsilon_3 \end{array}\right\}$$

Estas ecuaciones se pueden expresar en forma matricial, con la misma estructura que las ecuaciones (15.3) y (15.4), de la siguiente forma:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} \lambda_{11}^x & 0 & 0 \\ \lambda_{21}^x & 0 & 0 \\ 0 & \lambda_{32}^x & 0 \\ 0 & \lambda_{42}^x & 0 \\ 0 & 0 & \lambda_{53}^x \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \end{bmatrix}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} \lambda_{11}^y & 0 \\ 0 & \lambda_{22}^y \\ 0 & \lambda_{32}^y \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix}$$

¿Qué es estimar el CB-SEM? Obtener los parámetros que hacen que la matriz de varianzas y covarianzas implicada por el modelo se parezca al máximo a la muestral, luego necesitamos saber qué contenido ha de tener la primera. Veamos, a continuación, por tanto, cuál es la matriz de varianzas y covarianzas $\mathbf{\Sigma}$ que se deriva de los componentes estructural y de medida. Siguiendo la notación de Long (1983b), y dado que las variables están medidas en desviaciones respecto a sus medias, la matriz de varianzas y covarianzas poblacional $\mathbf{\Sigma}$ vendrá dada por:

$$\Sigma = E\left[\left[\frac{\mathbf{y}}{\mathbf{x}}\right]\left[\frac{\mathbf{y}}{\mathbf{x}}\right]'\right] = E\left[\begin{array}{c|c} \mathbf{yy}' & \mathbf{yx}' \\ \hline \mathbf{xy}' & \mathbf{xx}' \end{array}\right]$$

donde $[\mathbf{y} \mid \mathbf{x}]$ es el vector de dimensión $(p+q) \times 1$ que se obtiene al disponer los vectores $\mathbf{x}$ e $\mathbf{y}$ uno junto al otro. Si sustituimos las ecuaciones (15.3) y (15.4) en la expresión anterior tenemos:

$$\mathbf{\Sigma} = E\left[\begin{array}{c|c} (\mathbf{\Lambda_y}\eta + \varepsilon)(\mathbf{\Lambda_y}\eta + \varepsilon)' & (\mathbf{\Lambda_y}\eta + \varepsilon)(\mathbf{\Lambda_x}\xi + \delta)' \\ \hline (\mathbf{\Lambda_x}\xi + \delta)(\mathbf{\Lambda_y}\eta + \varepsilon)' & (\mathbf{\Lambda_x}\xi + \delta)(\mathbf{\Lambda_x}\xi + \delta)' \end{array}\right]$$

Efectuando las multiplicaciones de matrices se obtiene que:

Estadísticos e-Books & Papers

$$\mathbf{\Sigma} = E \left[\begin{array}{c|c} \mathbf{\Lambda_y}\eta\eta'\mathbf{\Lambda_y'} + \varepsilon\varepsilon' & \mathbf{\Lambda_y}\eta\xi'\mathbf{\Lambda_x'} + \varepsilon\delta' \\ +\mathbf{\Lambda_y}\eta\varepsilon' + \varepsilon\eta'\mathbf{\Lambda_y'} & +\mathbf{\Lambda_y}\eta\delta' + \varepsilon\xi'\mathbf{\Lambda_x'} \\ \hline \mathbf{\Lambda_x}\xi\eta'\mathbf{\Lambda_y'} + \delta\varepsilon' & \mathbf{\Lambda_x}\xi\xi'\mathbf{\Lambda_x'} + \delta\delta' \\ +\mathbf{\Lambda_x}\xi\varepsilon' + \delta\eta'\mathbf{\Lambda_y'} & +\mathbf{\Lambda_x}\xi\delta' + \delta\xi'\mathbf{\Lambda_x'} \end{array} \right]$$

Aplicando el operador esperanza matemática, y teniendo en cuenta las hipótesis subyacentes de los cuadros 15.1 y 15.2, se llega a que:

$$\mathbf{\Sigma} = \left[\begin{array}{c|c} \mathbf{\Lambda_y B^{-1}}\left(\mathbf{\Gamma\Phi\Gamma'} + \mathbf{\Psi}\right)\mathbf{B'^{-1}\Lambda_y'} + \mathbf{\Theta}_\varepsilon & \mathbf{\Lambda_y B^{-1}\Gamma\Phi\Lambda_x'} \\ \hline \mathbf{\Lambda_y B^{-1}\Gamma\Phi\Lambda_x'} & \mathbf{\Lambda_x\Phi\Lambda_x'} + \mathbf{\Theta}_\delta \end{array} \right] \tag{15.5}$$

Es muy importante que se sea consciente de cómo la matriz (15.5) contiene todos los parámetros que al final han de estimarse en el modelo: las cargas factoriales del instrumento de medida, tanto de las variables latentes dependientes ($\mathbf{\Lambda_y}$) como independientes ($\mathbf{\Lambda_x}$), los coeficientes de regresión de la parte estructural del modelo, tanto los que unen factores independientes con los dependientes ($\mathbf{\Gamma}$) como los que unen factores dependientes entre sí ($\mathbf{B}$), covarianzas entre los factores independientes y varianzas de los mismos ($\mathbf{\Phi}$), varianzas, varianzas de los errores de los indicadores de factores dependientes ($\mathbf{\Theta}_\varepsilon$) y de los factores independientes ($\mathbf{\Theta}_\delta$), varianzas y covarianzas de los errores de los factores dependientes ($\mathbf{\Psi}$). Pues bien, la estimación de un modelo CB-SEM pasa, como se ha señalado, por buscar aquellos estimadores que hacen que la matriz estimada según la expresión (15.5) se parezca lo más posible a la matriz de varianzas-covarianzas muestral $\mathbf{S}$.

15.3. Identificación del modelo de ecuaciones estructurales

Todo lo que se indicó respecto a la identificación de un AFC es válido para un CB-SEM, aunque en este último caso tiene mayor complejidad debido a que el número de parámetros suele ser muy superior. De nuevo existen dos tipos de condiciones: las necesarias y suficientes y las que solamente son necesarias.

Al primer tipo de condiciones, **necesarias y suficientes**, corresponden las siguientes:

1. La matriz de información debe ser definida positiva. La matriz de información es la formada por las derivadas de segundo orden de la función de ajuste que se emplee (véase el epígrafe 13.4) con respecto a los parámetros que se han de estimar en el modelo. Si el modelo está identificado, el rango de esta matriz debe ser igual al número de parámetros libres del modelo (esto es, la matriz debe ser definida positiva). Esto es equivalente a la comprobación que se hace en el modelo de regresión de que el rango de la matriz de varianzas y covarianzas de los regresores debe ser igual al

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

número de regresores. De todas formas, en el caso del CB-SEM, y como ya señalábamos en el apartado anterior, diversos autores, como por ejemplo McDonald (1982), ponen en duda que esta condición sea necesaria y suficiente. En todo caso, los programas estadísticos la comprueban e informan de qué parámetros plantean problemas de identificación, por lo que debe prestarse atención a esta parte de sus salidas.

2. Bekker *et al.* (1994) presentaron un enfoque que supone evaluar el jacobiano (derivadas de primer orden de la función de ajuste con respecto a los parámetros libres), demostrando que en caso de ser definida positiva el modelo estaría identificado. Este enfoque, sin embargo, no ha sido todavía implementado en los programas estadísticos al uso y no es, por ello, aplicable.

Por lo expuesto, lo habitual en la literatura es recurrir a **condiciones necesarias,** que se resumen en las siguientes (véase Hatcher, 1994 y Ullman, 2001):

1. El número de datos (varianzas y covarianzas muestrales) debe ser siempre superior al número de parámetros por estimar.

2. Debe establecerse la escala de los factores dependientes e independientes. Esto se logra fijando a 1 la carga factorial asociada a una de las variables observadas o fijando a 1 la varianza de un factor. Esto último solo es aplicable a los factores independientes, porque en los dependientes la varianza no es un parámetro para estimar directamente.

3. Hay que asegurar la identificabilidad del componente de medida. Si solo hay un factor, el modelo estará identificado si el factor tiene al menos tres variables observadas que carguen sobre él. Si hay dos o más factores hay que fijarse en cuántas variables cargan sobre cada uno. Si hay tres o más, el modelo estará identificado si los errores asociados con los indicadores no están correlacionados, cada variable carga solo sobre un factor y los factores pueden covariar entre ellos. Si solo hay dos indicadores por factor, el modelo puede estar identificado si los errores asociados con cada indicador no están correlacionados, cada indicador carga solo sobre un factor y ninguna de las covarianzas entre los factores está fijada a cero. Si solo hay un indicador, las varianzas de los términos de error del indicador se han de fijar a cero.

4. Los parámetros del coeficiente de regresión de la variable observada sobre el término de error se fijan arbitrariamente a 1.

Estadísticos e-Books & Papers

Apliquemos lo expuesto a nuestro ejemplo. Vamos a utilizar ahora un paquete de R distinto, recurriremos a la función **sem{sem}**[1](Fox, 2006). Veamos cómo identificar el modelo.

En primer lugar, ¿de cuántos datos disponemos? Contamos con 8 variables observadas, por lo tanto, con $8(8+1)/2 = 36$ varianzas y covarianzas muestrales. Si, por el contrario, contamos con los parámetros que hay que estimar antes de que el modelo esté identificado (véase la figura 15.2) tendríamos que estimar 40, que se corresponden con (véanse los * de dicha figura):

- 5 varianzas de los términos de error δ correspondientes a indicadores de factores independientes y 3 correspondientes a varianzas de indicadores de factores dependientes (ε).
- 2 varianzas de los términos de error (ζ) de factores dependientes.
- 3 varianzas de los factores independientes ξ.
- 8 coeficientes de regresión de las variables observadas sobre los términos de error δ y ε, aunque ya vimos en el CFA que por simplicidad la identificación es directa fijándolos a 1 desde el principio.
- 2 coeficientes de regresión de los términos de error ζ sobre los factores dependientes, que al igual que en el caso anterior se fijan a 1 desde el principio.
- 4 covarianzas, 3 entre los factores independientes (ϕ) y 1 entre los errores de los factores dependientes (ζ) dado que están conectados por una relación no recursiva (doble flecha), si no fuera el caso, esta covarianza no existiría.
- 8 coeficientes de regresión entre los factores y las variables observadas (λ).
- 5 entre factores dependientes e independientes (β y γ).

Por lo tanto, antes de aplicar las restricciones adicionales, este modelo está claramente infraidentificado.

A continuación mostramos la sintaxis de **sem{sem}** que resuelve el CB-SEM de nuestro ejemplo. Sigámosla para explicar los pasos de identificación. Véase la figura 15.3 para la síntesis de los pasos de identificación efectuados.

```
modelo<-specifyEquations()
#Medida=======================================
x1=1*desempeno
x2=1*satisfaccion
```

[1]https://cran.r-project.org/package=sem

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

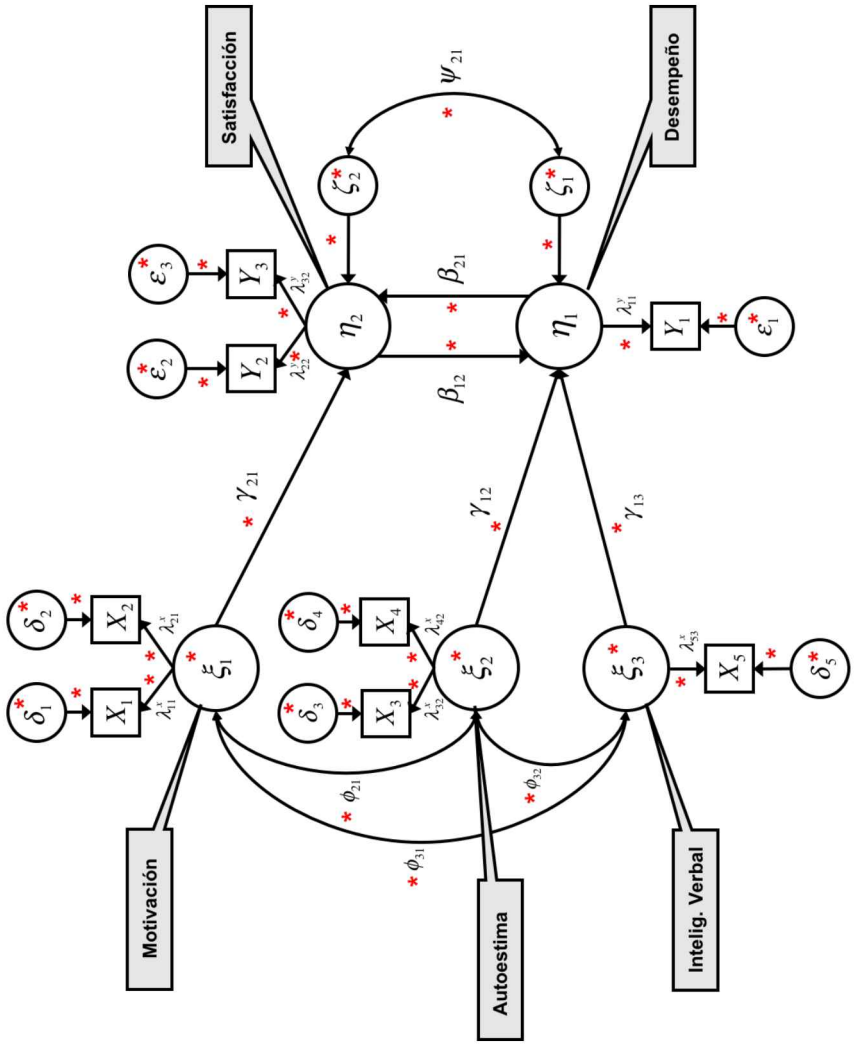


Figura 15.2.: Parámetros iniciales que se deben estimar en el modelo

Estadísticos e-Books & Papers

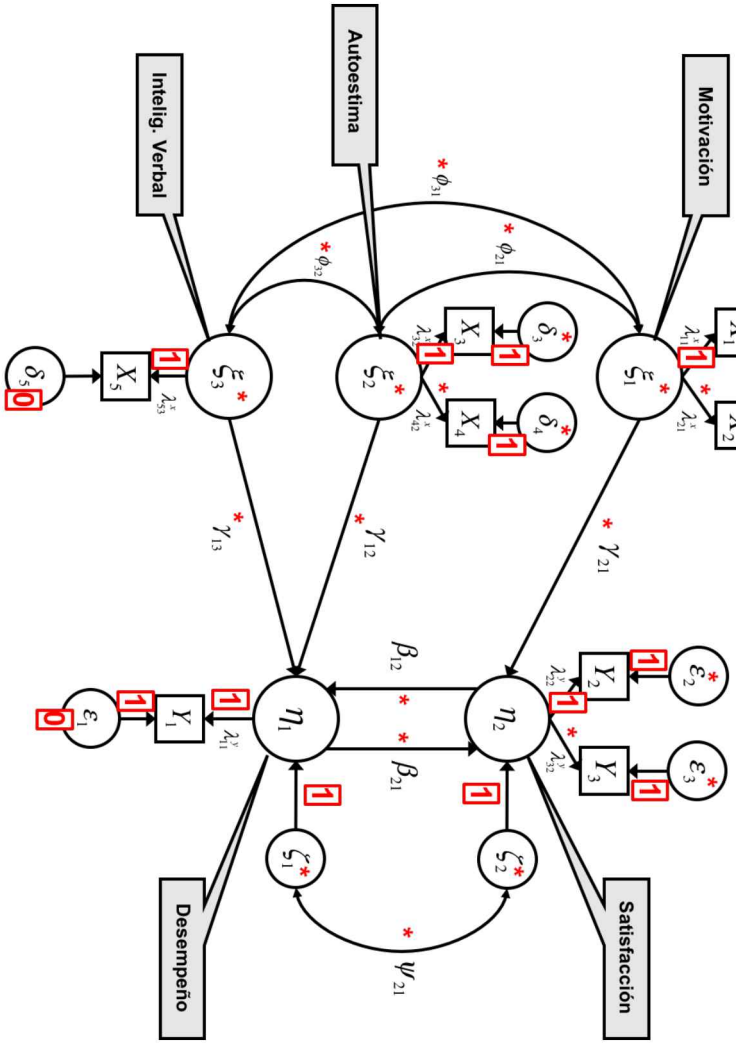


Figura 15.3.: Parámetros para estimar tras la identificación y parámetros restringidos

© Ediciones Paraninfo

Estadísticos e-Books & Papers

```
x3=lam1*satisfaccion
x4=1*motivacion
x5=lam2*motivacion
x6=1*autoestima
x7=lam3*autoestima
x8=1*iverbal
#Estructural==================================
satisfaccion=beta1*desempeno+gam1*motivacion
desempeno=beta2*satisfaccion+gam2*autoestima+gam3*iverbal
#Varianzas factores============================
V(motivacion)=phi1
V(autoestima)=phi2
V(iverbal)=phi3
#Varianzas errores=============================
V(x1)=0
V(x2)=the22
V(x3)=the33
V(x4)=the44
V(x5)=the55
V(x6)=the66
V(x7)=the77
V(x8)=0
#Varianzas Disturbances========================
V(desempeno)=psi1
V(satisfaccion)=psi2
#Covarianzas error relacion no recursiva======
C(desempeno,satisfaccion)=psi12
#Covarianzas factores=========================
C(motivacion,autoestima)=phi12
C(motivacion,iverbal)=phi13
C(autoestima,iverbal)=phi23
```

En primer lugar, fijamos a 1 los *coeficientes de regresión entre las variables observadas y sus términos de error*. Lo mismo hacemos para los factores dependientes y sus términos de error. En **sem{sem}** ni siquiera se escriben en las ecuaciones los términos de error, por lo que este paso está implícito en la sintaxis. Puede verse como no se añade a la ecuación el error δ_3 del indicador X_3 ni el ζ_2 del factor satisfacción.

```
x3=lam1*satisfaccion
satisfaccion=beta1*desempeno+gam1*motivacion
```

En segundo lugar, establecemos la *escala de los factores* dependientes e independientes. Lo hacemos fijando a 1 el coeficiente de regresión de una de las variables observadas que los determinan. Una por cada factor. Vemos como

Estadísticos e-Books & Papers

para el factor satisfacción hemos fijado a 1 (`1*`) la carga del primer indicador, mientras que estimamos dándole una etiqueta al parámetro (`lam1`) la carga del segundo.

```
x1=1*desempeno
x2=1*satisfaccion
x3=lam1*satisfaccion
x4=1*motivacion
x5=lam2*motivacion
x6=1*autoestima
x7=lam3*autoestima
x8=1*iverbal
```

A continuación se fijan a cero las varianzas de los términos de error asociados a las variables observadas cuando un factor está explicado solo por un indicador (caso de X_1 y X_8). Véase cómo en el resto de varianzas se les asocia una etiqueta al parámetro por estimar (`the22`, por ejemplo, para X_2):

```
V(x1)=0
V(x2)=the22
V(x3)=the33
V(x4)=the44
V(x5)=the55
V(x6)=the66
V(x7)=the77
V(x8)=0
```

Finalmente este caso tiene la poco habitual circunstancia de una relación no recursiva entre los factores de satisfacción y desempeño. Cuando esto ocurre, para que el modelo esté identificado, es necesario hacer covariar sus errores ζ_1 y ζ_2 :

```
C(desempeno,satisfaccion)=psi12
```

Aunque hemos dado a los parámetros etiquetas relacionadas con la matriz de la que forman parte según el desarrollo matemático del apartado anterior, esto no es necesario, pueden etiquetarse como se desee.

Podemos ahora comprobar que **el modelo está sobreidentificado**, lo que no ocurría antes de introducir las restricciones, ya que había 36 varianzas y covarianzas muestrales, pero 40 parámetros por estimar. Sin embargo, se han fijado a 1 los 10 coeficientes de regresión de los términos de error. Asimismo se han fijado a 1, para establecer la unidad de medida, los coeficientes de regresión entre X_1 y motivación, X_3 y autoestima, X_5 e inteligencia verbal, Y_2 y satisfacción e Y_1 y desempeño (5 en total) y, finalmente, se han fijado a cero las varianzas de ε_1 (Y_1) y δ_5 (X_5), por lo tanto, los parámetros por estimar son

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

$40 - 10 - 5 - 2 = 23$. Nuestro modelo está ahora sobreidentificado y tenemos $36 - 23 = 13$ grados de libertad. Identificado el modelo, pasemos ahora a la estimación del mismo

15.4. Estimación del modelo de ecuaciones estructurales

Para el análisis de los distintos procedimientos de estimación de los parámetros y los indicadores acerca de la bondad de ajuste se remite al lector al epígrafe 13.4, dado que todo lo señalado para el CFA es válido para el MEC. En este capítulo repetiremos brevemente en qué consiste el procedimiento de estimación y lo ilustraremos con el ejemplo que venimos manejando.

Una vez se ha identificado el modelo, este puede estimarse por cualquiera de los procedimientos señalados en el capítulo anterior: máxima verosimilitud (ML), mínimos cuadrados generalizados (GLS), mínimos cuadrados sin ponderar (ULS), etc. Las estimaciones de los parámetros se harán de forma que se minimice la diferencia entre la matriz de varianzas y covarianzas muestrales $\mathbf{S}$ y la matriz de varianzas y covarianzas predicha $\hat{\mathbf{\Sigma}}$ que, recordemos, venía dada por la expresión (15.5). Cada procedimiento de estimación define de una manera distinta esa "diferencia" o función de ajuste. Un listado de las funciones de ajuste de cada método se ofreció en el apartado 13.4.

Si nos fijamos en la sintaxis de `sem{sem}` comprobamos que en este caso el método de estimación no aparece especificado. Esto se debe a que la opción por defecto es la máxima verosimilitud. ¿Qué estimaciones de los parámetros se han obtenido? En primer lugar, el programa ofrece las estimaciones de los coeficientes de regresión entre los factores y las variables observadas, ya sean estos dependientes o independientes, para, inmediatamente, ofrecer las estimaciones de los parámetros que vinculan los factores dependientes con los independientes (cuadro 15.3), replicando las ecuaciones que mostrábamos en el apartado anterior. También se ofrecen varianzas de errores, covarianzas... en general todos los parámetros que hemos señalado en la subsección anterior. Entraremos en la interpretación más adelante, pero, en esta fase de la investigación, lo relevante es la parte estructural del modelo que contrastan las hipótesis planteadas, es decir, nos centraríamos en los parámetros β y γ. Por ejemplo el efecto de la autoestima sobre el desempeño es positivo y significativo ($\gamma_{12} = 0{,}95; t = 3{,}46; p < 0{,}01$). También se ofrecen las estimaciones estandarizadas de esos parámetros (cuadro 15.4)

15.5. Bondad de ajuste del modelo estimado

Al igual que ocurría con la identificación del modelo, los principales indicadores de medida de su ajuste fueron abordados en el capítulo dedicado al CFA. Por ello, en este nos centraremos solo en su aplicación al ejemplo.

Estadísticos e-Books & Papers

Cuadro 15.3.: Resultados no estandarizados de la estimación

```
Parameter Estimates
      Estimate   Std Error  z value    Pr(>|z|)
lam1   0.9331989 0.16216556  5.7546058 8.684419e-09 x3 <--- satisfaccion
lam2   1.0980063 0.34691186  3.1650873 1.550364e-03 x5 <--- motivacion
lam3   0.9235472 0.14923590  6.1885056 6.073724e-10 x7 <--- autoestima
beta1  0.3081530 0.14619312  2.1078487 3.504408e-02 satisfaccion <--- desempeno
gam1   0.5188620 0.22862412  2.2694983 2.323804e-02 satisfaccion <--- motivacion
beta2 -0.2582784 0.48723818 -0.5300866 5.960519e-01 desempeno <--- satisfaccion
gam2   0.9590623 0.27698101  3.4625561 5.350703e-04 desempeno <--- autoestima
gam3  -0.1786967 0.10218716 -1.7487200 8.033943e-02 desempeno <--- iverbal
phi1   0.3427943 0.14325246  2.3929385 1.671404e-02 motivacion <--> motivacion
phi2   0.5911988 0.14149006  4.1783768 2.935970e-05 autoestima <--> autoestima
phi3   1.0000000 0.12856487  7.7781746 7.357848e-15 iverbal <--> iverbal
the22  0.3281175 0.11061623  2.9662688 3.014369e-03 x2 <--> x2
the33  0.4148843 0.10380383  3.9968116 6.420136e-05 x3 <--> x3
the44  0.6572057 0.13516903  4.8621028 1.161453e-06 x4 <--> x4
the55  0.5867210 0.14763139  3.9742289 7.060762e-05 x5 <--> x5
the66  0.4088012 0.09494782  4.3055356 1.665822e-05 x6 <--> x6
the77  0.4957432 0.09279611  5.3422842 9.178263e-08 x7 <--> x7
psi1   0.6000819 0.29900741  2.0069131 4.475892e-02 desempeno <--> desempeno
psi2   0.4158819 0.11732949  3.5445641 3.932628e-04 satisfaccion <--> satisfaccion
psi12  0.1420264 0.24287629  0.5847686 5.587033e-01 satisfaccion <--> desempeno
phi12  0.1823314 0.07602293  2.3983732 1.646808e-02 autoestima <--> motivacion
phi13 -0.2041769 0.08172714 -2.4982749 1.247994e-02 iverbal <--> motivacion
phi23 -0.2503970 0.08755973 -2.8597279 4.240047e-03 iverbal <--> autoestima
iverbal
```

Cuadro 15.4.: Resultados estandarizados de la estimación

```
          Std. Estimate
1            1.0000000            x1 <--- desempeno
2            0.8196844         x2 <--- satisfaccion
3   lam1     0.7649285         x3 <--- satisfaccion
4            0.5854864           x4 <--- motivacion
5   lam2     0.6428678           x5 <--- motivacion
6            0.7688945           x6 <--- autoestima
7   lam3     0.7101104           x7 <--- autoestima
8            1.0000000              x8 <--- iverbal
9  beta1     0.3759410   satisfaccion <--- desempeno
10  gam1     0.3706142  satisfaccion <--- motivacion
11 beta2    -0.2117068   desempeno <--- satisfaccion
12  gam2     0.7374178     desempeno <--- autoestima
13  gam3    -0.1786967        desempeno <--- iverbal
14  phi1     1.0000000    motivacion <--> motivacion
15  phi2     1.0000000    autoestima <--> autoestima
16  phi3     1.0000000          iverbal <--> iverbal
17           0.0000000                    x1 <--> x1
18 the22     0.3281175                    x2 <--> x2
19 the33     0.4148843                    x3 <--> x3
20 the44     0.6572057                    x4 <--> x4
21 the55     0.5867210                    x5 <--> x5
22 the66     0.4088012                    x6 <--> x6
23 the77     0.4957432                    x7 <--> x7
24           0.0000000                    x8 <--> x8
25  psi1     0.6000819       desempeno <--> desempeno
26  psi2     0.6189801 satisfaccion <--> satisfaccion
27 psi12     0.1732697   satisfaccion <--> desempeno
28 phi12     0.4050212    autoestima <--> motivacion
29 phi13    -0.3487303        iverbal <--> motivacion
30 phi23    -0.3256584        iverbal <--> autoestima
```

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 15.5.: Residuos estandarizados de la estimación

	x1	x2	x3	x4	x5	x6	x7	x8
x1	0.0000	-0.0017	0.0023	-0.0242	0.0208	0.0000	0.0046	0.0056
x2	-0.0017	0.0000	0.0000	-0.0231	0.0369	0.0188	-0.0172	0.0617
x3	0.0023	0.0000	0.0000	0.0560	-0.0226	0.0793	0.0880	0.1651
x4	-0.0242	-0.0231	0.0560	0.0000	-0.0114	0.0187	0.0036	0.0052
x5	0.0208	0.0369	-0.0226	-0.0114	0.0000	-0.0392	-0.0109	-0.0528
x6	0.0000	0.0188	0.0793	0.0187	-0.0392	0.0000	0.0000	-0.0436
x7	0.0046	-0.0172	0.0880	0.0036	-0.0109	0.0000	0.0000	0.0573
x8	0.0056	0.0617	0.1651	0.0052	-0.0528	-0.0436	0.0573	0.0000

En lo que respecta a la matriz residual de varianzas y covarianzas, `sem {sem}` ofrece la versión estandarizada de esta (cuadro 15.5), y nosotros podemos generar un gráfico de su distribución (figura 15.4). Recordemos que esta matriz es la diferencia entre la matriz estimada y la muestral y en el caso ideal debería ser una matriz nula. En todo caso, se espera que los residuos sean pequeños y estén simétricamente distribuidos. Residuos grandes asociados a ciertas variables pueden indicar una incorrecta especificación del modelo.

```
a<-round(standardized.residuals(fit),4)
b<-c(a[lower.tri(a)])

qplot(b,
      geom="histogram",
      binwidth = 0.01,
      main = "Grafico de residuos",
      xlab = "Tamano del residuo",
      ylab = "frecuencia",
      breaks=seq(-0.5,0.5,by=0.05),
      fill=I("black"),
      col=I("white"),
      alpha=I(.7),
      xlim=c(-0.5,0.5))
```

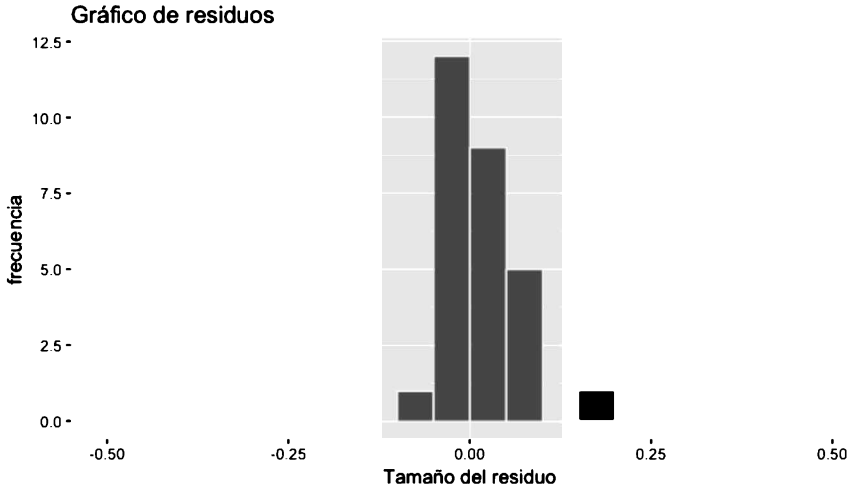
El análisis de esta información muestra que los residuos no son demasiado altos y su distribución es prácticamente simétrica y centrada en cero. Debemos prestar atención al hecho de que la mayoría de los residuos más elevados van asociados a la variable X_8. Este hecho apunta a algún problema de especificación de alguno de los parámetros del modelo, lo que deberá considerarse cuando se contemplen los indicadores para la reespecificación del mismo. El cuadro 15.6 ofrece el resto de indicadores de bondad de ajuste.

Limitándonos a los que se comentaron en el capítulo anterior, vemos que el estadístico χ^2 del modelo propuesto (15.05), además de no ser significativo (p=0,30), permite aceptar la hipótesis nula de igualdad entre las matrices de covarianzas muestral y estimada.

Por otra parte, respecto a los estadísticos *ad hoc* que presentamos en el capí-

Estadísticos e-Books & Papers

Figura 15.4.: Gráfico de residuos estandarizados



Cuadro 15.6.: Indicadores de ajuste del modelo

```
Model Chisquare =  15.05892   Df =  13 Pr(>Chisq) = 0.3036888
 Goodness-of-fit index =  0.9698424
 Adjusted goodness-of-fit index =  0.9164866
 RMSEA index =  0.03617893   90% CI: (NA, 0.1007046)
 Bentler-Bonett NFI =  0.9412905
 Tucker-Lewis NNFI =  0.9805924
 Bentler CFI =  0.9909893
 Bentler RNI =  0.9909893
 Bollen IFI =  0.9915444
 SRMR =  0.04166824
 AIC =  61.05892
 AICc =  26.32423
 BIC =  -47.39335
 CAIC =  -60.39335
```

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

tulo anterior, vemos que todos cumplen los requisitos exigidos. Los estadísticos NFI (0,9413), TLI (0,9806), CFI (0,9910), GFI (0,9698) y AGFI (0,9165) toman valores superiores a 0,9. Estos estadísticos confirman, también, el buen ajuste del modelo. Constatado que el ajuste del modelo es razonable, debe procederse a la interpretación del mismo y, si así se deriva de los resultados, a su posible reespecificación.

15.6. Interpretación del modelo

La interpretación del modelo pasa por establecer qué parámetros estimados han resultado ser significativos. Aunque esta información ya está mostrada en el cuadro, la mostramos de nuevo en los cuadros 15.3 (versión no estandarizada) y 15.4 (versión estandarizada); aprovecharemos esta ocasión para presentar los resultados de tal manera que se ilustre claramente la conexión entre los parámetros estimados y los componentes de la matriz de varianzas estimada $\hat{\mathbf{\Sigma}}$ que había que estimar para minimizar su diferencia con la matriz muestral $\mathbf{S}$ y que recogíamos en la expresión (15.5). En el caso global que veremos a continuación ya haremos una presentación de resultados acorde con los requerimientos habituales de las publicaciones de investigación. El cuadro 15.7 resume esta información.

Si recordamos los objetivos del trabajo de Bagozzi (1980), este autor trataba de determinar, en primer lugar, si la satisfacción en el trabajo era antecedente o consecuencia del desempeño. Para ello en el modelo había planteado la doble relación causal entre estos dos factores latentes, cuyos coeficientes de regresión (véase la figura 15.1) eran β_{12} (influencia de la satisfacción sobre el desempeño) y β_{21} (influencia del desempeño sobre la satisfacción). Si nos fijamos en el valor del estadístico t para ambos parámetros en el cuadro 15.7, comprobamos como β_{12} ($t = -0{,}53$) no es significativo, mientras que β_{21} ($t = 2{,}10$) sí que lo es para $p < 0{,}05$. Esto nos permite concluir que la satisfacción en el trabajo es una de las variables determinantes del desempeño de un vendedor en su puesto de trabajo.

La otra variable que influye directa y positivamente sobre la satisfacción en el lugar de trabajo de un vendedor es la motivación para el logro ($\gamma_{21} = 0{,}519; t = 2{,}26$), es decir, que, cuanto más valor da a las recompensas que puede recibir un vendedor por hacer bien su trabajo, mayor es su satisfacción.

Indirectamente, es decir, a través del desempeño como variable mediadora, influyen en la satisfacción en el trabajo la autoestima ($\gamma_{12} = 0{,}959; t = 3{,}46$) y la inteligencia verbal ($\gamma_{13} = -0{,}179; t = -1{,}74$). En el primer caso, el buen concepto que un vendedor tiene de sí mismo le lleva a obtener mejores resultados y, a través de los mismos, mejorar su satisfacción. Sin embargo puede sorprender el signo negativo del coeficiente asociado a la inteligencia verbal de hecho era contrario al hipotetizado por Bagozzi (1980). El autor explica este signo, aunque la variable no es estrictamente significativa, a partir del hecho de que aquellos individuos de gran inteligencia pueden llegar a encontrar aburrido

Estadísticos e-Books & Papers

Cuadro 15.7.: Parámetros estimados

Matriz	Parámetro	Estimación		
		No std	Std	t
$\mathbf{\Lambda_x}$	λ_{11}^x	1,000†	0,585	–
	λ_{21}^x	1,098	0,643**	3,16
	λ_{32}^x	1,000†	0,769	–
	λ_{42}^x	0,924	0,710**	6,18
	λ_{53}^x	1,000†	1,000	–
$\mathbf{\Lambda_y}$	λ_{11}^y	1,000†	1,000	–
	λ_{22}^y	1,000†	0,820	–
	λ_{32}^y	0,933	0,765**	5,75
$\mathbf{B}$	β_{12}	-0,258	-0,212	-0,53
	β_{21}	0,308	0,376*	2,10
$\mathbf{\Gamma}$	γ_{12}	0,959	0,737**	3,46
	γ_{13}	-0,179	-0,179	-1,74
	γ_{21}	0,519	0,371*	2,26
$\mathbf{\Psi}$	ψ_{11}	0,600	0,600*	2,01
	ψ_{22}	0,416	0,619**	3,54
	ψ_{12}	0,142	0,142	0,58
$\mathbf{\Phi}$	ϕ_{11}	0,343	1,000*	2,39
	ϕ_{22}	0,591	1,000**	4,18
	ϕ_{33}	1,000	1,000**	7,78
	ϕ_{12}	0,182	0,405*	2,40
	ϕ_{13}	-0,204	-0,349*	-2,50
	ϕ_{23}	-0,250	-0,326	-2,86
$\mathbf{\Theta}_\delta$	θ_1^δ	0,657	0,657	4,86
	θ_2^δ	0,587	0,587	3,97
	θ_3^δ	0,409	0,409**	4,31
	θ_4^δ	0,496	0,496**	5,34
	θ_5^δ	0,000†	0,000	–
$\mathbf{\Theta}_\varepsilon$	θ_1^ε	0,000†	0,000	–
	θ_2^ε	0,328	0,328**	2,97
	θ_3^ε	0,415	0,415**	4,00

†Fijados para identificación

© Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 15.8.: Índices de modificación

```
5 largest modification indices, A matrix (regression coefficients):
   iverbal<-x3 autoestima<-x3         x8<-x3    x3<-iverbal         x3<-x8
      7.190381       6.747341       6.722907       4.941608       4.941608

  5 largest modification indices, P matrix (variances/covariances):
            iverbal<->x3                   x8<->x3           autoestima<->x3
                5.948519                  5.191360                  4.675407
                 x5<->x4 motivacion<->satisfaccion
                4.012981                  4.012967
```

y poco desafiante su trabajo, llevándoles este hecho a poner menos esfuerzo, obteniendo peores resultados. Estaríamos ante un fenómeno equivalente a la sobrecualificación de muchos trabajadores que se ha demostrado que es un elemento incentivador de la movilidad laboral.

Hasta aquí los principales resultados. Si nos fijamos en el cuadro 15.7, comprobaremos que todo el resto de parámetros son significativos a excepción de la covarianza entre los términos de error asociados a los factores satisfacción y desempeño ($\psi_{21} = 0{,}142; t = 0{,}58$). Recordemos que esta covarianza se introdujo con fines de identificación de la relación no recursiva. Al no resultar significativa una de las relaciones entre satisfacción y desempeño, parece natural que esta covarianza entre los errores desaparezca.

Todo lo expuesto podría ser suficiente para avalar el trabajo realizado por el investigador: buen ajuste del modelo y resultados significativos y relevantes en su conjunto. Sin embargo cabe la posibilidad de que el modelo sea susceptible de algún tipo de reespecificación que lo mejore, dotando de mayor peso a los resultados obtenidos.

15.7. Reespecificación del modelo

El lector debe remitirse al epígrafe equivalente del tema anterior para revisar los motivos que pueden llevar a reespecificar un modelo (básicamente mejora del ajuste o contraste de alguna hipótesis teórica) y las precauciones que deben tenerse respecto a la incorporación o eliminación de relaciones no soportadas por marco teórico alguno. La forma de solicitar los índices de modificación en `sem{sem}` es la siguiente:

```
MI<-modIndices(fit) print(MI,n.largest=5,round=3)
```

La salida se muestra en el cuadro 15.8.

Vemos, en primer lugar, que, en la propuesta sobre la matriz que denomina A y que contiene las propuestas de modificación de los coeficientes de regresión, existen una serie de coeficientes que no nos planteamos modificar en la estimación del modelo estructural puesto que, de ser necesario considerarlo, se

Estadísticos e-Books & Papers

debería haber realizado en la estimación del CFA por afectar al instrumento de medida que se validó entonces. Nos referimos a sugerencias como que el indicador X_3 cargue también sobre la inteligencia verbal (`iverbal<-x3`) o sobre la autoestima (`autoestima<-x3`) e incluso a relaciones estructurales entre los indicadores (`x8<-x3`). No se realiza sugerencia alguna de relaciones entre factores que serían las que cabría considerar en esta fase.

Si nos fijamos en la matriz $\boldsymbol{P}$, que contiene las propuestas sobre las varianzas y covarianzas, se plantea la introducción de covarianzas entre errores, que tampoco consideraremos porque implica reconocer que existen elementos fuera de nuestro modelo que provocan esa covarianza y que no hemos sido capaces de identificar, máxime cuando no tenemos problemas de ajuste.

15.8. Un ejemplo completo de modelo de ecuaciones estructurales

El caso 15.1 pretendía ser un caso con un número limitado de indicadores que permitiera la ilustración de la notación y la formulación matemática sin que la complejidad del mismo entorpeciera esta tarea. También pretendía introducir en la fase de identificación dos situaciones que, no siendo frecuentes, conviene que el investigador sepa cómo abordar: variables latentes con un único indicador y relaciones no recursivas.

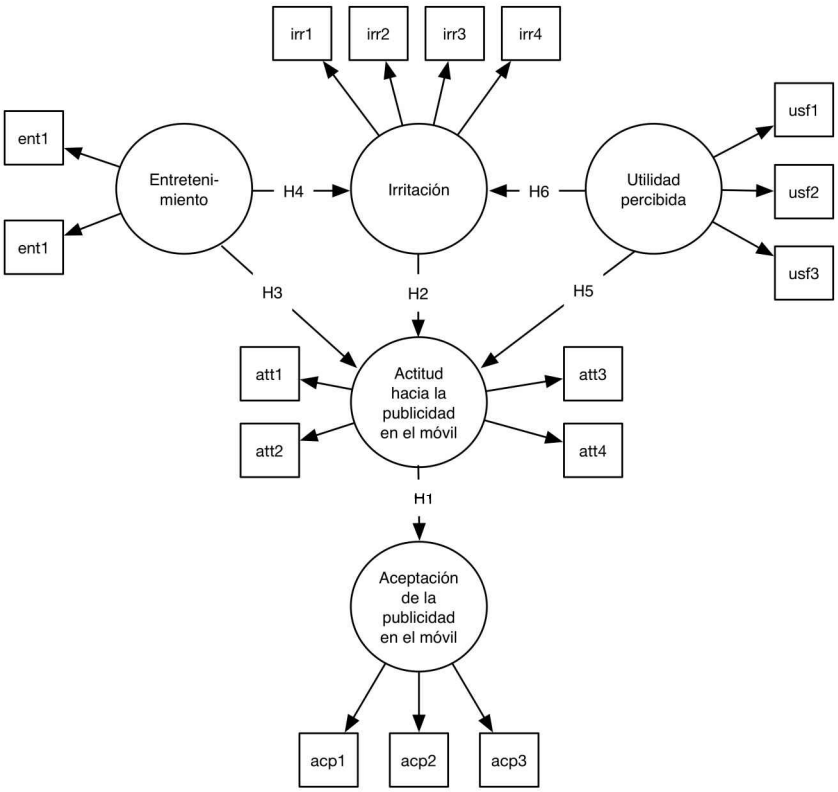
Para que el lector no pierda la visión de conjunto de los tres últimos temas como un proceso que se inicia con la estimación de un CFA, se usan sus resultados para validar el instrumento de medida, antes de proceder a la estimación del modelo estructural, vamos a cerrar este tema con el caso que hemos venido usando de manera transversal a lo largo de esta secuencia. Con el instrumento de medida ya validado, nos centraremos únicamente en los cambios que hay que realizar en la sintaxis del CFA para identificar el modelo, además de para introducir las ecuaciones estructurales, y en los resultados de la estimación de estas últimas incluyendo la forma habitual de presentación de los mismos en las publicaciones.

Caso 13.2. Determinantes de la aceptación de la publicidad en el móvil
Recordemos que el caso que vamos a cerrar era el que presentábamos en la figura 13.1 y que reproducimos de nuevo en la figura 15.5 para que el capítulo sea autocontenido y que se corresponde con el trabajo de Aldás *et al.* (2013). Remitimos a este trabajo para la justificación teórica de las relaciones, pero, básicamente, lo que planteaba era que la aceptación de la publicidad en el móvil viene condicionada porque el usuario perciba una utilidad en la misma y que el entretenimiento que genere su creatividad sea capaz de minorar la irritación que la intrusión pueda causar. Para estimar el modelo estructural que aparece en la figura 13.1 era necesario, previamente —y así lo hicimos— estimar un CFA para, con la información que proporciona (capítulo 13), evaluar la fiabilidad y validez del instrumento de medida (capítulo 14). El modelo lo continuaremos

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 15.5.: Ejemplo de modelo de ecuaciones estructurales



Fuente: Aldás *et al.* (2013)

estimando con `lavaan{lavaan}`.

Partimos de la sintaxis del CFA estimado en el capítulo 13 y, sobre ella, introduciremos las modificaciones necesarias que se derivan de su transformación en un modelo de ecuaciones estructurales.

```
modelo.cfa <- '

# Modelo de medida
actitud          =~ att1+att2+att3+att4
entretenimiento  =~ ent1+ent2
utilidad         =~ usf1+usf2+usf3
irritacion       =~ irr1+irr2+irr3+irr4
aceptacion       =~ acc1+acc2+acc3
```


Estadísticos e-Books & Papers

```
#Varianzas de los factores
actitud~~actitud
entretenimiento~~entretenimiento
utilidad~~utilidad
irritacion~~irritacion
aceptacion~~aceptacion

#Covarianzas
actitud~~entretenimiento
actitud~~utilidad
actitud~~irritacion
actitud~~aceptacion
entretenimiento~~utilidad
entretenimiento~~irritacion
entretenimiento~~aceptacion
utilidad~~irritacion
utilidad~~aceptacion
irritacion~~aceptacion

#Varianzas de los t<U+00E9>rminos de error
att1~~att1 att2~~att2 att3~~att3 att4~~att4 ent1~~ent1 ent2~~ent2
usf1~~usf1 usf2~~usf2 usf3~~usf3 irr1~~irr1 irr2~~irr2 irr3~~irr3
irr4~~irr4 acc1~~acc1 acc2~~acc2 acc3~~acc3
'
```

El primer paso es la incorporación de las ecuaciones estructurales que se corresponden con las hipótesis de la figura 15.5. Esta introducción desencadenará una serie de cambios en la identificación que habrá que cubrir. Las ecuaciones estructurales son las siguientes:

```
# Relaciones estructurales

irritacion   ~ entretenimiento+utilidad
actitud      ~ entretenimiento+irritacion+utilidad
aceptacion   ~ actitud
```

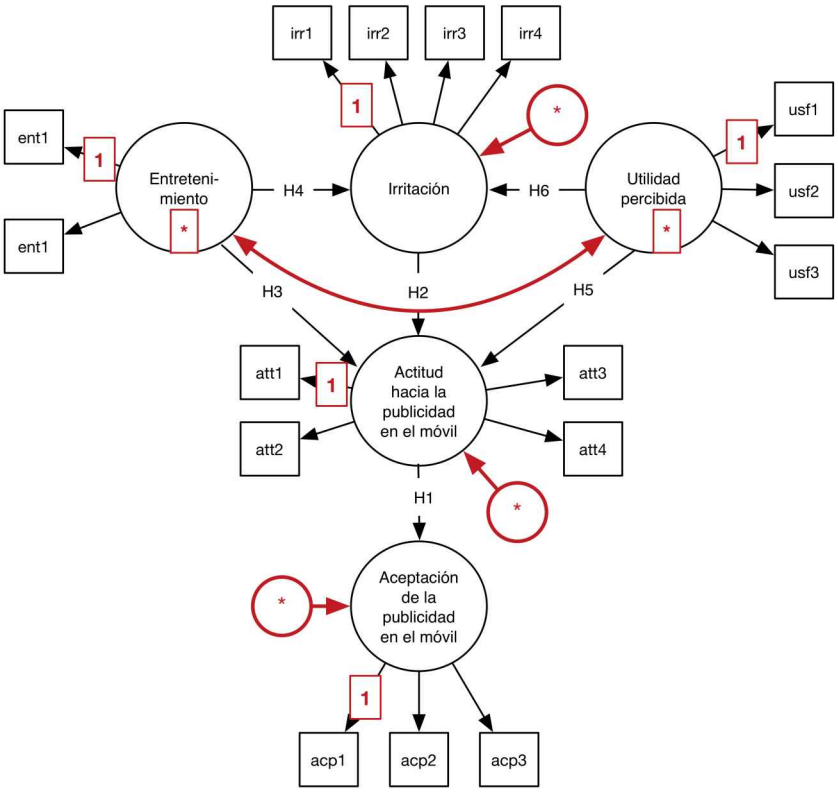
Al introducir estas relaciones se generan estos cambios teóricos que ilustramos en la figura 15.6 e iremos trasladando a la sintaxis:

- En el CFA todos los factores eran independientes y por eso planteábamos covarianzas entre ellos, sin embargo, al introducir las relaciones estructurales, la irritación, la actitud y la aceptación pasan a ser **factores dependientes**.
- Al ser factores dependientes será necesario i**ntroducir sus términos de error**, cuyas varianzas habrá que estimar:

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 15.6.: Ilustración de la identificación del modelo



Estadísticos e-Books & Papers

```
# Varianzas de los errores de los factores

actitud~~actitud
irritacion~~irritacion
aceptacion~~aceptacion
```

- Realmente en términos de sintaxis no cambia nada porque, si se revisa la sintaxis del CFA, le estábamos pidiendo que estimara las varianzas de esos factores, pero el programa entiende que al ser factores dependientes no le estamos pidiendo ahora sus varianzas —que no pueden tener— sino las de sus errores. Lo que ocurre es que no todos los programas son tan flexibles y es importante que el investigador sepa que deben eliminarse de la estimación las varianzas de los factores dependientes y pasar a estimar las de sus errores.
- Al haberse convertido esos factores en dependientes, **tampoco puede estimarse ninguna covarianza donde estén implicados**, por lo que habrá que quitarlas de la sintaxis. Nosotros somos partidarios de quitar las covarianzas y cualquier otro parámetro convirtiéndolo en comentario (#) de tal forma que se pueda siempre seguir qué se ha hecho. Vemos que solo queda una única covarianza entre los dos únicos factores independientes: entretenimiento y utilidad.

```
#Covarianzas
#actitud~~entretenimiento
#actitud~~utilidad
#actitud~~irritacion
#actitud~~aceptacion

entretenimiento~~utilidad

#entretenimiento~~irritacion
#entretenimiento~~aceptacion
#utilidad~~irritacion
#utilidad~~aceptacion
#irritacion~~aceptacion
```

- El cambio fundamental se produce en la **identificación de la escala**. Recordemos que en el CFA, como todos los factores eran independientes, su varianza era estimable e identificábamos la escala fijándola a 1 (se consigue en la sintaxis escribiendo `1*`) y estimando todas las cargas, porque nos era necesario para la validación. Pero ahora tres factores —irritación, actitud y aceptació— son dependientes, no se estima su varianza y

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

la única forma de identificar es fijando a 1 una carga para cada uno de estos factores. Nosotros, por simplicidad, preferimos fijar a 1 una carga también para los que son independientes y podríamos optar por fijar la varianza. Lo que no se nos debe olvidar es, en la sintaxis, quitar la opción de `std.lv=TRUE` (poner `std.lv=FALSE`), que ya explicamos en el CFA y que fija las varianzas a 1.

```
actitud          =~ 1*att1+att2+att3+att4
entretenimiento  =~ 1*ent1+ent2
utilidad         =~ 1*usf1+usf2+usf3
irritacion       =~ 1*irr1+irr2+irr3+irr4
aceptacion       =~ 1*acc1+acc2+acc3

fit <- lavaan(modelo.cfa, data=datos, std.lv=FALSE, mimic="eqs",
estimator="ML", verbose=TRUE, warn=TRUE)
```

Con todos estos cambios introducidos, la sintaxis resultante es la siguiente (hemos obviado las líneas de comentario de los parámetros que no se estiman para que sea más sintética, pero el lector la encontrará completa en la página web del libro):

```
modelo.cfa <- '
# Modelo de medida
actitud          =~ 1*att1+att2+att3+att4
entretenimiento  =~ 1*ent1+ent2
utilidad         =~ 1*usf1+usf2+usf3
irritacion       =~ 1*irr1+irr2+irr3+irr4
aceptacion       =~ 1*acc1+acc2+acc3
# Relaciones estructurales
irritacion   ~entretenimiento+utilidad
actitud      ~entretenimiento+irritacion+utilidad
aceptacion   ~actitud
#Varianzas de los factores
entretenimiento~~entretenimiento  utilidad~~utilidad
#Covarianzas
entretenimiento~~utilidad
#Varianzas de los terminos de error
att1~~att1 att2~~att2 att3~~att3 att4~~att4 ent1~~ent1 ent2~~ent2
usf1~~usf1 usf2~~usf2 usf3~~usf3 irr1~~irr1 irr2~~irr2 irr3~~irr3
irr4~~irr4 acc1~~acc1 acc2~~acc2 acc3~~acc3
#errores factores dependientes
actitud~~actitud
irritacion~~irritacion
aceptacion~~aceptacion
```


Estadísticos e-Books & Papers

Cuadro 15.9.: Ajuste del modelo

```
lavaan (0.5-22) converged normally after  41 iterations

  Number of observations                           353

  Estimator                                         ML
  Minimum Function Test Statistic              261.485
  Degrees of freedom                                97
  P-value (Chi-square)                           0.000

Model test baseline model:

  Minimum Function Test Statistic             2993.948
  Degrees of freedom                               120
  P-value                                        0.000

User model versus baseline model:

  Comparative Fit Index (CFI)                    0.943
  Tucker-Lewis Index (TLI)                       0.929


Root Mean Square Error of Approximation:

  RMSEA                                          0.069
  90 Percent Confidence Interval          0.059  0.080
  P-value RMSEA <= 0.05                          0.001

Standardized Root Mean Square Residual:

  SRMR                                           0.079
```

```
'
#Estimacion del modelo
fit <- lavaan(modelo.cfa, data=datos, std.lv=FALSE, mimic="eqs",
estimator="ML", verbose=TRUE, warn=TRUE)
```

El primer paso, como ocurría con el CFA, es evaluar el ajuste del modelo (cuadro 15.9), comprobar que tenemos grados de libertad, porque la sobreidentificación no la hemos comprobado al identificar el modelo en los pasos anteriores, y fijarnos ya en la parte estructural de los resultados, dado que la parte correspondiente al instrumento de medida (cuadro 15.10) ya fue objeto de análisis en el proceso de evaluación del mismo.

Podemos comprobar que, aunque la ji-cuadrado es significativa [$\chi^2(97) = 261{,}49; p < 0{,}01$], los índices comparativos son superiores a 0,90 ($CFI = 0{,}943$ y $TLI = 0{,}929$), los residuos son inferiores en promedio a 0,08 ($SRMR = 0{,}079$) y el RMSEA es también inferior a 0,08 lo que se considera aceptable ($RMSEA[90CI] = 0{,}068[0{,}059|0{,}080]$). Por lo tanto, el ajuste del modelo es suficientemente sólido como para pasar a interpretar los resultados, cuya salida está en el cuadro 15.10 pero que nosotros hemos elaborado como suele mostrarse en las publicaciones en el cuadro 15.11. La figura 15.7, obtenida de la salida de `lavaan`, sirve al investigador para visualizar los resultados, pero no sería adecuada para su publicación.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

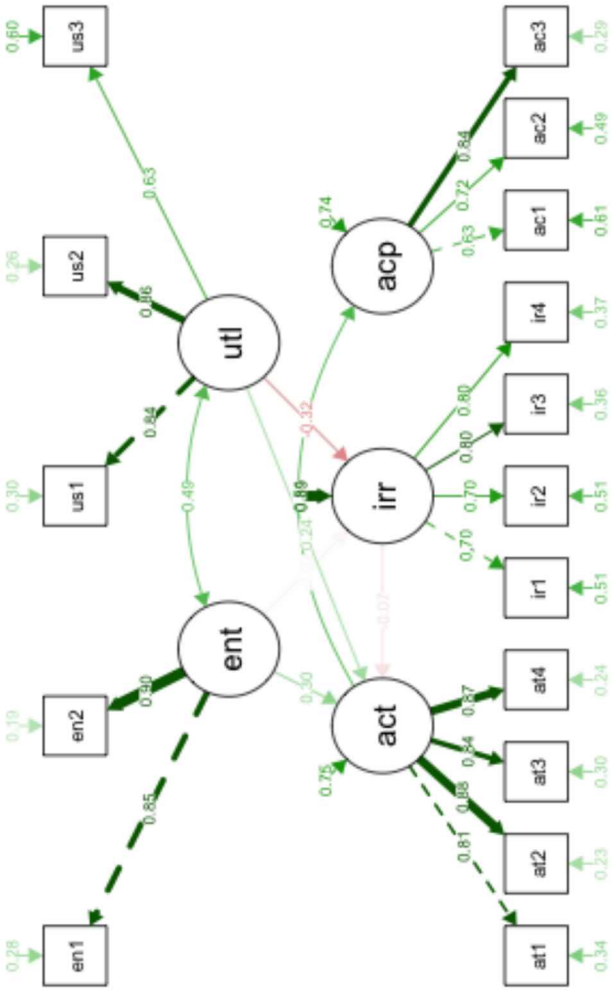


Figura 15.7.: Gráfico de resultados

Estadísticos e-Books & Papers

Cuadro 15.10.: Estimación de la parte estructural

```
Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  irritacion ~
    entretenimient   -0.029    0.089   -0.326    0.745   -0.023   -0.023
    utilidad         -0.329    0.079   -4.170    0.000   -0.315   -0.315
  actitud ~
    entretenimient    0.291    0.064    4.541    0.000    0.299    0.299
    irritacion       -0.058    0.046   -1.267    0.205   -0.074   -0.074
    utilidad          0.199    0.057    3.498    0.000    0.242    0.242
  aceptacion ~
    actitud           0.523    0.072    7.294    0.000    0.512    0.512
```

Como síntesis podemos concluir que la actitud positiva hacia la publicidad móvil es fundamentalmente una tarea que hay que lograr consiguiendo que añada un valor que genere percepción de utilidad. El efecto de la utilidad es significativo y positivo sobre la actitud ($H5 : \beta = 0{,}242; p < 0{,}01$) y también como reductor de la irritación que esta pueda generar ($H6 : \beta = -0{,}315; p < 0{,}01$), aunque la irritación no actúa de manera directa sobre la actitud ($H2 : \beta = -0{,}074; p > 0{,}05$). El papel del entretenimiento, es decir, que la publicidad utilice soportes que aporten valor lúdico al sujeto, también es importante para generar una actitud positiva hacia este soporte ($H3 : \beta = 0{,}299; p < 0{,}01$), aunque no como reductor de la irritación ($H4 : \beta = -0{,}023; p > 0{,}05$). Como en todos los modelos de aceptación de la tecnología, la actitud es un buen predictor de la aceptación ($H1 : \beta = 0{,}512; p < 0{,}01$).

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 15.11.: Contraste de hipótesis

Hipótesis	β estandarizado	Valor t
H1: Actitud publicidad móvil→Aceptación publicidad móvil	0.512**	7.29
H2: Irritación→Actitud publicidad móvil	-0.074	-1.27
H3: Entretenimiento→Actitud publicidad móvil	0.299**	4.54
H4: Entretenimiento→Irritación	-0.023	-0.33
H5: Utilidad percibida→Actitud publicidad móvil	0.242**	3.50
H6: Utilidad percibida→Irritación	-0.315**	-4.17

$\chi^2(97) = 261{,}49^{**}; CFI = 0{,}943; TLI = 0{,}929; RMSEA(90\,\%CI) = 0{,}069(0{,}059|0{,}080)$
$^{**}p < 0{,}01;\ ^{*}p < 0{,}05$

© Ediciones Paraninfo

Estadísticos e-Books & Papers

Estadísticos e-Books & Papers

16. Modelos de ecuaciones estructurales: modelos de estructuras de varianza (PLS-SEM)

16.1. Introducción

Cuando Joreskog (1967) crea un algoritmo operativo para la estimación por máxima verosimilitud de modelos estructurales con variables latentes, pero, sobre todo, desde que posteriormente implementa ese algoritmo en un *software* comercial razonablemente sencillo de utilizar, LISREL (Jöreskog, 1970), estaba sentando las bases para que los modelos basados en covarianzas (CBSEM) se convirtieran en la principal —si no la única— aproximación utilizada por los investigadores para modelizar los modelos de ecuaciones estructurales (SEM).

Sin embargo, años atrás, Herman Wold, que había sido el director de tesis de Jöreskog, había cuestionado que el enfoque de máxima verosimilitud (ML) fuera una alternativa razonable para abordar la estimación de estos modelos. El algoritmo ML es un enfoque muy paramétrico, plantea muchas restricciones respecto al tamaño muestral y las propiedades que han de tener los datos. Como señalan Tenenhaus *et al.* (2005), Wold consideraba que era necesario un enfoque menos exigente (que denominó *soft modelling* en contraposición al enfoque *hard modelling* de Jöreskog), más cercano a los tamaños muestrales y características reales que tienen los datos con los que trabajamos habitualmente. A este enfoque lo denominó *Partial Least Squares* (PLS) que fue rápidamente concretado (Wold, 1973) en un algoritmo operativo que llamó NIPALS (=*Non linear Iterative PArtial Least Squares*). Hay que esperar hasta 1982 para encontrar una presentación más ordenada de este enfoque aplicado a modelos estructurales con variables latentes (Wold, 1982).

El enfoque de PLS (que denotaremos a partir de ahora como PLS-SEM) apenas había sido utilizado desde su creación para resolver problemas de SEM. La razón, probablemente, quepa buscarla en la inexistencia de un *software* sencillo de utilizar hasta fechas recientes. `LVPLS 1.8` (Lohmoller, 1987), ha sido la única herramienta de la que se ha dispuesto durante mucho tiempo y no era, desde luego, una alternativa intuitiva para un usuario estándar. Con el comienzo de siglo, los esfuerzos de Chin (2000) con `PLS-Graph` (beta), Ringle *et al.* (2005) con `SmartPLS` (beta), Fu con `VisualPLS`, ScriptWarp Inc, con `WarpPLS` (softwa-

Estadísticos e-Books & Papers

re comercial) o Tenenhaus, Vizi, Chatelin y Lauro con `XLSTAT-PLSPM` (software comercial) han permitido que los investigadores dispongan de herramientas razonablemente sencillas para abordar sus estimaciones, provocando la extensión de la técnica, como ponen de manifiesto trabajos como los de Henseler *et al.* (2009), Reinartz *et al.* (2009), Hair *et al.* (2012a) o Hair *et al.* (2012b). Existen también varios paquetes de `R` que permiten la aplicación de PLS-SEM en este entorno: `plspm`[1], `matrixpls`[2] y `semPLS`[3].

En mi opinión, sin embargo, hay otras razones más allá de la disponibilidad de *software* que explican el interés creciente en esta herramienta. Para mí, la clave tiene sus orígenes en el trabajo de Diamantopoulos y Winklhofer (2001), quienes alertaron que los investigadores estábamos modelizando de manera automática como reflectivas las variables latentes que incorporábamos a nuestras investigaciones como consecuencia, probablemente, de la inercia derivada del enfoque de la teoría clásica de la medición en el campo de la psicología. Otros autores pusieron cifras a estos errores. Así Jarvis *et al.* (2003) cuantificaron en casi un tercio de los trabajos los errores de especificación reflectiva en las principales revistas de las áreas de marketing y dirección de empresas.

Ante esta situación, los investigadores se encuentran con una grave dificultad a la hora de incorporar los constructos formativos a sus modelos con el enfoque clásico de los CBSEM: por definición, un constructo formativo no está identificado. La única forma de hacerlo es, a grandes rasgos, o incorporar indicadores reflectivos que convivan con los formativos (modelo MIMIC) o que el constructo formativo esté unido por relaciones estructurales con al menos dos constructos reflectivos. En el primer caso, el investigador tiene que haber previsto la situación incorporando preguntas adicionales a su cuestionario. En el segundo caso, el modelo que se quiere confirmar puede tener esa característica o no. Por lo tanto, la dificultad para que el enfoque CBSEM aborde los constructos formativos es muy elevada y, en ese momento, la academia volvió su mirada al enfoque alternativo del PLS-SEM.

Como demostraremos al explicar su algoritmo, PLS-SEM apenas se ve influido por el hecho de que la variable latente sea formativa o reflectiva y esto incrementó inmediatamente su popularidad. Diamantopoulos *et al.* (2008) demostraron, en un trabajo que revisaba el uso de constructos formativos, que casi la mitad de ellos ya había empleado PLS-SEM como herramienta analítica.

El impulso final a la creciente extensión y aceptación de PLS-SEM como enfoque de modelización estructural ha venido de una serie de trabajos firmados por relevantes investigadores en el ámbito del marketing que han fijado las buenas prácticas en la estimación de modelos mediante PLS-SEM (Hair *et al.*, 2012a, 2013), pero, sobre todo, de la elaboración del primer manual introductorio que da guías claras a los investigadores con una formación básica en estadística (Hair *et al.*, 2014b).

Sin embargo, es muy importante señalar que PLS-SEM es una herramien-

[1] `https://cran.r-project.org/web/packages/plspm/index.html`.
[2] `https://cran.r-project.org/web/packages/matrixpls/index.html`.
[3] `https://cran.r-project.org/web/packages/semPLS/index.html`.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

ta que también sufre de importantes limitaciones que serán abordadas en este capítulo. Su algoritmo necesita que todas las variables latentes tengan asignados indicadores. Pero, en el campo del marketing, por ejemplo, se dan con mucha frecuencia variables latentes que están formadas por dimensiones, es decir, constructos de segundo orden. Por ejemplo, es habitual que la confianza se modelice en marketing como una variable latente que tiene tres dimensiones que también son variables latentes: honestidad, benevolencia y competencia. Las dimensiones se convierten en el instrumento de medida de la confianza y esta no tiene indicadores específicos. Nos encontramos ante una situación que es necesario resolver si queremos que PLS-SEM pueda enfrentarse a problemas habituales en nuestras áreas de trabajo. Tampoco puede estimar modelos con relaciones no recursivas (dobles relaciones estructurales) ni modelos en que las variables latentes no estén conectadas con relaciones estructurales, lo que impide estimar modelos equivalentes a los del análisis factorial confirmatorio, hecho que nos lleva a preguntarnos —y luego responder— cómo se valida el instrumento de medida cuando se estima mediante PLS-SEM.

Si el lector está interesado en una introducción en español a PLS-SEM pero utilizando *software* comercial como `SmartPLS 3.0` puede consultar el libro de Aldás (2016), en el que también se abordan cuestiones no tratadas en este manual como los constructos de segundo orden. La estructura que seguimos es muy similar y este capítulo es, en ese sentido, deudor del mencionado trabajo.

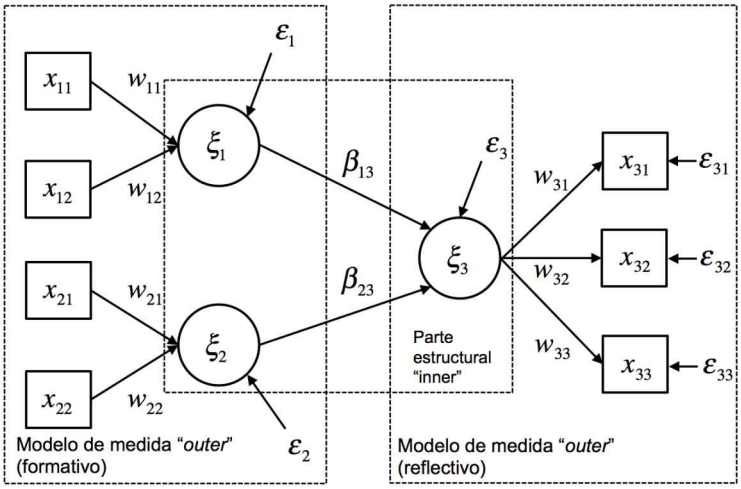
16.2. El algoritmo de estimación de los modelos PLS-SEM

Para ilustrar el algoritmo de PLS-SEM, del que nacen las ventajas y desventajas de esta técnica, lo aplicaremos a la estimación de un modelo muy sencillo representado en la figura 16.1. Es importante destacar que la figura 16.1 es un simple ejemplo, es decir, nada obliga a que el modelo que vaya a estimarse por PLS-SEM tenga que tener constructos formativos y reflectivos o que los formativos tengan que ser las variables independientes del modelo. Cualquier combinación es posible: todos los constructos podrían ser reflectivos, los reflectivos podrían ser los dependientes, etc.

La terminología clásica que se aplica en la estimación de modelos CBSEM se aplica básicamente también en los modelos PLS-SEM. Así, al instrumento de medida, es decir, el conjunto de indicadores que miden cada variable latente, aquí se renombran como la parte externa o *outer* del modelo, mientras que a la parte estructural, esto es, la que une mediante relaciones de dependencia/independencia a las variables latentes, el algoritmo las denota como parte interna o *inner* del modelo. Es muy importante señalar que la decisión de modelizar como formativo o reflectivo un constructo determinado no es una decisión técnica, sino conceptual. Depende de qué indicadores se elijan (porque un constructo no es por naturaleza ni formativo ni reflectivo, depende de los indicadores seleccionados). Para tener una guía para decantarse por una formu-

Estadísticos e-Books & Papers

Figura 16.1.: Ejemplo de un modelo estructural estimable mediante PLS-SEM



lación (formativa) u otra (reflectiva) para un constructo determinado, pueden revisarse los trabajos de Diamantopoulos y Winklhofer (2001), Jarvis *et al.* (2003) o Diamantopoulos *et al.* (2008).

Cuando un modelo se estima por el enfoque clásico de estimación basada en covarianzas CBSEM, la lógica de la estimación —como vimos en el capítulo correspondiente— es la siguiente. Se calcula la matriz de varianzas y covarianzas teórica que se deriva del modelo dibujado. A continuación se calcula la matriz de varianzas y covarianzas muestral entre los indicadores del modelo (esto es, la representación estadística de la realidad) y el algoritmo de máxima verosimilitud lo que hace, en esencia, es dar valores a los parámetros para estimar en la matriz teórica (cargas, coeficientes de regresión, covarianzas entre variables independientes, etc.) para intentar que la matriz resultante (matriz reproducida) se parezca lo máximo posible a la realidad (matriz muestral). En la medida en que esto se consigue, el ajuste es bueno y el modelo teórico se considera una representación plausible de la realidad. Realmente, el algoritmo de máxima-verosimilitud lo que hace es calcular los parámetros que minimizan las diferencias entre ambas matrices (Chin y Newsted, 1999). En ningún momento el algoritmo necesita realizar estimación alguna de los valores de los constructos, que continúan como variables latentes.

El enfoque de PLS-SEM es radicalmente contrario. En lugar de buscar ajuste entre las matrices de varianzas y covarianzas teórica y muestral, lo que se intenta es maximizar la varianza explicada de aquellas variables latentes dependientes por parte de las variables latentes independientes (Haenlein y Kaplan, 2004). Esto exige que primero se haya de obtener una estimación de las variables latentes, lo que se consigue como una combinación lineal de los indicadores

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

que los forman (Fornell y Bookstein, 1982). Los pesos que se utilizan para realizar estas combinaciones lineales se obtienen de tal forma que se maximice esa varianza explicada. Cuando esto se ha logrado, entonces se realiza un conjunto de regresiones para determinar los coeficientes de regresión de la parte estructural.

Probablemente la forma más sencilla de entender este algoritmo es aplicarlo al ejemplo de la figura 16.1. La descripción que realizamos está basada en Henseler *et al.* (2009) y la estructuramos en siete pasos sucesivos. Si se quiere una versión más detallada y con una mayor fundamentación estadística del algoritmo con algunas de sus variantes, recomendamos los trabajos de Tenenhaus *et al.* (2005) y Fornell y Cha (1994).

Paso 1. Inicialización

El primer paso es obtener una primera aproximación a los valores de las variables latentes (LV) a partir de sus indicadores. Lo más sencillo es, en esta primera iteración, que el factor se calcule como la suma de los indicadores, que es lo mismo que una combinación lineal donde los pesos de la combinación serían todos 1. Dependiendo del programa que se utilice, existe la posibilidad de otorgar en esta etapa otros pesos arbitrarios. El superíndice *outer,* que aparece en la estimación de los factores $\hat{\xi}^{outer}$, hace referencia a que esa estimación está realizada utilizando la parte *outer* del modelo, es decir, el instrumento de medida, los indicadores de los factores, tal y como se ilustraba en la figura 16.1. Las estimaciones de las variables latentes así obtenidas se estandarizan para tener una media de 0 y una desviación típica de 1.

$$\begin{gathered}\hat{\xi}_1^{outer} = x_{11} + x_{21} \\ \hat{\xi}_2^{outer} = x_{12} + x_{22} \\ \hat{\xi}_3^{outer} = x_{13} + x_{23} + x_{33} \\ w_{11} = w_{21} = w_{12} = w_{22} = w_{13} = w_{23} = w_{33} = 1\end{gathered} \tag{16.1}$$

Paso 2. Estimación *inner* de los coeficientes de regresión

En el paso 2 se estiman los coeficientes de regresión que unen LV; en el ejemplo de la figura 16.1, serían los coeficientes de regresión de $\xi_1 \rightarrow \xi_3$ y $\xi_2 \rightarrow \xi_3$. Hay diferentes formas de hacer esto. De acuerdo con el método del centroide (*centroid scheme*) (Wold, 1982) se usa el signo de la correlación entre las estimaciones de las LV, con el método de ponderación de factores (*factor weighting scheme*) (Lohmoller, 1989) se usa la correlación entre ellos, y con el método de la ponderación de *path* (*path weighting scheme*) se tiene en cuenta la dirección del *path*, es decir cuál es dependiente y cuál independiente. La ilustración la realizaremos mediante el método de ponderación de factores. Si dos LV son adyacentes (están unidas por un *path*) el coeficiente se calcula como la correlación entre las puntuaciones de las dos LV. Si no son adyacentes (no están unidos por ningún *path*, como es el caso de ξ_1 y ξ_2 en nuestro ejemplo, el coeficiente

Estadísticos e-Books & Papers

se fija a 0. Es decir, en el modelo ilustrado en la figura 16.1, las correlaciones (e_{13}, e_{23}) serían las estimaciones para los coeficientes de regresión β_{13}, β_{23}.

$$e_{jh} = \begin{cases} \text{cov}\left(\hat{\xi}_j^{outer}, \hat{\xi}_h^{outer}\right) & \xi_j, \xi_h \text{ adyacentes} \\ 0 & \text{resto de casos} \end{cases} \tag{16.2}$$

Paso 3. Estimación *inner* de las LV

En el paso 1, las variables latentes o factores se estimaban a partir de los indicadores, por eso denominábamos a esa estimación *outer*. En el paso 3 volvemos a estimar las variables latentes, pero ahora usando la parte estructural del modelo, es decir, las estimaciones de los coeficientes de regresión de los *paths* que hemos calculado en el paso 2, esto es:

$$\hat{\xi}_h^{inner} = \sum_j e_{jh} \hat{\xi}_j^{outer} \tag{16.3}$$

Que, aplicado a nuestro ejemplo resultaría en:

$$\begin{aligned} \hat{\xi}_1^{inner} &= e_{13} \hat{\xi}_3^{outer} \\ \hat{\xi}_2^{inner} &= e_{23} \hat{\xi}_3^{outer} \\ \hat{\xi}_3^{inner} &= e_{13} \hat{\xi}_1^{outer} + e_{23} \hat{\xi}_2^{outer} \end{aligned} \tag{16.4}$$

Paso 4. Estimación *outer* de los pesos

La estimación de los pesos en la etapa de inicialización había sido evidentemente *naíf*, porque solo pretendía ser un punto de partida en el proceso de optimización, y por eso se habían fijado arbitrariamente a 1. Llega el momento de que busquemos una estimación mejor, lo que implica que de alguna manera se ha de introducir un criterio de optimización, esto es, de maximización o minimización de algo. Lo que se hace con este fin es realizar una regresión por mínimos cuadrados ordinarios que tiene una estructura distinta en función de que la variable latente sea reflectiva o formativa (es decir, estimación de pesos que maximiza la varianza explicada de factores en los constructos formativos o indicadores en los reflectivos). En el caso de variables latentes reflectivas se realizan tantas regresiones con una variable explicativa como indicadores tenga el factor. Cada indicador es la variable dependiente y, el factor, la independiente. En el caso de un constructo formativo, se realiza una única regresión donde el factor es la variable dependiente y los indicadores son las independientes. Es decir:

$$x_{ij} = c_{ij} + w_{ij} \hat{\xi}_j^{inner} + \varepsilon_{ij} \ \text{(LV reflectiva)} \tag{16.5}$$

$$\hat{\xi}_j^{inner} = c_j + \sum_i w_{ij} x_{ij} + \varepsilon_j \ \text{(LV formativa)} \tag{16.6}$$

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Que aplicado al ejemplo de la figura 16.1:

$$\begin{aligned} x_{13} &= c_{13} + w_{13}\hat{\xi}_3^{inner} + \varepsilon_{13} \\ x_{23} &= c_{23} + w_{23}\hat{\xi}_3^{inner} + \varepsilon_{23} \\ x_{33} &= c_{33} + w_{33}\hat{\xi}_3^{inner} + \varepsilon_{33} \\ &\rightarrow \hat{w}_{13}, \hat{w}_{23}, \hat{w}_{33} \end{aligned} \tag{16.7}$$

$$\begin{aligned} \hat{\xi}_1^{inner} &= c_1 + w_{11}x_{11} + w_{21}x_{21} + \varepsilon_1 \\ \hat{\xi}_2^{inner} &= c_2 + w_{12}x_{12} + w_{22}x_{22} + \varepsilon_2 \\ &\rightarrow \hat{w}_{11}, \hat{w}_{21}, \hat{w}_{12}, \hat{w}_{22} \end{aligned} \tag{16.8}$$

Es importante avanzar que es de esta etapa del algoritmo de la que se derivan algunas de las grandes ventajas de PLS-SEM. Nótese que lo único que ha tenido que hacer el algoritmo para tratar un constructo formativo es una regresión, es decir, exactamente lo mismo que ha tenido que hacer si el constructo es reflectivo (con la diferencia de que es una regresión algo más compleja porque tiene más variables independientes, pero nada más). También de esta etapa del algoritmo se desprende por qué, como veremos, las necesidades de tamaño muestral cuando un modelo estructural se estima mediante PLS-SEM son menores: para realizar una regresión con dos variables explicativas, como muestra la expresión 16.8, no hace falta un tamaño muy superior a 30 o 40 casos, y esa es la regresión más compleja que se va a realizar para el ejemplo de la 16.1.

Paso 5. Estimación *outer* de los valores de las variables latentes

En este momento estamos de nuevo como en el paso 1, solo que en lugar de tener una estimación de los pesos igual a 1, tenemos una estimación de los pesos mejorada ($\hat{w}_{ij}$), fruto de maximizar mediante regresión la varianza explicada de indicadores (constructos reflectivos) o factores (constructos formativos). Simplemente volvemos a calcular una estimación *outer* de las variables latentes usando estos nuevos pesos:

$$\hat{\xi}_j^{outer} = \sum_i \hat{w}_{ij} x_{ij} \tag{16.9}$$

que en nuestro ejemplo daría lugar a:

Estadísticos e-Books & Papers

$$\begin{aligned}\hat{\xi}_1^{outer} &= \hat{w}_{11}x_{11} + \hat{w}_{21}x_{21} \\ \hat{\xi}_2^{outer} &= \hat{w}_{12}x_{12} + \hat{w}_{22}x_{22} \\ \hat{\xi}_3^{outer} &= \hat{w}_{13}x_{13} + \hat{w}_{23}x_{23} + \hat{w}_{33}x_{33}\end{aligned} \tag{16.10}$$

Paso 6. Criterio de parada

Y ahora los cinco pasos anteriores se irían repitiendo con resultados mejorados de la estimación de los pesos hasta que se alcanzara algún criterio de parada. Este criterio tiene que ver, lógicamente, con que en dos pasos sucesivos la estimación de los pesos que se obtienen apenas difieran. El valor de esa diferencia (criterio de parada) dependerá del *software* y, normalmente, podrá personalizarse. En el caso de la versión 3 de `SmartPLS` (Ringle *et al.*, 2015), ese valor por defecto es 10^{-7} y es, como señalamos, modificable. De una manera más formal el criterio quedaría:

$$\sum_{i,j} \left| w_{ij}^{(k)} - w_{ij}^{(k-1)} \right| < 10^{-7} \tag{16.11}$$

Paso 7. Solución final

Tras la última iteración, las cargas de los constructos formativos y reflectivos y los coeficientes de regresión de la parte estructural se calculan de la siguiente forma. Las puntuaciones factoriales (valores de las LV) son la última estimación *outer* que sale del algoritmo:

$$\hat{\xi}_j = \hat{\xi}_j^{outer} \tag{16.12}$$

La estimación de las cargas de los constructos formativos, que llamamos pesos (π), y las cargas de los constructos reflectivos (λ) se calculan con la fórmulas 16.7 y 16.8 que vimos en el paso 4, es decir:

$$x_{ij} = c_{ij} + \lambda_{ij}\hat{\xi}_j + \varepsilon_{ij} \ \text{(LV reflectiva)} \tag{16.13}$$

$$\hat{\xi}_j = c_j + \sum_i \pi_{ij}x_{ij} + \varepsilon_j \ \text{(LV formativa)} \tag{16.14}$$

Y para los coeficientes de regresión de la parte estructural del modelo, dado que ya están estimadas las variables latentes, solo es necesario realizar el conjunto de regresiones parciales en las que pueda descomponerse el modelo estructural (tantas como LV dependientes existan):

$$\hat{\xi}_h = \sum_j \beta_{jh}\hat{\xi}_j + \zeta_h \tag{16.15}$$

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

16.3. Cuándo usar PLS-SEM: fortalezas y debilidades

En este capítulo hemos dedicado la sección 16.2 a describir el algoritmo de PLS-SEM porque es imposible entender de dónde se derivan sus propiedades positivas y negativas, sus fortalezas y debilidades, si no se tiene, al menos, la intuición del funcionamiento de dicho algoritmo. Con lo descrito en la sección 16.2 en mente, es muy sencillo comprender sus fortalezas y debilidades.

16.3.1. Fortalezas

1. PLS-SEM impone pocas o **ninguna restricción respecto a la distribución que deben seguir los datos** (Fornell y Bookstein, 1982), relajando el supuesto de normalidad multivariable que es fundamental en la estimación por máxima verosimilitud de los CBSEM. Esta propiedad no parece aflorar de manera evidente del algoritmo. Si se realizan regresiones mediante mínimos cuadrados ordinarios (OLS), estas regresiones exigen normalidad. La explicación reside en que esa normalidad en la regresión es necesaria no para estimar el parámetro, sino para estimar su significatividad y lograr que el estadístico t siga una distribución conocida y podamos evaluarlo frente a sus valores críticos. PLS-SEM utiliza las estimaciones de los parámetros obtenidos por la regresión OLS pero, como veremos más adelante en el libro, utiliza procedimientos alternativos de remuestreo (*bootstrapping*) para evaluar la significatividad de esos parámetros, por esa razón la asunción de normalidad multivariante no es necesaria.

2. PLS-SEM funciona bien con **muestras relativamente reducidas**. Si se revisa el algoritmo, veremos que todo acaba descomponiéndose en regresiones OLS. En el caso de los constructos reflectivos, estas regresiones solo tienen una variable explicativa, y en el caso de los formativos, tantas como indicadores. Luego la regresión más compleja correspondería al constructo formativo con más indicadores o, revisando la parte estructural, a la regresión con la LV dependiente que estuviera relacionada con más LV independientes. Las reglas aproximadas clásicas (Hair *et al.*, 1998) hablan de 15-20 casos por cada variable explicativa, lo que haría que el modelo de nuestra figura 16.1 pudiera ser estimado con unos 40 casos. Ese mismo modelo, estimado mediante CBSEM requeriría de 200 casos o más (Boomsma, 1985). Si no se dispusiera de ellos, generaría estimadores sesgados (Hu y Bentler, 1995), soluciones inadmisibles (p. ej. casos Heywood) y problemas de identificación, especialmente en modelos complejos (Chin y Newsted, 1999). Una precisión es muy importante: este buen comportamiento con tamaños muestrales reducidos no libera al investigador de conseguir muestras que sean representativas, obviamente. Más adelante en el capítulo veremos criterios de evaluación del tamaño

Estadísticos e-Books & Papers

muestral mucho más precisos que las reglas aproximadas a que hacíamos referencia anteriormente y que tienen que ver con la potencia de la prueba.

3. Los **constructos formativos** se incorporan con mucha facilidad al algoritmo de PLS-SEM, de hecho, como hemos visto, se tratan como regresiones OLS igual que los reflectivos. Es cierto que los constructos formativos también pueden incorporarse a modelos estimados mediante CBSEM, pero las restricciones (modelos MIMIC, relación del constructo formativo con otros reflectivos) son muchas y no siempre la estimación es viable (puede revisarse Jarvis *et al.* 2003 o Diamantopoulos y Winklhofer, 2001 para detalles acerca de cómo incorporar constructos formativos en modelos CBSEM).

4. El **algoritmo de PLS-SEM es muy eficiente** (las regresiones son fáciles de implementar), no está sujeto a las restricciones de identificación del algoritmo de máxima verosimilitud (p. ej. identificación de escala) por muy complejo que sea el modelo (Hair *et al.*, 2011).

16.3.2. Debilidades

1. El algoritmo, a diferencia de los CBSEM, no estima los parámetros ajustando una matriz de varianzas y covarianzas teórica a una matriz de varianzas y covarianzas muestral, por lo tanto, **no va a contar con indicadores de ajuste global del modelo**[4], lo que limita, como señalan Hair *et al.* (2012b), su utilidad para comparar modelos alternativos competidores y le obliga, también, a recurrir a procedimientos no paramétricos (*blindfolding*) para evaluar la parte estructural del modelo, tal y como veremos en epígrafes posteriores.

2. Cada LV debe estar conectada, al menos, a otra LV mediante un *path* (relación estructural). Esto hace que **no puedan estimarse análisis factoriales confirmatorios (CFA)** mediante PLS-SEM. Recordemos que los CFA son necesarios en el enfoque de CBSEM para evaluar las propiedades psicométricas de los instrumentos de medida (fiabilidad, validez convergente, validez discriminante, etc.) de acuerdo con el enfoque clásico en dos pasos propuesto por Anderson y Gerbing (1988). Esto obligará a que la validación del instrumento de medida deba realizarse con

[4]Recordemos que, cuando se estima un CBSEM por máxima verosimilitud, la función de máxima verosimilitud se construye, básicamente, como la diferencia entre la matriz de varianzas y covarianzas teórica $\sum$ y la matriz de varianzas y covarianzas muestral $\mathbf{S}$, es decir, $F_{ML}(\mathbf{S};\sum) = tr\left(\mathbf{S};\sum^{-1}\right) + [\log|\sum| - \log|\mathbf{S}|] - q$, donde q es el número de indicadores. Pues bien, el estadístico que evalúa el ajuste del modelo y que se puede utilizar para comparar modelos anidados es la χ^2, que se basa directamente en esa función de máxima verosimilitud minimizada para obtener la estimación de los parámetros: $\chi^2 = (N-1)\,F_{ML}$, donde N es el tamaño muestral.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

un planteamiento alternativo al tradicional, que también abordaremos en este capítulo.

3. Todas las variables latentes deben tener, al menos, un indicador. Por definición, un **constructo de segundo orden** no tiene indicadores por cuanto son sus dimensiones de primer orden las que actúan como instrumento de medida del segundo, lo que es un problema, dado que en las áreas de marketing y dirección de empresas, la existencia de constructos de segundo orden es muy frecuente. Aunque no lo abordaremos en este manual puede consultarse Aldás (2016) para analizar la forma de resolver esta cuestión en un libro equivalente en la metodología al presente.

4. No se pueden estimar **modelos no recursivos** ($A \rightleftharpoons B$) ni tampoco cuando la parte estructural cuenta circularidades entre las LV: ($A \rightarrow B; B \rightarrow C; C \rightarrow A$)[5].

5. **La estimación de los parámetros realizada mediante PLS-SEM no es óptima en términos de sesgo y consistencia**. Como señalan Chin (1998) o Haenlein y Kaplan (2004), las puntuaciones factoriales de las LV son la agregación de indicadores que contienen un término de error y no convergerán con los valores reales salvo el caso en que el número de indicadores y el tamaño muestral tienda a infinito (McDonald, 1996). Este sesgo tiende a provocar estimaciones de las cargas y pesos más altas que las poblacionales y estimaciones más bajas en la parte estructural del modelo (coeficiente de regresión). En este momento hay una modificación del algoritmo de PLS-SEM propuesta por Dijkstra y Henseler (2015) e implementada en `SmartPLS 3.0`, que permite obtener estimaciones consistentes y, además, abre las puertas a eliminar otra de las limitaciones de PLS-SEM al proponer una medida de ajuste global que sería equivalente conceptualmente a la χ^2 de los CBSEM. Sin embargo solo puede aplicarse en modelos donde todos los constructos sean reflectivos.

16.3.3. Criterios de elección entre CBSEM y PLS-SEM

Como señalan Hair *et al.* (2011), la discusión entre CBSEM y PLS-SEM no debe plantearse en términos excluyentes. Hemos visto en el apartado anterior que muchos de los puntos fuertes de PLS-SEM coinciden con puntos débiles de CBSEM y viceversa y así lo vieron ya sus precursores (Joreskog y Wold, 1982). Estamos ante herramientas que son complementarias y es tarea del investigador decidir, dados los objetivos de su estudio, la estructura de su modelo y las características de sus datos, por uno u otro enfoque para la estimación. Por lo tanto, la pregunta ha de ser ¿cuándo deberíamos decantarnos por una u otra herramienta? Reinartz *et al.* (2009) revisan más de 30 trabajos en las áreas

[5]En el momento de redacción de este manual, hay una propuesta en fase preliminar de estimación consistente y asintóticamente normal para PLS que permitiría incorporar las relaciones no recursivas (Dijkstra y Henseler, 2015).

Estadísticos e-Books & Papers

de marketing y dirección de empresas indagando sobre el argumento que esgrimieron los autores para decantarse por PLS-SEM. También Joreskog y Wold (1982) plantean una propuesta para guiar la elección entre ambas herramientas. Podemos resumirlas del siguiente modo:

1. Usar PLS-SEM si el objeto de la investigación es relativamente nuevo y la teoría no está consolidada o los instrumentos de medida son muy preliminares (Chin, 1998; p. 333) si estamos en un enfoque exploratorio o ante una extensión de teoría estructural ya existente (Hair *et al.*, 2011; p. 144).

2. Usar PLS-SEM si el modelo es muy complejo, con gran número de indicadores y/o variables latentes (Chin, 1998; p. 333; Chin y Newsted 1999; p. 314; Hair *et al.*, 2011; p. 144).

3. Usar PLS-SEM si el modelo incorpora constructos formativos junto a los reflectivos (Chin, 1998; p. 333; Fornell y Cha, 1994; p. 73; Chin y Newsted 1999; p. 314; Hair *et al.*, 2011; p. 144).

4. Usar PLS-SEM si los datos no cumplen las condiciones de normalidad, independencia o tamaños muestrales mínimos para aplicar CBSEM (Chin, 1998; p. 333; Chin y Newsted 1999; p. 314; Hair *et al.*, 2011; p. 144).

5. Usar PLS-SEM si el objetivo principal del investigador es la predicción, no la estimación de los parámetros estructurales (Fornell y Cha, 1994; p. 73; Chin y Newsted 1999; p. 314; Hair *et al.*, 2011; p. 144).

6. Usar PLS-SEM si se van a utilizar las puntuaciones factoriales de las LV para análisis subsiguientes (Hair *et al.*, 2011; p. 144).

7. Usar CBSEM si el objetivo es someter teoría a confirmación o comparar teorías alternativas (Hair *et al.*, 2011; p. 144).

8. Usar CBSEM si el modelo tiene relaciones no recursivas (Hair *et al.*, 2011; p. 144).

16.4. Etapas en la estimación de un modelo estructural mediante PLS-SEM

Desde el conocido trabajo de Anderson y Gerbing (1988), se asume, como vimos para los modelos basados en covarianzas, que la estimación de un modelo estructural exige dos pasos: (1) validar el instrumento de medida y asegurarnos de que las variables latentes están correctamente aproximadas por sus indicadores y (2) estimar la parte estructural del modelo y derivar de ella las conclusiones que afectan al contraste de las hipótesis. En el capítulo 14 vimos que la primera fase requiere de la estimación de un análisis factorial confirmatorio (CFA) y la

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

aplicación de los criterios de fiabilidad, validez convergente y validez discriminante que detallamos en su momento. Pero al analizar el algoritmo de PLS-SEM hemos visto que la realización de un CFA no es posible en cuanto que implica variables latentes no relacionadas estructuralmente. Por este motivo el enfoque que se suele seguir es el siguiente (Hair *et al.*, 2014b):

1. Estimar el modelo completo, pero ignorar los resultados de la parte estructural y centrarse en la adecuación del instrumento de medida con criterios similares a los vistos en el capítulo 14.

2. Una vez depurado el instrumento de medida, volver a estimar el modelo y —ahora sí— analizar los resultados de la parte estructural con criterios también similares a los vistos en el capítulo 15.

3. En ambos casos, se requerirán criterios específicos para abordar cuestiones como los constructos formativos, la ausencia de normalidad o la inexistencia de indicadores de ajuste.

Los pasos que se han de seguir han sido descritos con detalle en el trabajo de Hair *et al.* (2014b) y se muestran a continuación. Para ilustrarlos adecuadamente los aplicaremos al caso 16.1.

1. Instrumento de medida. Constructos reflectivos.
   - *a*) Consistencia interna y fiabilidad: fiabilidad compuesta (CR) y alfa de Cronbach (CA) superiores a 0,70.
   - *b*) Validez convergente: las cargas deben ser significativas y superiores a 0,70. La varianza extraída promedio (AVE) debe ser superior a 0,50.
   - *c*) Validez discriminante:
     - 1) La AVE de cada variable latente debe ser superior al cuadrado de la correlación más grande que esa variable latente tenga con cualquier otra variable latente.
     - 2) La ratio HTMT no puede ser superior a 0,90.
2. Instrumento de medida. Constructos formativos.
   - *a*) Si un indicador tiene un peso significativo se mantiene. Si su peso y su carga son ambos no significativos, el indicador se suprime.
   - *b*) Multicolinealidad: el índice de inflación de la varianza (VIF) de cada indicador debe ser inferior a 5.
3. Modelo estructural.
   - *a*) R^2 de 0,5, 0,50 o 0,25 puede considerarse, respectivamente, como relevantes (*substantial*), moderados (*moderate*) o débiles (*weak*).
   - *b*) Utilícense los valores de los estadísticos t obtenidos mediante *bootstrapping* para establecer la significatividad de las relaciones estructurales.
   - *c*) Q^2 positivas obtenidas por *blindfolding* son indicadores de la relevancia predictiva del modelo.

Estadísticos e-Books & Papers

Caso 16.1. La dependencia del medio en Internet

Este caso está basado en el trabajo de Aldás *et al.* (2008). Partiendo de un modelo de aceptación de la tecnología clásico (Davis, 1989), incorpora una variable a la que se ha prestado poca atención en el campo del B2C, la dependencia del internauta del medio Internet, generándose un modelo integrador que se ilustra en la figura 16.2, con mejor capacidad explicativa que el TAM clásico.

En este modelo debemos prestar atención al constructo dependencia del medio, sobre todo para que se entienda su carácter formativo. El individuo tiene sus propios objetivos y algunos de ellos requieren el acceso a recursos de información que son controlados por los medios masivos (Ball-Rokeach 1985, 1989). La dependencia del individuo del medio se compone de tres dimensiones o categorías: entendimiento, orientación y entretenimiento (Ball-Rokeach 1985, 1989). Cada una de esas categorías se divide a su vez en dimensión personal y social, proporcionando seis niveles de relación de dependencia del medio.

El *entendimiento personal* es el proceso de interpretar las creencias, comportamientos y autoconcepto de uno mismo, mientras que el *entendimiento social* es la comprensión de otros individuos, culturas y acontecimientos en el mundo que nos rodea. La *orientación a la acción* se refiere a la necesidad de obtener una guía para desarrollar comportamientos personales específicos, como, por ejemplo, comprar, y la *orientación a la interacción* se refiere a la forma de cumplir los objetivos del individuo relativos a cómo comportarse y relacionarse correctamente con otras personas, incluyendo tanto aquellas con las que el consumidor ha desarrollado fuertes lazos afectivos como aquellas con las que no tiene ningún contacto previo. Adicionalmente, el *entretenimiento personal* se refiere al uso del medio para evadirse y liberarse de la tensión, mientras que el *entretenimiento social* es una vía importante a través de la cual se aprenden los roles sociales, normas y valores, utilizando el medio para entretenerse junto con otras personas (Grant *et al.*, 1991).

Pensemos en Internet. Una persona puede ser muy dependiente de Internet porque, todas las mañanas, necesita consultar la edición digital de los diarios antes de ir a su trabajo para saber qué ocurre en el mundo, en su país o en su ciudad (entendimiento social) pero puede no volver a utilizar Internet para nada más a lo largo del día. Otro individuo, por ejemplo, puede ser también muy dependiente del medio porque cuando llega a su casa ha de jugar una partida de videojuego con otros jugadores que están conectados por Internet (entretenimiento social), sin embargo no ha consultado los diarios digitales por la mañana. Ambas personas pueden ser igual de dependientes con combinaciones muy distintas de origen de esa dependencia (nada de entendimiento social y mucho de entretenimiento personal y viceversa). Démonos cuenta de que, si el constructo se relacionara reflectivamente con sus indicadores, esto no sería posible. Solo podría haber dependencia del medio si simultáneamente hubiera dependencia por motivos de entretenimiento personal, entretenimiento social, entendimiento personal, entendimiento social, orientación a la acción y orientación a la interacción. Como la variable latente causa sus indicadores, todas tienen que darse simultáneamente. Pero, dado que esto no es así conceptual-

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

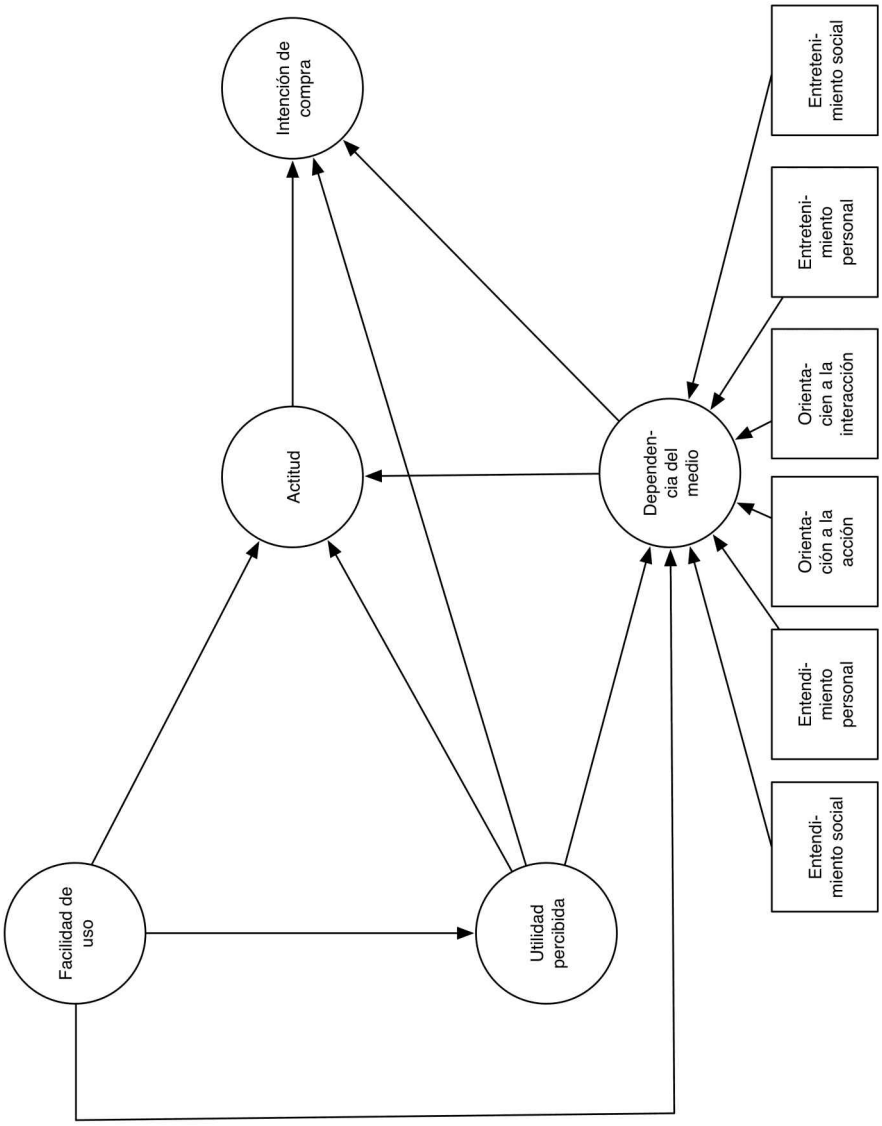


Figura 16.2.: Incorporación de la dependencia del medio a un modelo TAM

Estadísticos e-Books & Papers

mente (puede haber el mismo nivel de dependencia estando esta provocada mucho por una dimensión y nada por otra), el constructo ha de modelizarse de manera formativa, puesto que este planteamiento sí que permite esta red nomológica.

16.4.1. Validación del instrumento de medida

Al presentar el algoritmo ya señalamos que no era posible seguir el enfoque clásico de dos pasos propuesto por Anderson y Gerbing (1988) debido a que este enfoque se basa en la estimación de un análisis factorial confirmatorio como paso previo a la estimación del modelo estructural y en PLS-SEM no pueden estimarse CFA debido a que las variables latentes han de estar necesariamente unidas por una relación estructural (y las covarianzas del CFA no lo son).

La solución que se adopta es sencilla. Se estima el modelo completo (con la parte estructural, por tanto) mediante PLS-SEM, pero no se presta ninguna atención a la parte estructural (puesto que todavía no podemos confiar en el instrumento de medida). Solamente prestaremos atención a las cargas y a los pesos aplicando los criterios explicitados anteriormente. Tras depurar el instrumento de medida, prestaremos atención al modelo estructural. Revisemos los criterios.

A. Consistencia interna y fiabilidad (constructos reflectivos)

Para cada variable latente se calcula, como vimos en el capítulo 14:

- El α de Cronbach (1951) ha de ser superior al valor recomendado de 0,70 (Churchill, 1979), siendo:

$$\alpha = \frac{k\rho}{1 + (k-1)\,\rho} \tag{16.16}$$

  siendo k el número de indicadores de la LV, y ρ, la media de las correlaciones entre esos indicadores. Hair *et al.* (2012b), sin embargo, recomiendan no ofrecer en las estimaciones de PLS-SEM esta información como criterio de fiabilidad o consistencia interna y así lo haremos en este capítulo.

- La fiabilidad compuesta CR (Werts *et al.*, 1974) debería ser también superior a 0,70 (Fornell y Larcker, 1981).

$$CR_i = \frac{\left(\sum_i \lambda_{ij}\right)^2}{\left(\sum_i \lambda_{ij}\right)^2 + \sum_j var\left(\varepsilon_{ij}\right)} \tag{16.17}$$

  donde λ_{ij} es la estimación estandarizada de la carga del indicador j de la i-ésima LV. $Var\left(\varepsilon_{ij}\right)$ está relacionada con las cargas del siguiente modo:

$$Var\left(\varepsilon_{ij}\right) = 1 - \lambda_{ij}^2 \tag{16.18}$$

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

B. Validez convergente

- Aquí estamos exigiendo que la varianza que una variable latente explica de los indicadores que lo conforman sea, al menos, de la mitad, es decir, que sea superior a la varianza residual, a la asociada al término de error. Esto se consigue con una varianza extraída promedio (AVE) superior a 0,50 (Fornell y Larcker, 1981). Si vemos la expresión de la AVE (la notación es la misma de CR, siendo k el número de indicadores):

$$AVE_i = \frac{\sum_j \lambda_{ij}^2}{\sum_j \lambda_{ij}^2 + \sum_j var\left(\varepsilon_{ij}\right)} = \frac{\sum_j \lambda_{ij}^2}{k} \tag{16.19}$$

  dado que λ^2 es la varianza del indicador j explicada por la LV i, entonces la AVE es la media de las varianzas de los indicadores que explica el factor que están midiendo, y es a ese promedio al que exigimos que sea al menos la mitad del total.

- También exigimos que las cargas λ sean significativas (difícilmente un indicador j será una buena media de la LV i si no guarda ninguna relación con él, y si no es significativa es lo mismo que decir que no es estadísticamente distinta de 0) y que tengan un tamaño mínimo de 0,70. La lógica de 0,70 tiene que ver con la lógica de AVE, si una carga es 0,70, su cuadrado (la varianza explicada) es al menos del 50 % ($0{,}70^2 \simeq 0{,}50$).

C. Validez discriminante

De una manera sintética, dado que ya se trató el tema en el capítulo 14, dos conceptos son representados por dos LV distintas porque, aunque puedan estar relacionados (correlacionados), esta correlación no debería ser tan grande que nos haga dudar de que esos conceptos realmente sean distintos. En los modelos CBSEM vimos que la forma de contrastar, entre otras, la validez discriminante es evaluar que las correlaciones entre las LV que salen del CFA sean estadísticamente distintas de 1. Pero en PLS-SEM no podemos estimar un CFA. La alternativa por la que optamos es asegurarnos de que la AVE de cada par de factores es superior al cuadrado de la correlación entre ellos. La lógica es la siguiente, hemos dicho que la AVE es la varianza de los indicadores de una LV explicada por el factor que miden. El cuadrado de la correlación entre dos factores puede entenderse como la varianza de los indicadores de un factor que es explicada por el otro factor. Cuando exigimos:

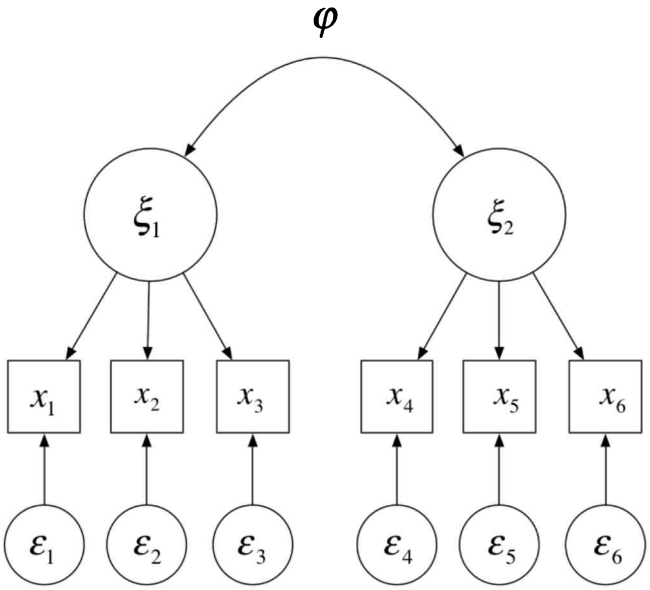
$$\begin{array}{l} AVE_i > \rho_{ij}^2 \\ AVE_j > \rho_{ij}^2 \end{array} \tag{16.20}$$

lo que estamos exigiendo es precisamente eso, que los indicadores de un factor estén más vinculados a ese factor que a otro para el que no se concibieron.

Recientemente, Henseler *et al.* (2014) han propuesto un procedimiento complementario para evaluar la validez discriminante que, en el fondo, está basado

Estadísticos e-Books & Papers

Figura 16.3.: Modelo de dos factores para ilustrar la el criterio de la ratio HTMT



Fuente: Henseler *et al.* (2014).

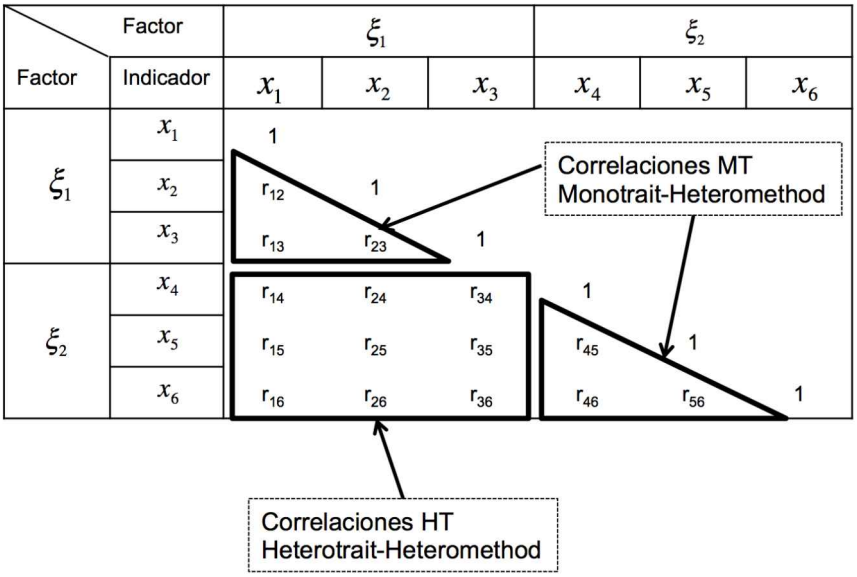
en la misma lógica. Si partimos de un ejemplo de dos factores, como el de la figura 16.3, en el que los factores, ξ_1 y ξ_2, son medidos a partir de tres indicadores cada uno de ellos, podemos construir la matriz de correlaciones entre esos indicadores, que está recogida en la figura 16.4. Esta matriz recoge dos tipos de correlaciones, por un lado, las correlaciones de los indicadores de un mismo factor, que Henseler *et al.* (2014) denominan *monotrait-heteromethod correlations*, cuya media denominaremos MT y, por otro lado, las correlaciones entre los ítems de un factor con los del otro factor, que se denominan *heterotrait-heteromethod* y cuya media denominaremos HT. De una manera simplificada, la ratio $HTMT$ sería la ratio HT/MT. Parece lógico que, si la media de las correlaciones entre los indicadores de dos constructos distintos es más grande que la que tienen los indicadores del constructo al que corresponden, entonces tendremos un problema de validez discriminante (en ese caso $HT/MT > 1$). Esto será poco habitual, lo normal es que las medias de las correlaciones entre los indicadores de los factores a los que van asociados sean más altas, la cuestión es cuán parecidas pueden ser a las asociadas a factores distintos. El criterio que se propone es el de Gold *et al.* (2001), quienes plantean que la ratio $HT/MT < 0{,}90$ para cada par de factores, de no ser así podemos estar ante un problema de validez discriminante.

Apliquemos ahora la teoría a la validación de la parte reflectiva del instrumento de medida del caso 16.1. Utilizaremos para ello el paquete `plspm`, cuya

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Figura 16.4.: Matriz de correlaciones para ilustrar la ratio HTMT



Fuente: Henseler *et al.* (2014).

única dificultad lógica está en traducir la parte estructural del modelo a una matriz de conexiones entre las variables latentes. Veamos primero la sintaxis paso a paso. Leemos en primer lugar la base de datos y cargamos las librerías necesarias:

```
library(haven)
datos <- read_sav(Datos_16_1_Caso.sav")
library(lavaan)
library(matrixpls)
library(plspm)
library(semPLS)
```

A continuación definimos la matriz que contiene las relaciones estructurales. En principio conviene ordenar los factores desde la parte izquierda del grafo (figura 16.2) hacia la parte derecha. En nuestro caso lo hemos hecho con el orden: 1. *facilidad de uso*, 2. *utilidad percibida*, 3. *dependencia*, 4. *actitud* y 5. *intención de compra*. Generamos para cada factor un vector que indica, siguiendo este orden, de qué factor recibe (1) o no recibe (0) flecha. Así vemos que el vector de *actitud* es $(1, 1, 1, 0, 0)$, dado que la actitud recibe flecha (es dependiente) del primer factor (facilidad de uso), del segundo (utilidad percibida) y del tercero (dependencia), pero no del cuarto (actitud) ni del quinto (intención de compra).

Estadísticos e-Books & Papers

```
facuso  =c(0,0,0,0,0)
utiper  =c(1,0,0,0,0)
dep     =c(1,1,0,0,0)
actitud =c(1,1,1,0,0)
intcomp =c(0,1,1,1,0)
```

A continuación juntamos todos los vectores en una matriz que vamos a llamar `modelo.path` y le damos como nombre de las filas y columnas los nombres de los factores.

```
# creamos la matriz fusionando filas
modelo.path = rbind(facuso,utiper,dep,actitud,intcomp)
# ponemos nombres a las columnas (optcional)
colnames(modelo.path) = rownames(modelo.path)
```

Llega el momento de decirle al programa qué indicadores de nuestra base de datos forman cada variable latente. El cuadro 16.1 muestra en la base de datos qué nombre tiene cada indicador y el orden que ocupa en la mencionada base. Generamos un vector (`modelo.blocks`) que define los "bloques", es decir, los indicadores que van juntos en un mismo factor, el orden en que hay que ponerlos es en el que se han definido en la matriz, por eso necesitamos el apoyo del cuadro 16.1 para asignar los indicadores:

```
modelo.blocks = list(7:12, 1:6, 24:29, 13:22, 23:23)
```

Ya solo resta para tener definido el modelo que le digamos al programa cuál es la modalidad ("modos") de medida de cada factor. En nuestro caso, todos son reflectivos (se indica como una A), salvo el tercer factor de la lista, que es la dependencia del medio que es formativo (se señala con una B):

```
modelo.modes = c("A","A","B","A","A")
```

Estamos en disposición de estimar nuestro modelo. La instrucción no necesita mucha aclaración, llamamos `modelo.pls` al objeto de `R` que contendrá la salida, `datos` es nuestro fichero de datos, `modelo.path` contiene la matriz con la parte estructural del modelo, `modelo.blocks`, la asignación de indicadores a variables latentes, `modes=modelo.modes`, el vector con el carácter formativo o reflectivo de cada "bloque" o variable latente, `scheme=path` indica el criterio seguido en el paso 2 del algoritmo para calcular factores a partir de indicadores. Las opciones `boot.val=TRUE` y `br=5000` corresponden al *bootstrapping* para estimar la significatividad de los parámetros que explicaremos posteriormente, `tol=1e-06` es el criterio de parada del algoritmo, y `maxiter=100`, el número máximo de iteraciones en que se permite la convergencia.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 16.1.: Estructura de la base de datos

Indicador	Orden en la base	Indicador	Orden en la base	Indicador	Orden en la base
utiper1	1	actitud1	12	dep1	24
utiper2	2	actitud2	14	dep2	25
utiper3	3	actitud3	15	dep3	26
utiper4	4	actitud4	16	dep4	27
utiper5	5	actitud5	17	dep5	28
utiper6	6	actitud6	18	dep6	29
facuso1	7	actitud7	19		
facuso2	8	actitud8	20		
facuso3	9	actitud9	21		
facuso4	10	actitud10	22		
facuso5	11	intcomp	23		
facuso6	12				

```
modelo.pls = plspm(datos, modelo.path, modelo.blocks,
modes= modelo.modes,scheme="path", boot.val=TRUE,
br=5000,tol = 1e-06, maxiter = 100)
summary(modelo.pls)
```

Comencemos analizando el tamaño de las cargas. Conviene siempre empezar por esta fase porque la mayoría de indicadores de fiabilidad y validez convergente están condicionados por las mismas. Eliminar una carga baja implica con toda seguridad la mejora de los mismos. Recordemos que nos estamos centrando en este momento en la parte reflectiva del instrumento, por lo que ignoraremos las cargas del constructo formativo *dependencia*. El cuadro 16.2 ofrece la salida. Observamos inmediatemente numerosas cargas por debajo de 0,7, sin embargo, es muy importante contener el impulso de la eliminación automática de las mismas. Recordemos que hemos asumido la validez de contenido del instrumento de medida porque son escalas que han sido validadas en trabajos previos. Eliminar muchos indicadores puede poner en peligro esa asunción. El criterio que se suele seguir es el siguiente (Hair *et al.*, 2014b):

1. Si la carga es inferior a $0,40$ se elimina.

2. Si está entre $0,40 - 0,70$, se observa si teníamos algún problema de fiabilidad y validez convergente en los indicadores CR o AVE. Si no los teníamos, se deja la carga. Si lo teníamos y al eliminarla el problema se soluciona, las eliminamos. Pero, si al eliminarlas, el problema persiste, se dejan en el modelo para no añadir el problema de validez de contenido al de fiabilidad y validez convergente.

Estadísticos e-Books & Papers

Cuadro 16.2.: Análisis del tamaño de las cargas

OUTER MODEL

	weight	loading	communality	redundancy
facuso				
1 facuso1	0.3255	0.819	0.6714	0.0000
1 facuso2	0.0721	0.462	0.2135	0.0000
1 facuso3	0.2492	0.690	0.4765	0.0000
1 facuso4	0.3295	0.745	0.5543	0.0000
1 facuso5	0.0499	0.395	0.1563	0.0000
1 facuso6	0.3538	0.743	0.5523	0.0000
utiper				
2 utiper1	0.1685	0.571	0.3264	0.0781
2 utiper2	0.2412	0.792	0.6267	0.1500
2 utiper3	0.2262	0.820	0.6723	0.1610
2 utiper4	0.2303	0.810	0.6555	0.1569
2 utiper5	0.2070	0.778	0.6046	0.1448
2 utiper6	0.2255	0.798	0.6366	0.1524
dep				
3 dep1	0.2728	0.519	0.2690	0.1095
3 dep2	0.0328	0.323	0.1046	0.0426
3 dep3	-0.1179	0.424	0.1795	0.0731
3 dep4	0.7410	0.929	0.8632	0.3514
3 dep5	0.1637	0.614	0.3767	0.1534
3 dep6	0.1842	0.591	0.3493	0.1422
actitud				
4 actitud1	0.1212	0.596	0.3548	0.1679
4 actitud2	0.1777	0.643	0.4130	0.1954
4 actitud3	0.0247	0.255	0.0649	0.0307
4 actitud4	0.1685	0.751	0.5645	0.2671
4 actitud5	0.1964	0.826	0.6829	0.3231
4 actitud6	0.2187	0.797	0.6347	0.3004
4 actitud7	0.1427	0.747	0.5584	0.2642
4 actitud8	0.1443	0.577	0.3330	0.1576
4 actitud9	0.1062	0.481	0.2313	0.1095
4 actitud10	0.1569	0.658	0.4332	0.2050
intcomp				
5 intcomp1	1.0000	1.000	1.0000	0.3314

© Ediciones Paraninfo

Estadísticos e-Books & Papers

Procedemos de esta manera prudente eliminando *facuso2* y *facuso5* y *actitud3* y *actitud9*. La forma de hacer operativa esta eliminación es localizar el orden que ocupan en la base de datos y no contemplarlos al definir los bloques del modelo.

```
modelo.blocks = list(c(7,9,10,12), 1:6,
24:29, c(13,14,16,17,18,19,20,22),23:23)
```

Vemos que hemos eliminado los indicadores que ocupaban el lugar 8 (*facuso2*), 11 (*facuso5*), 15 (*actitud3*) y 21 (*actitud9*). Con esta única modificación, reestimamos el modelo y vemos que las cargas ya son superiores a 0,7 (cuadro 16.3) y, cuando no lo son, no generan problemas en el alfa de Cronbach, la fiabilidad compuesta (CR `= DG.rho` en `plspm`) ni la varianza extraída promedio (AVE), como se aprecia en el cuadro 16.4. Ahora todos estos indicadores tienen valores superiores a los niveles mínimos exigibles que planteábamos con anterioridad.

En cuanto a la validez discriminante, el paquete `plspm` no calcula de manera directa la ratio HTMT, pero sí que da la información para aplicar el criterio de Fornell y Larcker (1981). El cuadro 16.5 nos ofrece la matriz de correlaciones entre los factores mientras que los AVE ya fueron calculados con anterioridad y aparecían en el cuadro 16.4. Por lo tanto, basta con comprobar que todas las correlaciones al cuadrado son inferiores al valor de los AVE de las variables latentes implicadas en la correlación o, como veremos que es práctica habitual en la presentación de los resultados, comprobar que las correlaciones son inferiores a la $\sqrt{AVE}$. Por ejemplo, centrándonos en la correlación más grande, la que corresponde a los factores utilidad percibida y actitud, esta toma el valor $\rho = 0,656$, que es un valor inferior a $\sqrt{AVE_{utiper}} = \sqrt{0,587} = 0,7662$ y también a $\sqrt{AVE_{actitud}} = \sqrt{0,502} = 0,7085$.

Como hemos señalado, el paquete `plspm` no ofrece la ratio HTMT, pero el paquete `matrixpls` sí que lo hace. La ventaja que tiene este paquete es que la definición del modelo, tanto la parte estructural como la de medida, no necesita de la construcción de las matrices que veíamos para `plspm` sino que define el modelo directamente mediante la sintaxis que utilizaría `lavaan` y que vimos en el capítulo 14. Al trabajar directamente con las matrices de covarianzas, este paquete es mucho más eficiente, aunque la presentación de los resultados en la salida es algo más confusa, razón por la que hemos presentado hasta ahora los resultados mediante el paquete `plspm`. A partir de este momento alternaremos las salidas cuando la presentación lo requiera. Para el caso que nos ocupa la sintaxis sería la siguiente:

```
library(lavaan)
library(matrixpls)
library(plspm)
library(semPLS)
#==================================
```


Estadísticos e-Books & Papers

Cuadro 16.3.: Análisis del tamaño de las cargas en la segunda estimación

```
-----------------------------------------------------------
OUTER MODEL
                    weight   loading   communality   redundancy
facuso
  1 facuso1         0.3426    0.816       0.6654       0.0000
  1 facuso3         0.2626    0.695       0.4832       0.0000
  1 facuso4         0.3445    0.749       0.5616       0.0000
  1 facuso6         0.3724    0.752       0.5649       0.0000
utiper
  2 utiper1         0.1678    0.571       0.3258       0.0844
  2 utiper2         0.2422    0.792       0.6276       0.1626
  2 utiper3         0.2266    0.820       0.6730       0.1744
  2 utiper4         0.2305    0.810       0.6554       0.1699
  2 utiper5         0.2072    0.778       0.6046       0.1567
  2 utiper6         0.2243    0.797       0.6357       0.1648
dep
  3 dep1            0.2799    0.524       0.2747       0.1119
  3 dep2            0.0243    0.315       0.0992       0.0404
  3 dep3           -0.1196    0.420       0.1766       0.0719
  3 dep4            0.7404    0.928       0.8609       0.3508
  3 dep5            0.1608    0.609       0.3711       0.1512
  3 dep6            0.1878    0.591       0.3490       0.1422
actitud
  4 actitud1        0.1271    0.589       0.3470       0.1639
  4 actitud2        0.1886    0.660       0.4357       0.2058
  4 actitud4        0.1776    0.762       0.5800       0.2739
  4 actitud5        0.2074    0.834       0.6951       0.3283
  4 actitud6        0.2307    0.796       0.6332       0.2990
  4 actitud7        0.1502    0.748       0.5596       0.2643
  4 actitud8        0.1519    0.565       0.3192       0.1507
  4 actitud10       0.1660    0.667       0.4444       0.2098
intcomp
  5 intcomp1        1.0000    1.000       1.0000       0.3355

-----------------------------------------------------------
```

Cuadro 16.4.: Análisis del tamaño de las cargas

```
-----------------------------------------------------------
BLOCKS UNIDIMENSIONALITY
         Mode  MVs  C.alpha  DG.rho  eig.1st  eig.2nd
facuso      A    4    0.748   0.841     2.28    0.668
utiper      A    6    0.856   0.894     3.52    0.780
dep         B    6    0.000   0.000     2.84    0.917
actitud     A    8    0.855   0.889     4.03    0.926
intcomp     A    1    1.000   1.000     1.00    0.000

-----------------------------------------------------------
-----------------------------------------------------------
SUMMARY INNER MODEL
               Type     R2  Block_Communality  Mean_Redundancy    AVE
facuso    Exogenous  0.000              0.569            0.000  0.569
utiper   Endogenous  0.259              0.587            0.152  0.587
dep      Endogenous  0.407              0.355            0.145  0.000
actitud  Endogenous  0.472              0.502            0.237  0.502
intcomp  Endogenous  0.336              1.000            0.336  1.000

-----------------------------------------------------------
```


© Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 16.5.: Correlación entre las variables latentes

```
-----------------------------------------------------------
CORRELATIONS BETWEEN LVs
          facuso  utiper    dep  actitud  intcomp
facuso     1.000   0.509  0.458    0.472    0.456
utiper     0.509   1.000  0.616    0.656    0.516
dep        0.458   0.616  1.000    0.529    0.412
actitud    0.472   0.656  0.529    1.000    0.530
intcomp    0.456   0.516  0.412    0.530    1.000

-----------------------------------------------------------
```

```
# Modelo mediante sintaxis Lavaan
#===================================
modelo.lavaan<-'

#Modelo de medida notese "<~" para indicar que
#el constructo es formativo
facuso=~facuso1+facuso3+facuso4+facuso6
utiper=~utiper1+utiper2+utiper3+utiper4+utiper5+utiper6
actitud=~actitud1+actitud2+actitud4+actitud5
        +actitud6+actitud7+actitud8+actitud10
intcomp=~intcomp1
dep<~dep1+dep2+dep3+dep4+dep5+dep6

#Parte estructural
intcomp~actitud+utiper+dep
actitud~dep+utiper+facuso
dep~facuso+utiper utiper~facuso
'
modelo.lavaan.out <- matrixpls(cov(datos),modelo.lavaan)
summary(modelo.lavaan.out)
```

El cuadro 16.6 ofrece la matriz con la ratio HTMT, donde se observa que ninguna ratio toma valores superiores al nivel crítico de 0,90 que señalábamos con anterioridad. Nótese que el constructo formativo dependencia del medio no aparece, en cuanto que no se le aplica el criterio ni tampoco la intención de uso medida con un único indicador y que, estrictamente por ello, no es una escala.

16.4.2. Determinación de la significatividad de los parámetros: *bootstrapping*

Hemos visto en el apartado anterior que para evaluar la validez convergente es necesario que las cargas factoriales sean significativas, lo que se comprueba viendo el valor del estadístico t. Pero también vimos que, si utilizáramos los

Estadísticos e-Books & Papers

Cuadro 16.6.: Matriz con los ratios HTMT

```
Heterotrait-monotrait matrix
           facuso    utiper   actitud
facuso  0.0000000
utiper  0.6290618 0.0000000
actitud 0.5638162 0.7567196 0.0000000
```

estadísticos t que se obtienen al estimar las regresiones al aplicar el algoritmo, sería necesario exigir normalidad multivariante.

PLS-SEM estima la significatividad de los parámetros con un planteamiento distinto: realiza un remuestreo basado en *bootstrapping*. Veamos el procedimiento. Se seleccionan N submuestras de manera aleatoria que tienen el mismo tamaño de la muestra original (esto es posible porque el muestreo se realiza con reemplazo). El número de submuestras, siguiendo a Hair *et al.* (2012b) ha de ser de al menos 5.000 submuestras a no ser que el tamaño muestral original sea superior a ese número en cuyo caso también el número de submuestras será superior.

El modelo se estima ahora para esas N submuestras, con lo que se obtienen N estimaciones de cada uno de los parámetros del modelo, ya sean coeficientes de regresión, cargas o pesos. Para cada parámetro estimado N veces podremos calcular, por tanto, la media de las estimaciones y su error típico. De acuerdo con Chin (1998), el estadístico t obtenido por *bootstrapping* que sirve para contrastar la hipótesis nula de que el parámetro (β, λ o w) es 0 se puede calcular de la siguiente forma:

$$t = \frac{\hat{\beta}}{SE\,(\beta)} \tag{16.21}$$

donde el estadístico t se distribuye como una t de Student con $N-1$ grados de libertad, $\hat{\beta}$ es la estimación del coeficiente de regresión (o la carga o el peso) obtenido de la muestra original y $SE\,(\beta)$ es el error estándar de las N estimaciones de ese mismo parámetro. De esta manera se valora la significatividad de todos los parámetros estimados en el modelo.

Henseler *et al.* (2009) señalan una importante consideración técnica que es necesario tener en cuenta a la hora de aplicar este procedimiento. Dependiendo del procedimiento de cálculo inicial de los pesos, el signo de la estimación de la variable latente puede cambiar (aunque sea la misma en valor absoluto). Como el remuestreo implica muchas estimaciones del mismo modelo, este cambio de signo puede acercar la media de las estimaciones a cero, sesgando el error estándar al alza y disminuyendo las probabilidades de rechazo de la hipótesis nula (parámetro no significativo). Por ese motivo, como también lo hacen Hair *et al.* (2012b), estos autores recomienda utilizar la opción de cambio individual del signo durante el proceso de remuestreo.

En un trabajo posterior, Hair *et al.* (2014b) plantean una variación sobre

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

la recomendación del cambio de signo en el procedimiento de *bootstrapping*. Su recomendación es comenzar por la opción más conservadora (la que genera valores t más bajos, es decir, permitir cambios de signo). Si el coeficiente analizado es significativo, lo damos como resultado final. Si no lo es, entonces recomiendan pasar a la opción de, cuando el signo cambia, volver a cambiarlo para hacerlo coherente con la estimación original a nivel individual porque es la que genera valores t más altos, si sigue entonces sin ser significativo, hay que darlo como no significativo y en caso de ser significativo recomiendan pasar a la opción de cambio de signo a nivel de constructo como criterio de compromiso para decidir.

La solicitud en el paquete `plspm` de la realización del *bootstrapping* es sencilla y la vimos con anterioridad, aunque no la comentamos, mediante `boot.val=TRUE` le señalamos que realice el *bootstrapping* y mediante `br=5000` que escoja 5.000 submuestras que siempre serán del mismo tamaño de la submuestra original.

```
modelo.pls = plspm(datos, modelo.path, modelo.blocks,
modes= modelo.modes,scheme="path", boot.val=TRUE,
br=5000,tol = 1e-06, maxiter = 100)
summary(modelo.pls)
```

Pues bien, el cuadro 16.7 nos da los resultados del *bootstrapping* para evaluar la significatividad de las cargas factoriales. El paquete `plspm` no ofrece calculados los estadísticos t ni la significatividad —aunque su cálculo es elemental con la información que proporciona mediante la ecuación 16.21— sino que opta por ofrecer los intervalos de confianza. La carga será significativa cuando el intervalo no contenga al cero, como es el caso para todas las cargas de nuestro ejemplo.

16.4.3. Validez y fiabilidad del instrumento de medida (constructos formativos)

Cuando el constructo es formativo, los criterios de validez y fiabilidad planteados con anterioridad no son aplicables. Lo que en un constructo reflectivo es deseable —la alta correlación entre los indicadores, que muestra que están reflejando adecuadamente una causa común, la variable latente que quieren medir— se convierte en un problema en cuanto que, como vimos al explicar el algoritmo, este va a realizar una regresión con tantas variables explicativas como indicadores formativos compongan la variable latente, pudiendo provocar problemas de multicolinealidad que distorsionen la interpretación. Si el constructo es realmente formativo, esta correlación debería ser baja, de ser elevada, estaríamos probablemente ante un constructo de carácter reflectivo. Lo contrario aplica en la evaluación de los constructos reflectivos, si sus cargas son sistemáticamente bajas, lo más probable es que nos encontremos ante una incorrecta modelización reflectiva del mismo. Puede consultarse Aldás (2014)

Estadísticos e-Books & Papers

Cuadro 16.7.: Significatividad de las cargas factoriales obtenida mediante *boostrapping*

```
-----------------------------------------------------------
BOOTSTRAP VALIDATION
loadings
                    Original  Mean.Boot  Std.Error  perc.025  perc.975
facuso-facuso1         0.816      0.816   1.96e-02     0.775     0.852
facuso-facuso3         0.695      0.694   3.87e-02     0.608     0.757
facuso-facuso4         0.749      0.751   2.58e-02     0.696     0.794
facuso-facuso6         0.752      0.752   2.60e-02     0.698     0.801
utiper-utiper1         0.571      0.567   4.47e-02     0.479     0.647
utiper-utiper2         0.792      0.790   2.01e-02     0.749     0.822
utiper-utiper3         0.820      0.819   1.77e-02     0.784     0.848
utiper-utiper4         0.810      0.811   1.62e-02     0.776     0.841
utiper-utiper5         0.778      0.776   2.27e-02     0.732     0.824
utiper-utiper6         0.797      0.801   1.94e-02     0.762     0.834
dep-dep1               0.524      0.525   6.79e-02     0.378     0.664
dep-dep2               0.315      0.309   7.70e-02     0.141     0.448
dep-dep3               0.420      0.414   7.08e-02     0.276     0.549
dep-dep4               0.928      0.914   2.63e-02     0.853     0.957
dep-dep5               0.609      0.609   6.22e-02     0.490     0.717
dep-dep6               0.591      0.583   5.81e-02     0.476     0.679
actitud-actitud1       0.589      0.587   4.19e-02     0.494     0.670
actitud-actitud2       0.660      0.658   3.69e-02     0.590     0.719
actitud-actitud4       0.762      0.759   2.69e-02     0.706     0.808
actitud-actitud5       0.834      0.833   1.54e-02     0.800     0.859
actitud-actitud6       0.796      0.794   1.97e-02     0.753     0.829
actitud-actitud7       0.748      0.743   2.60e-02     0.682     0.788
actitud-actitud8       0.565      0.565   3.66e-02     0.488     0.626
actitud-actitud10      0.667      0.671   3.86e-02     0.597     0.740
intcomp-intcomp1       1.000      1.000   6.49e-17     1.000     1.000
```

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

para una discusión más profunda sobre el carácter formativo o reflectivo de los constructos y técnicas para objetivar la determinación de su carácter.

Como plantean Hair *et al.* (2014b), el primer paso es evaluar el **nivel de colinealidad** que se da entre los indicadores formativos del constructo —en nuestro caso la dependencia del medio— y ver si este alcanza niveles problemáticos. El segundo paso, que abordaremos después, será evaluar sus pesos para establecer la significatividad y relevancia de los indicadores formativos del constructo evaluado.

El procedimiento para *evaluar los problemas de multicolinealidad* es el mismo que se sigue en una regresión. Supongamos que $X_1 \ldots X_n$ son los indicadores formativos de una determinada variable latente. Para evaluar el grado de multicolinealidad se regresa cada uno de los indicadores sobre el resto. El nivel de relación vendrá determinado por la R^2 de cada una de esas regresiones:

$$\begin{cases} X_1 = c_1 + b_{12}X_2 + b_{13}X_3 + \ldots b_{1n}X_n \rightarrow R_1^2 \\ X_2 = c_2 + b_{21}X_1 + b_{23}X_3 + \ldots b_{2n}X_n \rightarrow R_2^2 \\ \vdots \\ X_n = c_n + b_{n1}X_1 + b_{n2}X_2 + \ldots b_{nn-1}X_{n-1} \rightarrow R_n^2 \end{cases} \tag{16.22}$$

Si la R_1^2 fuera muy elevada (imaginemos $R_1^2 = 1$ como caso extremo) querría decir que X_1 podría ponerse como combinación lineal del resto de indicadores y estaríamos ante una situación de multicolinealidad perfecta. Obviamente estamos ante una cuestión de grado, es decir, a partir de qué valor de la R_1^2 la correlación es excesiva. Distintos trabajos ofrecen distintos niveles de corte. Hair *et al.* (2011) sugieren que estamos ante un problema de multicolinealidad cuando $R^2 \geq 0{,}80$. Desafortunadamente los programas estadísticos no ofrecen este valor, sino una transformación de los mismos, la tolerancia (TOL) y el índice de inflación de la varianza (VIF) se relacionan con la R^2 en sus niveles problemáticos de la siguiente forma:

$$TOL = 1 - R^2 \rightarrow R^2 \geq 0{,}80 \rightarrow TOL < 0{,}20 \tag{16.23}$$

$$VIF = \frac{1}{TOL} \rightarrow R^2 \geq 0{,}80 \rightarrow TOL < 0{,}20 \rightarrow VIF \geq \frac{1}{0{,}20} = 5 \tag{16.24}$$

es decir, que valores de la tolerancia inferiores a 0,20 o valores del índice de inflación de la varianza superiores a 5 serían indicativos de problema de colinealidad no asumible entre los indicadores del constructo formativo. Aunque parezca absurdo utilizar indicadores que, en el fondo, están midiendo lo mismo, no lo es. Cada uno tiene su propia interpretación, así la raíz cuadrada del índice de inflación de la varianza ($\sqrt{VIF}$) nos informa del grado en que el error estándar se ha incrementado debido a problemas de colinealidad. Si $VIF = 4$, entonces estaríamos diciendo que el error estándar se ha doblado por esa razón. La solución ante estas situaciones puede pasar por eliminar el indicador fuertemente correlacionado con otro (siempre que no afecte a la validez de contenido, lo que exige que los indicadores restantes incorporen el contenido del ítem, lo

Estadísticos e-Books & Papers

Cuadro 16.8.: Diagnóstico de colinealidad en los indicadores del constructo formativo. VIF.

```
    dep1     dep2     dep3     dep4     dep5     dep6
1.157629 1.349192 1.645306 1.396391 1.878545 1.604158
```

que no es muy difícil dado el grado de correlación que existía).

Los dos paquetes que estamos utilizando, `plspm` y `matrixpls`, no proporcionan de manera directa los indicadores de multicolinealidad, pero se trata de procedimientos estándar en el diagnóstico de las regresiones que pueden obtenerse mediante otros paquetes estándar como `car`. Si se observa la sintaxis hemos realizado una regresión donde puede utilizarse cualquier variable como dependiente (en nuestro caso *facuso1*) en la medida en que el test se aplica a las independientes que, eso sí, han de ser los indicadores del construto formativo (*dep1...dep6*). Los resultados del cuadro 16.8 no evidencian problemas de multicolinealidad.

```
library(car)
colinealidad <-
lm(facuso1~dep1+dep2+dep3+dep4+dep5, data=datos)
vif(colinealidad)
```

Si recordamos la etapa del algoritmo en el que se estimaban los pesos del constructo formativo (capítulo 2, paso 4), veíamos que las puntuaciones factoriales actuaban como variable dependiente y los indicadores como independientes, luego los coeficientes de regresión (los pesos) nos indican la contribución relativa de cada indicador al constructo, su importancia relativa. Lógicamente hay que exigir que estos pesos sean significativos (distintos estadísticamente de cero) lo que se logra averiguar, igual que con las cargas, mediante *bootstrapping*.

Hair *et al.* (2014b) hacen una importante precisión a la hora de determinar la relevancia relativa de cada indicador en un constructo formativo. Por construcción el valor máximo que un peso puede tomar en un constructo formativo viene condicionado por el número de indicadores del constructo en una relación $1/\sqrt{n}$, donde n es el número de indicadores. En el caso de nuestro ejemplo, con 6 dimensiones de segundo orden, el valor máximo que podrá tomar el peso es $1/\sqrt{6} = 0{,}41$. Por lo tanto si el tamaño máximo del peso decae al aumentar el número de indicadores, la probabilidad de que este no sea significativo también aumenta. En consecuencia Hair *et al.* (2014b) señalan que un peso no significativo no debería ser considerado necesariamente como un indicador de que el instrumento de medida no tiene la calidad suficiente. Abogan por no considerar solo la *contribución relativa* del indicador al constructo formativo (el peso), sino también la *contribución absoluta* que el indicador tendría si no se considerara ningún otro indicador (la carga factorial o *outer loading* también estimada en el algoritmo). Estos autores plantean la siguiente **secuencia para**

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

evaluar el instrumento de medida formativo:

1. Si el peso es significativo, se retiene y se interpreta la contribución relativa del indicador al constructo como un coeficiente de regresión estándar.

2. Cuando el peso de un indicador no sea significativo, pero su carga sea grande ($> 0{,}50$), se mantiene el indicador aunque no sea significativo el peso.

3. Cuando el peso de un indicador no es significativo, su carga es pequeña ($< 0{,}50$) y no significativa, se ha de eliminar el indicador formativo.

4. Cuando el peso de un indicador no es significativo, su carga es pequeña ($< 0{,}50$) pero es significativa, la decisión queda a criterio del investigador, que debe evaluar la relevancia conceptual y el posible solapamiento con otros indicadores. Si ese solapamiento conceptual existe, puede eliminarse, pues su contenido puede quedar recogido por otros indicadores, pero, si no es así, habría que mantenerlo y concluir que, en esta aplicación, ese indicador no juega un papel relevante en la configuración de la variable formativa[6].

Dado que en el proceso, como vimos en la subsección anterior, ya hemos realizado el *bootstrapping*, solo nos queda aplicar el procedimiento a los resultados del constructo formativo. Seleccionamos en el cuadro la información relevante de la salida. Vemos que hay dos pesos no significativos —su intervalo de confianza contiene al 0— los correspondientes a *dep2* y *dep3*. Si nos fijamos en sus respectivas cargas, estas son inferiores a 0,5 (0,315 para *dep2* y 0,420 para *dep3*) pero en ambos casos significativas. Nos encontramos por tanto en la situación 3, aquella en la que el investigador ha de decidir.

Y la decisión exige tener un conocimiento claro de qué es un constructo formativo. Estos constructos exigen que su red de indicadores sea exhaustiva. En la medida en que los indicadores definen al constructo, la ausencia de un indicador relevante puede afectar a la definición. Otra cuestión distinta es que, como vimos, en la aplicación a un caso concreto unos indicadores puedan ser relevantes —el caso del riesgo económico en el uso de la banca electrónica— y otros no serlo —el riesgo social en ese mismo caso, por ejemplo—. Por eso la

[6]Esta decisión puede chocar si mantenemos la lógica reflectiva, porque si la variable latente causa a los indicadores, todos deben estar correlacionados y darse simultáneamente, pero no así en la formativa. Imaginemos el concepto de riesgo percibido en Internet, que tiene dimensiones como el riesgo social (qué pensará mi entorno de que use Internet para un fin determinado) o el riesgo de tiempo (usar Internet para un fin dado puede suponer un consumo de tiempo excesivo si no estoy familiarizado con la tecnología). Pues bien, el riesgo de tiempo puede ser relevante en la configuración del riesgo si estamos analizando el uso de la banca electrónica en un colectivo de edad elevada y poco familiarizado con la tecnología y no ser así si estamos analizando un colectivo de edades intermedias familiarizados con Internet. El concepto es el mismo pero la significatividad de las dimensiones puede variar según la aplicación (lo que no tiene sentido en un constructo reflectivo).

Estadísticos e-Books & Papers

Cuadro 16.9.: Decisión sobre el mantenimiento o no de los indicadores no significativos del constructo formativo

```
---------------------------------------------------------
BOOTSTRAP VALIDATION
weights
                    Original  Mean.Boot  Std.Error  perc.025  perc.975
dep-dep1              0.2799     0.2886   7.21e-02    0.1615    0.4300
dep-dep2              0.0243     0.0191   6.45e-02   -0.1199    0.1551
dep-dep3             -0.1196    -0.1184   8.16e-02   -0.2641    0.0406
dep-dep4              0.7404     0.7225   6.23e-02    0.5714    0.8277
dep-dep5              0.1608     0.1749   8.57e-02    0.0116    0.3591
dep-dep6              0.1878     0.1801   8.18e-02    0.0287    0.3375

---------------------------------------------------------
BOOTSTRAP VALIDATION
loadings
                    Original  Mean.Boot  Std.Error  perc.025  perc.975
dep-dep1               0.524      0.525   6.79e-02     0.378     0.664
dep-dep2               0.315      0.309   7.70e-02     0.141     0.448
dep-dep3               0.420      0.414   7.08e-02     0.276     0.549
dep-dep4               0.928      0.914   2.63e-02     0.853     0.957
dep-dep5               0.609      0.609   6.22e-02     0.490     0.717
dep-dep6               0.591      0.583   5.81e-02     0.476     0.679
```

tendencia debe ser, en mi opinión, mantenerlos para que se vea, precisamente, que en la aplicación concreta del constructo esos indicadores no se activan. Por lo tanto, los dejaremos en el caso que estamos resolviendo.

16.4.4. Evaluación del modelo estructural

Como ya señalábamos al principio de este capítulo, la lógica de los modelos estimados por PLS-SEM es totalmente distinta a los CBSEM. Al no estimarse los parámetros otorgándoles aquellos valores que minimizan las diferencias entre la matriz de varianzas y covarianzas teórica y la muestral, no pueden derivarse estadísticos que evalúen la hipótesis nula de que esas matrices no son estadísticamente equivalentes (χ^2 en los CBSEM) ni todos los indicadores de ajuste que se derivan de ese estadístico (GFI, AGFI, TLI, RMSEA, etc.) y que ya presentamos en el capítulo 14.

Esta situación nos obliga a buscar algún tipo de indicador que nos diga si nuestro modelo teórico aporta o no algún tipo de valor en la explicación de la realidad, es decir, en qué medida los datos son compatibles con nuestro modelo. El enfoque que se sigue en PLS-SEM es intentar evaluar la capacidad de predecir los valores de las variables latentes dependientes que tiene el modelo. Para ello se evalúan tres aspectos (1) el valor de las R^2 de las LV dependientes, (2) la relevancia predictiva Q^2 del modelo y, en el caso de alcanzar valores satisfactorios, se pasa a (3) evaluar la significatividad de las relaciones estructurales (contraste de las hipótesis del modelo).

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

A. R^2 de las variables latentes dependientes en el modelo

La R^2 de una variable latente dependiente nos indica qué parte de la varianza de dicha variable es explicada por el conjunto de variables latentes que influyen sobre ella. Este valor está acotado entre 0 y 1 y podemos encontrarnos con una gran diversidad de reglas aproximadas para señalar qué es un valor razonable para el mismo. Así Chin (1998, p. 323) indica que un R^2 de 0,67 sería un indicador de que una parte relevante (*substantial*) de la LV está siendo explicada por la red conceptual de nuestro modelo, un valor de 0,33 indicaría una explicación moderada (*moderate*) y 0,19 o inferiores la explicación sería débil (*weak*). Otros autores como Hair *et al.* (2011), Hair *et al.* (2014b) o Henseler *et al.* (2009) dan los puntos de corte de 0,75, 0,50 o 0,25 para los mismos tres niveles (relevante, moderada y débil).

En nuestra opinión, y aunque como se comprueba en el cuadro 16.4 todos los R^2 muestran relaciones que podrían considerarse moderadas o fuertes, salvo quizás el de la facilidad de uso, estos criterios, siendo sin duda estándares razonables, no dejan de ser criterios arbitrarios. El problema habría que enfocarlo bajo la perspectiva de que nuestro tamaño muestral y la complejidad de la regresión con mayor número de variables explicativas sean compatibles con una potencia de prueba superior al 80 % (Cohen, 1988), es decir, que cuando rechazamos la hipótesis nula de que la R^2 se desvía significativamente de cero, tengamos una seguridad del 80 % de que es así porque esa hipótesis nula es falsa. La función `pwr.f2.test{pwr}` nos permite calcular, bien la potencia conseguida dado nuestro tamaño muestral o bien el tamaño muestral necesario para conseguir esa potencia del 80 %. Lo único que tiene que hacer el investigador es determinar la regresión más compleja que realiza el algoritmo, que siempre es la que tiene más número de variables independientes entre las siguientes (a) constructo formativo con mayor número de indicadores, en nuestro caso la dependencia con 6 o (b) la variable latente dependiente que recibe mayor número de relaciones estructurales (en nuestro ejemplo tanto actitud como intención de uso con 3). Por lo tanto, la regresión más compleja es la (a), con 6 variables independientes. La función tiene la estructura:

```
pwr.f2.test(u=,v=,f2=,sig.level=,power=)
```

y se estima siempre el parámetro que se declara como `NULL`. Los parámetros se corresponde al número de variables independientes $u = 6$; v es una función del número de regresores u y el tamaño muestral: $v = N - u - 1$, donde N es el tamaño muestral ($N = 464$), por lo tanto, $v = 464-6-1 = 457$; f^2 es el tamaño del efecto que se quiere ser capaz de detectar que normalmente es moderado y se introduce con el valor fijo $f^2 = 0,15$ (si se quisiera ser capaz de detectar un efecto elevado, se introduciría $f^2 = 0,35$; `sig.level` es la significatividad que también suele fijarse para $\alpha = 5\,\%$, `power` es la potencia de la prueba que la función calcula a partir de los parámetros anteriores y por lo tanto hay que indicar como `NULL`. Si quisiéramos saber qué muestra necesitaríamos para

Estadísticos e-Books & Papers

Cuadro 16.10.: Cálculo de la potencia lograda en la estimación y del tamaño que hubiera sido necesario para $1 - \beta = 0{,}80$

```
> pwr.f2.test(u =6, v = 457 , f2 =0.15 , sig.level = 0.05, power = NULL)

     Multiple regression power calculation

              u = 6
              v = 457
             f2 = 0.15
      sig.level = 0.05
          power = 0.9999999

> pwr.f2.test(u =6, v = NULL , f2 =0.15 , sig.level = 0.05, power = 0.80)

     Multiple regression power calculation

              u = 6
              v = 90.30998
             f2 = 0.15
      sig.level = 0.05
          power = 0.8
```

una potencia del 80 % introduciríamos `power=0.80` y dejaríamos que la función determinara `v=NULL`. Veámoslo.

```
# Potencia conseguida en nuestra estimacion
pwr.f2.test(u=6,v=457,f2=0.125,sig.level=0.05,power=NULL)

# Muestra necesaria para una potencia dle 80%
pwr.f2.test(u=6,v=NULL,f2=0.125,sig.level=0.05,power=0.80)
```

El cuadro 16.10 ofrece la salida de ambas funciones y muestra como, con nuestro tamaño muestral, la potencia es muy superior al 80 %, pero también pone de manifiesto la propiedad de PLS-SEM de no requerimiento de muestras elevadas en la medida en que bastaría una muestra —asumiendo que fuera representativa— de $N = 97$ casos para conseguir esa potencia del 80 %. Esta muestra se deriva del resultado como sigue:

$$v = N - u - 1$$

$$90,3 = N - 6 - 1 \rightarrow N \simeq 97$$

B. Relevancia predictiva Q^2 del modelo

Para intentar evaluar de una manera más objetiva la capacidad de realizar predicciones razonables de las variables latentes dependientes que la que proporciona el criterio bastante arbitrario de la R^2, se propone una segunda prueba

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

basada en un procedimiento de remuestreo denominado *blindfolding*. El procedimiento es muy exigente y su lógica bastante clara. El algoritmo habrá estimado las puntuaciones factoriales (valores) de todas las variables latentes dependientes del modelo, luego conoceremos sus valores. A continuación lo que se hace es omitir (borrar) uno de cada d datos de ese modelo, es decir, generamos valores perdidos artificialmente. Al valor d lo denominamos *distancia de omisión*. Esos valores perdidos pueden aproximarse de dos formas alternativas: utilizando el modelo estructural estimado (sus coeficientes de regresión) o sin utilizar el modelo estructural (por ejemplo utilizando la media de la LV para aquellos casos donde el valor no se ha omitido). En buena lógica nuestro modelo tendrá una capacidad de predecir razonable si al menos es capaz de cometer menos errores en la estimación de los valores perdidos que la estimación *naïve* de la media, que no utiliza el modelo para nada. No olvidemos que los valores obtenidos mediante la aplicación de la parte estructural del modelo no coincidirán con los originales porque el algoritmo los estimaba a partir del instrumento de medida.

De una manera algo más formal, el proceso sigue los siguientes pasos (puede encontrarse una descripción más detallada en Tenenhaus *et al.*, 2005):

1. La matriz de datos (puntuaciones de las variables latentes dependientes) se divide en d grupos (normalmente suelen ser 5 o 7, pero el único requerimento es que el número de grupos no sea un divisor exacto del tamaño muestral por las razones que luego veremos). Wold (1982) recomienda valores entre 5 y 10 grupos. A este valor d se le denomina *distancia de omisión*.

2. Se elimina el primer grupo de la base de datos, generándose así valores perdidos artificialmente. Si, por ejemplo, $d = 7$ se elimina uno de cada 7 casos.

3. Se estima el modelo mediante PLS-SEM sin esos valores perdidos. Como explicamos al introducir el algoritmo, el resultado de la aplicación será una estimación de los pesos, cargas, coeficientes de regresión y puntuaciones factoriales de las variables latentes. Supongamos que el modelo de la figura 16.5 se ha estimado mediante PLS-SEM y se han obtenido los coeficientes de regresión que allí se ilustran.

4. Mediante las puntuaciones factoriales y los coeficientes de regresión obtenidos, se estiman los valores faltantes de cada variable latente dependiente. Tendremos, para cada valor perdido, el verdadero valor antes de eliminarlo (y) y la estimación resultante de utilizar la parte estructural del modelo ($\hat{y}$), obtenido aplicando la expresión que se mostraba en el paso 7 del algoritmo:
$$\hat{\xi}_h = \sum_j \beta_{jh}\hat{\xi}_j + \zeta_h \tag{16.25}$$
En el ejemplo de la figura 16.5, el valor real que se omitió en la observación número 1 para el factor dependiente ξ_3 fue $\xi_3 = 9$, y la estimación que se

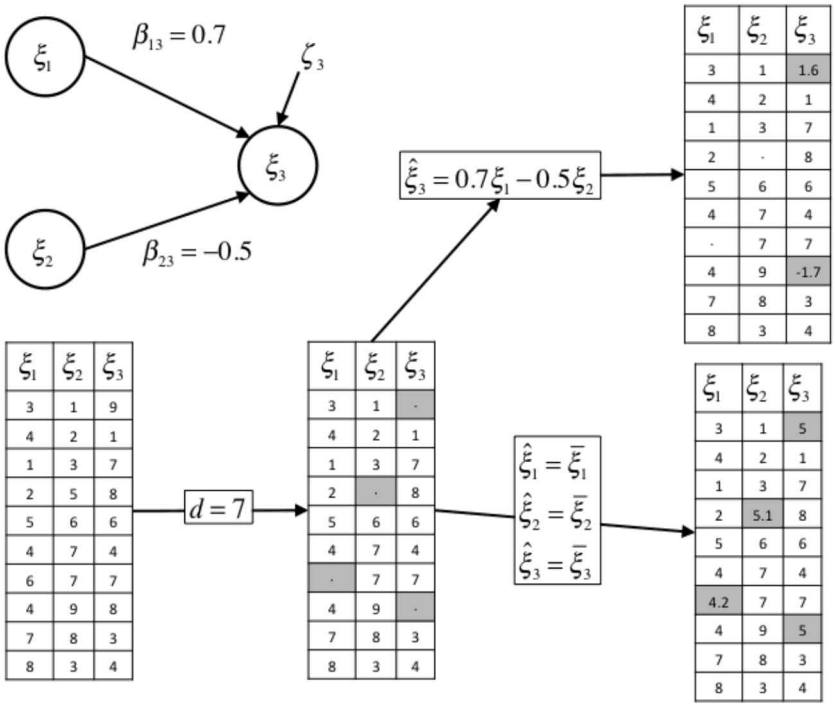
Estadísticos e-Books & Papers

realiza mediante el modelo estructural a partir de los valores estimados de ξ_1 y ξ_2 es $\hat{\xi}_3 = 0{,}7 \times 3 - 0{,}5 \times 1 = 1{,}6$.

5. A partir de las medias de las puntuaciones factoriales no omitidas de las variables latentes dependientes, se realiza la estimación *naïve* de ξ_3, es decir $\bar{\xi}_3 = (1 + 7 + 8 + 6 + 4 + 7 + 3 + 4)/9 = 5$.

6. El proceso se repite eliminando el siguiente $1/d = 1/7$ de la base de datos, así hasta haber eliminado y recalculado toda la matriz de datos original.

Figura 16.5.: Ilustración del procedimiento de *blindfolding*



Pues bien, si llamamos ξ_{kn} a la observación n del factor k que ha sido omitido, llamamos $\hat{\xi}_{kn}$ a la estimación de ese valor omitido utilizando la información del modelo estructural obtenido mediante PLS-SEM y $\bar{\xi}_{kn}$ a la estimación del valor omitido sin usar el modelo (la media de los valores no omitidos de ese factor), podremos calcular el error promedio que cometemos al estimar todos los valores perdidos de la siguiente forma (E_k utilizando PLS-SEM y O_k sin utilizarla):

$$E_k = \sum_{n=1}^{N} \left(\xi_{kn} - \hat{\xi}_{kn} \right)^2 \tag{16.26}$$

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 16.11.: Cálculo de la Q^2 de Stone-Geisser

```
Q2 predictive relevance statistics

 Block Q2
        dep       utiper     actitud     intcomp
 0.133182948  0.14148763  0.21483513  0.32063739
```

$$O_k = \sum_{n=1}^{N} \left(\xi_{kn} - \bar{\xi}_{kn}\right)^2 \tag{16.27}$$

De la lógica explicitada con anterioridad, es obvio que si se cometen menos errores usando el modelo que sin usarlo, este tendrá relevancia predictiva, es decir, la tendrá si $E_k < O_k$. Otra forma de escribirlo es mediante el indicador Q^2 de Stone-Geisser (Stone, 1974; Geisser, 1975):

$$Q_k^2 = 1 - \frac{E_k}{O_k} \tag{16.28}$$

donde el modelo tendrá relevancia predictiva si $Q^2 > 0$, que es equivalente a $E_k < O_k$.

El paquete `plspm` no ofrece el cálculo de la Q^2 de Stone-Geisser, pero sí el paquete `matrixpls`. La solicitud se realiza mediante la sintaxis siguiente, donde se ve claramente como la distancia de omisión se ha fijado a 7 en este caso. Es muy importante tener en cuenta que la distancia de omisión no puede ser un divisor exacto del tamaño muestral, dado que, en ese caso, el valor perdido no podría estimarse mediante los valores válidos del resto de variables de ese mismo caso porque todos serían casos perdidos.

```
predictions.blindfold <- matrixpls.crossvalidate(cov(datos),
model = modelo.lavaan, blindfold = TRUE,
predictionType = "redundancy",
groups = 7) q2(cov(datos),
predictions.blindfold, model=modelo.lavaan.pls)
```

El cuadro 16.11 confirma que todos los Q^2 son positivos, por lo que nuestro modelo exhibe relevancia predictiva.

C. Evaluación de la significatividad de las relaciones estructurales

Una vez comprobado que el modelo estimado tiene relevancia predictiva, el último paso es determinar qué hipótesis pueden confirmarse y cuáles no, es decir, determinar qué coeficientes de regresión son o no significativos. La explicación del procedimiento que sigue PLS-SEM para estimar la significatividad de dichos parámetros, el remuestreo mediante *bootstrapping* ya se ha presentado con

Estadísticos e-Books & Papers

Cuadro 16.12.: Resultados de la estimación del modelo estructural

```
---------------------------------------------------------
BOOTSTRAP VALIDATION
paths
                      Original  Mean.Boot  Std.Error  perc.025  perc.975
facuso -> utiper        0.5091      0.511     0.0378    0.4185     0.575
facuso -> dep           0.1954      0.198     0.0476    0.1144     0.295
facuso -> actitud       0.1539      0.157     0.0429    0.0815     0.246
utiper -> dep           0.5163      0.516     0.0413    0.4324     0.596
utiper -> actitud       0.4749      0.472     0.0443    0.3852     0.558
utiper -> intcomp       0.2546      0.249     0.0537    0.1456     0.348
dep -> actitud          0.1665      0.170     0.0437    0.0808     0.246
dep -> intcomp          0.0882      0.092     0.0556   -0.0150     0.197
actitud -> intcomp      0.3164      0.318     0.0512    0.2145     0.404
```

anterioridad y que calcula el estadístico t que contrasta la hipótesis nula de que un coeficiente de regresión determinado es nulo, igual que hacemos con una regresión normal. El cuadro muestra la salida, donde puede comprobarse como la única relación no significativa (el intervalo de confianza contiene al cero) se corresponde con el impacto directo de la dependencia del medio sobre la intención de compra ($\beta = 0,088$; $p > 0,05$). No entraremos a comentar los resultados que exigiría una presentación más detallada de las hipótesis, pero remitimos al lector al trabajo de Aldás *et al.* (2008), donde se discuten con profundidad.

16.4.5. El debate de los indicadores de ajuste global

Comentábamos al principio del capítulo que la elección entre los enfoques basados en covarianzas y varianzas para la estimación de modelos estructurales tienen que tener en cuenta las ventajas e inconvenientes de cada uno de ellos. Al final, muy a grandes rasgos, nos decantaremos por PLS-SEM cuando tengamos constructos formativos o muestras representativas pero pequeñas y por CB-SEM en el resto de casos porque si no es necesario querremos tener siempre estadísticos e indicadores del ajuste global de nuestro modelo y PLS-SEM no nos los ofrece.

Conscientes de la importancia de esta limitación, numerosos investigadores han propuesto —y siguen haciéndolo cada día— índices que pretenden sintetizar en un único indicador el ajuste global del modelo. En mi opinión esto supone ignorar el carácter de PLS-SEM, donde la parte estructural se descompone en regresiones parciales por lo que hablar de ajuste global es contradictorio con su esencia, a diferencia de los CB-SEM en los que, como vimos, los parámetros se estiman de manera simultánea para minimizar la diferencia entre las matrices de covarianzas muestrales y teóricas. Esa diferencia es, precisamente, el ajuste global.

Así, por ejemplo, y como se muestra en el cuadro 16.13, el paquete `plspm` implementa el denominado *goodness of fit index (GoF)* propuesto por Tenenhaus

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

et al. (2004) y que Henseler y Sarstedt (2013) demuestran que, ni el *GoF* ni el *GoF* relativo, (Esposito-Vinzi *et al.*, 2010) tienen capacidad alguna para validar los modelos, como sumo puede servir para ver si el modelo funciona de manera equivalente en dos conjuntos de datos distintos. En esta misma recomendación de no utilizar este criterio coinciden Hair *et al.* (2014b).

Recientemente Dijkstra y Henseler (2015) han abordado el problema de la inconsistencia de PLS-SEM al estimar de manera separada cargas y coeficientes estructurales que provoca que los primeros suelan estar sesgados al alza y los segundos a la baja. Esta inconsistencia solo desaparece cuando el número de indicadores por constructo es muy elevado (Gefen *et al.*, 2011). Introducen lo que denominan **consistent PLS** que, manteniendo las ventajas de PLS-SEM, logra estimar de manera consistente las cargas factoriales, correlaciones entre variables latentes y relaciones estructurales, no así los pesos, por lo que solo es aplicable a modelos en los que todos los constructos sean reflectivos.

La estimación consistente tiene varias ventajas, la primera de las cuales es que permite abordar de una manera mucho más sensata la búsqueda de indicadores de ajuste globales, incluso estando más cerca de un estadístico equivalente a la χ^2. Al estimar de manera consistente la matriz de correlaciones entre las variables latentes, como apuntan Henseler *et al.* (2016), puede analizarse el nivel de discrepancia entre la matriz que implica el modelo y la muestral. Se puede utilizar para ello un indicador que ya vimos al evaluar los modelos basados en covarianzas, el SRMR, que no es otra cosa que la raíz cuadrada de los residuos medios estandarizados y cuyo *benchmark* de <0,08 establecieron Hu y Bentler (1999). También proponen Dijkstra y Henseler (2015) que esas discrepancias se calculen mediante distancias euclídeas (d_{LS}) o mediante distancias geodésicas (d_G). El *bootstrapping* permitirá calcular estas distancias en submuestras de la muestra original. Cuando las discrepancias entre la matriz de correlación que implica el modelo y la muestral sean tan pequeñas que solo sean atribuibles al error del muestreo, será bastante plausible esperar que la muestra esté siendo extraída de una población de acuerdo con el modelo propuesto. Por lo tanto, igual que ocurre con el estadístico χ^2, el buen ajuste del modelo se establecería para $p > 0,05$.

Desgraciadamente, a día de hoy, ninguno de los dos paquetes de `R` que estamos utilizando proporciona estas distancias y sus intervalos de confianza obtenidos por *bootstrapping*. `matrixpls` sí que ofrece dos estimaciones distintas del SRMR. Por esta razón, solo a efectos informativos, mostramos en el cuadro 16.14 la salida de `SmartPLS 3.0,` que sí que los tiene implementados y que no podrían confirmar el buen ajuste del modelo, aunque ya sabemos que estos indicadores no les son aplicables en cuanto que solo son adecuados para modelos que pueden ser corregidos para lograr la consistencia, según la terminología de Dijkstra y Henseler (2015), es decir, aquellos donde todo el instrumento de medida sea reflectivo.

Estadísticos e-Books & Papers

Cuadro 16.13.: Indicadores de ajuste proporcionados por `matrixpls` y por `plspm`

`matrixpls`

```
Residual-based fit indices
                                    Value
Communality                     0.6730753
Redundancy                      0.1476459
SMC                             0.3907875
RMS outer residual covariance 0.3764617
RMS inner residual covariance 0.3871787
SRMR                            0.3616927
SRMR (Henseler)                 0.2942533

 Absolute goodness of fit: 0.5128639
```

`plspm`

```
--------------------------------------------
GOODNESS-OF-FIT
[1]  0.4283

--------------------------------------------
```

Cuadro 16.14.: Indicadores de ajuste para la estimación consistente proporcionados por `SmartPLS 3.0`

SRMR

Media, desviación estándar, valores t, p valores | Intervalos de confianza | Muestras | Copiar

	Muestra original (C	Media de la muest	Desviación estánd	Estadísticos t (| O/	P Valores
Modelo saturado	0,066	0,040	0,002	35,394	0,000
Modelo estimado	0,067	0,041	0,002	35,685	0,000

d_ULS

Media, desviación estándar, valores t, p valores | Intervalos de confianza | Muestras | Copia

	Muestra original (C	Media de la muest	Desviación estánd	Estadísticos t (| O/	P Valores
Modelo saturado	1,412	0,528	0,049	28,796	0,000
Modelo estimado	1,456	0,535	0,050	29,261	0,000

d_G

Media, desviación estándar, valores t, p valores | Intervalos de confianza | Muestras | Copiar

	Muestra original (C	Media de la muest	Desviación estánd	Estadísticos t (| O/	P Valores
Modelo saturado	0,417	0,205	0,020	21,218	0,000
Modelo estimado	0,426	0,208	0,020	21,254	0,000

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

16.5. Presentación de los resultados en una publicación

La presentación de los resultados siempre es un planteamiento muy personal y está, evidentemente, sujeta a las restricciones y exigencias de las normas editoriales de las revistas. En cualquier caso, el principio que debe regirlas es el de economía (utilizar el menor número posible de cuadros) y el de transparencia (que el lector disponga de toda la información necesaria para tener un criterio claro acerca de cuáles son los resultados y la calidad de los mismos).

A continuación ofrecemos nuestra propuesta para, en solamente tres cuadros, sintetizar toda la información relevante. Dejamos al lector, como un ejercicio que debería realizar, el localizar en las salidas que se han presentado a lo largo del capítulo la información necesaria que hemos trasladado a los mismos[7].

[7]Recuérdese que los valores de las t no están presentes directamente en las tablas pero se calculan de manera inmediata mediante la expresión 16.21.

Estadísticos e-Books & Papers

Cuadro 16.15.: Fiabilidad, consistencia interna y validez convergente del instrumento de medida

Factor	Indicador	λ	t	γ	t	VIF	CA	CR	AVE
Dependencia	dep1	0,524**	7,72	0,280**	3,88	1,158	–	–	–
	dep2	0,315**	4,09	0,024	0,38	1,349			
	dep3	0,420**	5,93	-0,120	-1,47	1,645			
	dep4	0,928**	35,29	0,740**	11,88	1,396			
	dep5	0,609**	9,79	0,161*	1,88	1,879			
	dep6	0,592**	10,17	0,188**	2,30	1,604			
Utilidad percibida	utiper1	0,571**	12,77				0,86	0,89	0,59
	utiper2	0,792**	39,40						
	utiper3	0,820**	46,33						
	utiper4	0,810**	50,00						
	utiper5	0,778**	34,27						
	utiper6	0,797**	41,08						
Facilidad de uso	facuso1	0,816**	41,63				0,75	0,84	0,57
	facuso3	0,695**	17,96						
	facuso4	0,749**	29,03						
	facuso6	0,752**	28,92						
Actitud	actitud1	0,589**	14,06				0,86	0,89	0,50
	actitud2	0,660**	17,89						
	actitud4	0,762**	28,33						
	actitud5	0,834**	54,16						
	actitud6	0,796**	40,41						
	actitud7	0,748**	28,77						
	actitud8	0,565**	15,44						
	actitud10	0,667**	17,28						

$^{**}p < 0{,}01$; $^{**}p < 0{,}05$; CA = α de Cronbach; CR = Fiabilidad Compuesta; AVE = Varianza extraída promedio

© Ediciones Paraninfo

Estadísticos e-Books & Papers

Cuadro 16.16.: Validez discriminante

Factor	F1	F2	F3	F4	F5
F1. Actitud	**0,709**	–	0,564	–	0,757
F2. Dependencia	0,529	–	–	–	–
F3. Facilidad de uso	0,472	0,458	**0,754**	–	0,629
F4. Intención de uso	0,530	0,412	0,456	**1,000**	0,559
F5. Utilidad percibida	0,656	0,616	0,509	0,516	**0,766**

Nota: Diagonal, raíz cuadrada de la varianza extraída.
Triángulo inferior: correlaciones entre las variables latentes
Triángulo superior: ratio HTMT

Estadísticos e-Books & Papers

Cuadro 16.17.: Estimación del modelo estructural

Hipótesis	β estandarizado	Valor t bootstrap
H1: Facilidad de uso→Utilidad percibida	0,509**	13,47
H2: Facilidad de uso→Actitud compra Internet	0,154**	3,59
H3: Utilidad percibida→Actitud compra Internet	0,475**	10,72
H4: Utilidad percibida→Intención compra Internet	0,255**	4,74
H5: Actitud compra Internet→Intención compra Internet	0,316**	6,18
H6: Utilidad percibida→Dependencia Internet	0,516**	12,50
H7: Facilidad uso→Dependencia Internet	0,195**	4,11
H8: Dependencia Internet→Actitud compra Internet	0,167**	3,81
H9: Dependencia Internet→Intención compra Internet	0,088	1,59

R^2(Actitud)=0.472;R^2(Dependencia)=0.407;R^2(Intención)=0.336;R^2(Utilidad)=0.259
Q^2(Actitud)=0.215; Q^2(Dependencia)=0.133;Q^2(Intención)=0.320;Q^2(Utilidad)=0.141;
**$p < 0{,}01$; *$p < 0{,}05$

© Ediciones Paraninfo
Estadísticos e-Books & Papers

Índice de figuras

1.1. Ilustración de una escala de intervalo 22
1.2. Técnicas de análisis de dependencia 25
1.3. Técnicas de análisis de interdependencia 27

2.1. Valores de la variable SYS estandarizados (ZSYS) 45
2.2. Relación ingresos-edad 48
2.3. Relación ingresos-antigüedad en el puesto 48
2.4. Relación ingresos-beneficio de la empresa 49
2.5. Sintaxis para la obtención de la D^2 de Mahalanobis y su significatividad en R 53
2.6. Casos atípicos de acuerdo con la D^2 de Mahalanobis 54
2.7. Casos atípicos de acuerdo con la D^2 de Mahalanobis 55
2.8. Distribuciones normal, asimétricas, platicúrticas y leptocúrticas 58
2.9. Gráfico Q-Q para la variable peso 60
2.10. Gráfico ji-cuadrado 64
2.11. Transformaciones en búsqueda de normalidad 67
2.12. Histogramas de la variable SYS y su logaritmo 68
2.13. Ejemplos de homocedasticidad y heterocedasticidad 70
2.14. Gráficos de dispersión bivariados 75

3.1. Proceso de realización de un análisis de conglomerados 78
3.2. Gráfico de dispersión de los datos hipotéticos 80
3.3. Ilustración de la distancia *city block* 82
3.4. Dendograma 92
3.5. Historial de conglomeración "vecino más cercano" 94
3.6. Historial de conglomeración "vecino más lejano" 95
3.7. Datos simulados de cuatro conglomerados 104
3.8. Índices gráficos para la elección del número de conglomerados 105
3.9. Reglas para la determinación del centroide inicial 108
3.10. Proceso de determinación del centroide inicial 108
3.11. Dendogramas mediante resultantes de aplicar distintos métodos de conglomeración 118
3.12. Proceso de determinación del centroide inicial 119
3.13. Dendogramas mediante resultantes de aplicar distintos métodos de conglomeración 120
3.14. Visualización de los resultados de un análisis de conglomerados 123

Estadísticos e-Books & Papers

4.1. Ilustración de la aplicación del MDS a la distancia entre las capitales españolas de provincia 126
4.2. Ilustración de la aplicación del MDS a los datos de la imagen de cadenas de electrodomésticos 128
4.3. Diagrama de Shepard 131
4.4. Contribución relativa de cada punto al *stress* 134
4.5. Diagrama de residuos 135
4.6. Obtención de la matriz individual de proximidades 138
4.7. Gráfico de Shepard 144
4.8. Valor del *stress* para distinto número de dimensiones 146
4.9. Representación bidimensional de las regiones producida por el CMDS 147
4.10. Dendograma del análisis de conglomerados sobre las coordenadas de los estímulos 149
4.11. Grupos formados a partir del dendograma 149
4.12. Mapa conjunto de estímulos derivados del WMDS 151
4.13. Mapa de pesos de los sujetos del WMDS 154
4.14. Mapas perceptuales original (`smacof`) y elaborado 158

5.1. Finalidad básica del análisis de correspondencias 160
5.2. Gráfico descriptivo de los datos 162
5.3. Ilustración geométrica del concepto de inercia 165
5.4. Descomposición de la inercia total 167
5.5. Representación de la solución bidimensional 169
5.6. Representación de la solución bidimensional con puntos fila y columna adicionales 180
5.7. Descripción de la base de datos 183
5.8. Gráfico de sedimentación 184
5.9. Mapa perceptual 185
5.10. Gráfico de contribuciones a las dimensiones 187
5.11. Mapa con los individuos 188

6.1. Ilustración de un ANOVA 192
6.2. Valores medios e intervalos de confianza al 95 % para las medias 209
6.3. Valores medios del consumo de televisión por sexo y por nivel educativo 218
6.4. Ilustración del efecto interacción. Ni efectos principales ni efecto interacción 219
6.5. Ilustración del efecto interacción. Solo uno de los efectos principales significativos. Sin interacción. 220
6.6. Ilustración del efecto interacción. Ambos efectos principales significativos. Sin interacción. 221
6.7. Ilustración del efecto interacción. Efecto interacción significativo. 221

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

7.1. Ventas de ron y cola por semana y tipo de zona 256

8.1. Ejemplo de regresión lineal simple 263
8.2. Información total, explicada por el modelo y residual 273
8.3. Contraste de una y dos colas 278
8.4. Ejemplo de homocedasticidad y heteroscedasticidad 287
8.5. Gráficos de diagnóstico 297
8.6. Gráficos de componentes y residuos para el diagnóstico de la linealidad 301
8.7. Test de Cook para los casos influyentes 304
8.8. *Hat values* para detectar casos influyentes 306
8.9. Gráficos *Added variables* 307
8.10. Gráficos de influencia 308
8.11. Variable cualitativa con dos modalidades e influencia aditiva . 310
8.12. Variable cualitativa con dos modalidades e influencia multiplicativa 316
8.13. Pendiente y término independiente diferentes 318

9.1. Funciones de distribución hipotéticas de 2 grupos 325
9.2. Elipses de concentración de funciones de distribución de frecuencias y su proyección sobre los ejes X_1 y X_2 328
9.3. Elipses de concentración de funciones de distribución de frecuencias y su proyección sobre el eje discriminante 329
9.4. Funciones de distribución de frecuencias de las puntuaciones sobre el eje discriminante 330
9.5. Representación gráfica de los datos de partida 334
9.6. Función de Fisher sobre la representación gráfica de los datos de partida 336
9.7. Representación de las puntuaciones discriminantes de los individuos sobre las funciones discriminantes 360
9.8. Proyecciones de los casos sobre los ejes discriminantes 361
9.9. Mapa territorial 362

10.1. Ajuste de una recta y una función logística a una variable dicotómica 366
10.2. Ejemplo de curvas ROC 384
10.3. Curva ROC para el caso del Titanic 385

11.1. Datos originales centrados y su proyección sobre X_1^* 398
11.2. Datos originales centrados y su proyección sobre X_1^* 400
11.3. Datos originales y nuevos ejes (componentes principales) . . . 402
11.4. Interpretación gráfica de autovectores y autovalores 404
11.5. Casos extremos de relación entre variables 406
11.6. Matriz de correlaciones entre las variables del caso y significatividad (p<0,05). 418

Estadísticos e-Books & Papers

11.7. Gráfico de sedimentación 421
11.8. Gráfico del análisis paralelo 423
11.9. Gráfico de las cargas sobre las dos primeras componentes . . 427
11.10. Representación gráfica de los países sobre las primeras dos componentes 428
11.11. Representación gráfica conjunta de los países y de los sectores sobre las primeras dos componentes 429

12.1. Visión conceptual diferencial del PCA y del EFA 432
12.2. Representación gráfica de las tres estimaciones 451
12.3. Criterios para determinar el número de factores 453
12.4. Solución rotada y sin rotar del caso 12.1 (ejes principales) . . 455
12.5. Rotación ortogonal (varimax) y oblicua (oblimin) del caso 12.1 (ejes principales) 460
12.6. Interpretación geométrica de las cargas factoriales y las cargas de estructura 462
12.7. Resultado gráfico del análisis paralelo 472
12.8. Resultado gráfico del análisis paralelo 474
12.9. Mapa perceptual de las bebidas 476

13.1. Ejemplo de modelo de ecuaciones estructurales 480
13.2. Ejemplo de modelo de ecuaciones estructurales 481
13.3. Modelo de análisis factorial exploratorio 483
13.4. Modelo de análisis factorial confirmatorio 485
13.5. Modelo estructural elemental 488
13.6. CFA antes y después de la identificación 493
13.7. Histograma de residuos 508
13.8. CFA por estimar en el caso 13.2 520
13.9. CFA por estimar en el caso 13.2 522
13.10. Histograma de errores 526
13.11. Grafo del CFA estimado por `lavaan` 528

14.1. Modelo de dos factores para ilustrar la el criterio de la ratio HTMT 548
14.2. Matriz de correlaciones para ilustrar la ratio HTMT 549
14.3. CFA por estimar en el caso 13.2 551

15.1. Modelo de desempeño de la fuerza de ventas 569
15.2. Parámetros iniciales que se deben estimar en el modelo 579
15.3. Parámetros para estimar tras la identificación y parámetros restringidos 580
15.4. Gráfico de residuos estandarizados 586
15.5. Ejemplo de modelo de ecuaciones estructurales 591
15.6. Ilustración de la identificación del modelo 593
15.7. Gráfico de resultados 597

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

16.1. Ejemplo de un modelo estructural estimable mediante PLS-SEM 604
16.2. Incorporación de la dependencia del medio a un modelo TAM 615
16.3. Modelo de dos factores para ilustrar la el criterio de la ratio HTMT 618
16.4. Matriz de correlaciones para ilustrar la ratio HTMT 619
16.5. Ilustración del procedimiento de *blindfolding* 636

Estadísticos e-Books & Papers

Estadísticos e-Books & Papers

Índice de cuadros

2.1. Preguntas de actitud hacia el tabaco 32
2.2. Respuestas simuladas al cuestionario 33
2.3. Media de V4 en función del tipo de valor perdido 34
2.4. Prueba t para muestras independientes 35
2.5. Matriz de correlaciones 36
2.6. Estadísticos descriptivos 38
2.7. Descriptivos en eliminación de casos por parejas 39
2.8. Imputación por regresión 40
2.9. Descripción de la base de datos 43
2.10. Observaciones atípicas para cada variable 44
2.11. Observaciones atípicas para cada variable 46
2.12. Matriz X de datos para la detección de *outliers* 50
2.13. Resultado del test de Mahalanobis 52
2.14. Resultado del test de Mahalanobis 59
2.15. Cálculos necesarios para el gráfico Q-Q 60
2.16. Valores críticos para el coeficiente de correlación del gráfico Q-Q bajo el supuesto de normalidad 61
2.17. Test de contraste de la normalidad univariante 62
2.18. Test de contraste de la normalidad univariante 64
2.19. Resultados de los test de normalidad multivariante 65
2.20. Test de normalidad antes y después de la transformación . . . 66
2.21. Test de normalidad antes y después de la transformación . . . 72
2.22. Matrices de covarianzas para fumadores (1) y no fumadores (2) 73
2.23. Test M de Box 73

3.1. Inversión en publicidad y ventas de 8 empresas hipotéticas . . 79
3.2. Matriz de distancias euclídeas para los datos del ejemplo . . . 81
3.3. Base de datos hipotética de variables binarias 83
3.4. Cálculo de similitudes 83
3.5. Matriz de distancias euclídeas para los datos del ejemplo . . . 85
3.6. Activo y número de trabajadores de 8 empresas hipotéticas . 86
3.7. Matriz de distancias euclídeas para los datos del ejemplo . . . 86
3.8. Activo y número de trabajadores de 8 empresas hipotéticas . 87
3.9. Matriz de distancias euclídeas para los datos del ejemplo tras la normalización 88
3.10. Matriz de distancias euclídeas al cuadrado para los datos del caso 3.1 90

Estadísticos e-Books & Papers

3.11. Datos en el paso 2 del proceso de conglomeración e historial de conglomeración 91
3.12. Historial de conglomeración método "vinculación promedio" . 94
3.13. Distancias entre pares de observaciones en la etapa 6 del método "vinculación promedio" 96
3.14. Decisión de los grupos a fusionar en la etapa 6 mediante el método de "Ward" 97
3.15. Aplicación de las tasas de variación de los coeficientes de conglomeración 99
3.16. Decisión de los grupos a retener mediante distintos indicadores 105
3.17. Decisión de los grupos a fusionar en la etapa 6 mediante el método de Ward 106
3.18. Asignación de observaciones en el primer paso 110
3.19. Asignación de observaciones en el segundo paso 110
3.20. Centroides finales 111
3.21. Análisis de varianza sobre los conglomerados finales 112
3.22. Equipamiento de los hogares en distintas comunidades autónomas 115
3.23. Resultados de la detección de *outliers* 116
3.24. Centroides resultantes del método jerárquico 119
3.25. Centroides resultantes del método no jerárquico 121
3.26. Significatividad de las diferencias entre los perfiles de los conglomerados 122

4.1. Solución bidimensional 130
4.2. Distancias entre las configuraciones 130
4.3. Matriz de disparidades 130
4.4. Interpretación del *stress* en términos de bondad de ajuste del MDS 132
4.5. Valor del *stress* y otros indicadores de ajuste para los datos del ejemplo 133
4.6. Información para el cálculo de los indicadores de bondad del MDS 136
4.7. Descriptores de destinos turísticos 137
4.8. Indicadores de desarrollo educativo 142
4.9. Valor del *stress* y otros indicadores de ajuste para los datos del caso 145
4.10. Coordenadas de los estímulos 147
4.11. Etiquetas asociadas a las comunidades autónomas 151
4.12. Pesos de la configuración para cada departamento 152
4.13. Indicadores objetivos del mapa comercial español 153
4.14. Datos para el estudio de posicionamiento 156

5.1. Relación entre puesto de trabajo y hábito de fumar 162
5.2. Perfiles fila y columna de la matriz de datos 163

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

5.3. Masas, inercia y coordenadas estandarizadas de los objetos representados . . . 164
5.4. Inercia total y autovalores . . . 166
5.5. Test χ^2 . . . 168
5.6. Contribución de los puntos a las dimensiones, de las dimensiones a los puntos y coordenadas para la representación . . . 170
5.7. Masas fila y columna . . . 172
5.8. Matriz de correspondencias . . . 172
5.9. Perfiles fila . . . 173
5.10. Perfiles columna . . . 173
5.11. Matriz de residuos estandarizados . . . 174
5.12. Matriz de valores singulares ($\mathbf{\Gamma}$) y vectores singulares izquierdos ($\mathbf{U}$) y derechos ($\mathbf{V}$) . . . 175
5.13. Relación entre puesto de trabajo y hábito de fumar con filas y columnas adicionales . . . 179
5.14. Calidad de la representación de los puntos fila y columna suplementarios . . . 181
5.15. Autovalores e inercia explicada . . . 182
5.16. Calidad de la representación de los niveles de las variables . . . 186

6.1. Datos de las puntuaciones obtenidas por los empleados adiestrados con distintos métodos . . . 191
6.2. Cálculos para la descomposición de la suma de cuadrados en el caso 6.1 . . . 195
6.3. Resultados de la estimación del ANOVA del caso 6.1 . . . 197
6.4. Test de Brown-Forsythe . . . 199
6.5. Test de Welch . . . 199
6.6. Test de Levene . . . 199
6.7. Cálculo del test de Kruskal-Wallis . . . 201
6.8. Resultados del test de Kruskal-Wallis . . . 202
6.9. Pruebas *post hoc* de Bonferroni y Holm . . . 206
6.10. Test *post hoc* de Tukey . . . 206
6.11. Test *post hoc* robustos . . . 207
6.12. Descriptivos de las variables dependientes en los grupos . . . 208
6.13. Test de Levene . . . 210
6.14. Contraste de hipótesis . . . 211
6.15. Pruebas *post hoc* . . . 212
6.16. Test de normalidad . . . 213
6.17. ANOVA no paramétrico de Kruskal-Wallis . . . 214
6.18. Estadísticos descriptivos . . . 217
6.19. Resultados de la estimación del ANOVA de dos factores . . . 217
6.20. Pruebas *post hoc* para el nivel de estudios . . . 225

7.1. Datos para la ilustración del MANOVA . . . 229
7.2. Cálculo de los productos cruzados de $\mathbf{T}$ . . . 232

Estadísticos e-Books & Papers

7.3. Matrices residual y factorial del MANOVA 235
7.4. Contraste de hipótesis con la λ de Wilks 237
7.5. $\mathbf{FW^{-1}}$ 238
7.6. Autovalores y autovectores de $\mathbf{FW^{-1}}$ 239
7.7. Traza V de Pillai-Bartlett 240
7.8. Traza T^2 de Hotelling 241
7.9. Raíz mayor de Roy 241
7.10. Pruebas de normalidad multivariante 243
7.11. Test M de Box para analizar la igualdad de las matrices de covarianzas 245
7.12. Test de esfericidad de Barlett 246
7.13. MANOVA para el contraste *post hoc* 248
7.14. Pruebas t para el contraste *post hoc* 249
7.15. Medias muestrales para cada combinación de los factores A y B 251
7.16. Datos sobre las ventas de dos productos de la cadena Mercanova 253
7.17. Matrices suma de cuadrados y productos cruzados 254
7.18. Estimación de la significatividad de los efectos 255
7.19. Test de normalidad multivariante de Shapiro 258
7.20. Test M de Box de igualdad de las matrices de covarianzas . . 259
7.21. Test de esfericidad de Barlett 260

8.1. Datos simulados para el caso 8.1 262
8.2. Parámetros estimados para la regresión lineal 266
8.3. Datos sobre cantidades y precios del café 267
8.4. Parámetros estimados para el caso 8.2 y análisis de la significatividad global 267
8.5. Parámetros estimados para el caso 8.3 271
8.6. Modelos general y restringido para el caso 8.4 280
8.7. Análisis de la varianza de los modelos general y restringido para el caso 8.4 281
8.8. Estadísticos descriptivos de *absen* 285
8.9. Análisis de la varianza en la estimación del caso 8.3 285
8.10. Análisis de la multicolinealidad con el VIF 291
8.11. Análisis de la multicolinealidad con el índice de condición . . 292
8.12. Test de Jarque y Bera de normalidad multivariante 295
8.13. Test de Shapiro y de Kolmogorov para analizar la normalidad de los residuos 296
8.14. Contrastes de homocedasticidad 300
8.15. Test de Durbin-Watson para el contraste de la autocorrelación 302
8.16. Test de Durbin-Watson para el contraste de la autocorrelación 303
8.17. Estimación del modelo $\ln(salario) = \beta_0 + \delta_0 mujer + \beta_1 educ + \varepsilon$ 311
8.18. Estimación del modelo $\ln(salario) = \beta_0 + \theta_1 mediana + \theta_2 grande + \beta_1 educ + \varepsilon$ 313
8.19. Anova de los modelos original y restringido 314

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

8.20. Estimación del modelo $\ln(salario) = \beta_0 + \beta_1 educ + \delta_1 mujer \times educ + \varepsilon$ 317
8.21. Estimación del modelo general y restringido 319
8.22. ANOVA del modelo general y restringido 320

9.1. Datos sobre características de préstamos concedidos en el Banco de Ademuz (datos en decenas de miles de euros) 323
9.2. Porcentaje de clasificaciones correctas e incorrectas utilizando la variable *patrimonio neto* 326
9.3. Porcentaje de clasificaciones correctas e incorrectas utilizando la variable *deudas pendientes* 326
9.4. Coeficientes de la función discriminante de Fisher 334
9.5. Aplicación de la función discriminante de Fisher 335
9.6. Aciertos y errores en la clasificación 337
9.7. Pertenencia al grupo predicha 337
9.8. Significatividad de la función discriminante 339
9.9. Autovalores del MANOVA 339
9.10. Matriz SCPC residual 340
9.11. Coeficientes estandarizados de la función discriminante 341
9.12. Matriz de correlaciones residual 342
9.13. Matriz de estructura 342
9.14. Test M de Barlett-Box 343
9.15. Contraste de la normalidad multivariante 344
9.16. Predicción de nuevos casos 345
9.17. Probabilidades a posteriori partiendo de probabilidades a priori iguales para los dos grupos 347
9.18. Probabilidades a posteriori partiendo de probabilidades a priori de 10 % para el grupo I y 90 % para el grupo II 348
9.19. Datos sobre los préstamos concedidos en el Banco de Buñol . 351
9.20. Medias de las variables explicativas en los tres grupos 355
9.21. Funciones discriminantes 356
9.22. Matriz de confusión 356
9.23. Significatividad conjunta de las funciones discriminantes . . . 357
9.24. Autovalores y proporción de la varianza explicada 357
9.25. Coeficientes estandarizados de las funciones discriminantes . . 359
9.26. Matriz de estructura 359
9.27. Comprobación de las hipótesis del modelo 363
9.28. Probabilidades a posteriori 364

10.1. Datos simulados para la ilustración 10.1 367
10.2. Resultados de -2LL 370
10.3. Resultados del contraste de significatividad global del modelo 370
10.4. Contraste individual de los coeficientes de regresión 371
10.5. *Odd ratio* en el modelo estimado 373
10.6. Matriz de confusión 376

Estadísticos e-Books & Papers

10.7. Indicadores de ajuste con funciones predefinidas 377
10.8. Descripción de los datos del caso 10.1 377
10.9. Relación univariante entre las variables independientes y la supervivencia de los pasajeros del Titanic 378
10.10. Significatividad global: ratio de máxima verosimilitud 379
10.11. Significatividad de las variables individuales y *odds ratio*. . . 380
10.12. Indicadores de ajuste y matriz de confusión 382
10.13. Test de Hosmer-Lemeshow 386
10.14. Descripción datos caso 10.2 389
10.15. Análisis bivariado de las variables dependiente e independientes 389
10.16. Resultados de la estimación del modelo 390
10.17. Significatividad global del modelo 391
10.18. Coeficientes de regresión del modelo y su significatividad . . . 391
10.19. Coeficientes de regresión del modelo y su significatividad . . . 392
10.20. *Risk ratios* del modelo estimado 393

11.1. Datos originales y en desviaciones respecto a la media 397
11.2. Datos originales centrados y nueva variable x_1^* para una rotación de $\theta = 10°$ 399
11.3. Varianzas explicadas por la nueva variable x_1^* para distintos ángulos de rotación θ 400
11.4. Datos originales centrados y nueva variable x_1^* para una rotación de $\theta = 10°$ 401
11.5. Autovalores y autovectores 407
11.6. Puntuaciones en las componentes principales 409
11.7. Cargas factoriales. Matriz factorial. 410
11.8. Variables de la base de datos de empleo 416
11.9. Descriptivos de las variables del caso 11.1 417
11.10. Test de esfericidad de Bartlett 418
11.11. Aplicación del criterio del autovalor >1 420
11.12. Análisis paralelo de Horn (1965) 422
11.13. Estadísticos de Bartlett (1950), Anderson (1963) y Lawley (1956) 424
11.14. Correlaciones de las variables con las componentes (cargas) . 425
11.15. Coordenadas de los objetos (países) en las primeras dos componentes 428

12.1. Matriz de correlaciones de las notas de 220 estudiantes en 6 asignaturas 433
12.2. Soluciones factoriales alternativas al caso 12.1 439
12.3. Autovalores y autovectores 441
12.4. Matriz $\mathbf{\Lambda}$ con las cargas factoriales 441
12.5. Solución de dos factores para el caso 12.1 con extracción por componentes principales 442
12.6. Estimación del EFA mediante `principal{psych}` 444
12.7. Estimación del EFA con extracción por ejes principales 446

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

12.8. Estimación del EFA con extracción por máxima verosimilitud 448
12.9. Cargas factoriales estimadas con una rotación ortogonal y extracción con ejes principales 456
12.10. Matriz de giro 458
12.11. Matrices de estructura y factoriales para una rotación oblicua y una ortogonal en el caso 12.1 463
12.12. Test de esfericidad de Barlett 464
12.13. Medidas KMO de adecuación muestral general e individual . 466
12.14. Matriz residual de la extracción con ejes principales y rotación varimax 467
12.15. Matriz **B** con los coeficientes de regresión para la estimación de las puntuaciones factoriales 469
12.16. Preguntas para la medición de las percepciones 470
12.17. Resultados del test de esfericidad de Barlett y el test KMO . 471
12.18. Resultados del análisis paralelo y el test de Barlett para la determinacion del número de factores que se deben extraer . . 472
12.19. Matriz de cargas rotadas 473
12.20. Puntuaciones factoriales 475
12.21. Promedio de las variables por tipo de bebida 477

13.1. Matriz de correlaciones entre las notas de los 275 estudiantes 482
13.2. Datos simulados para la ilustración 489
13.3. Resultados de la estimación 490
13.4. Estimación de las cargas factoriales 504
13.5. Estimación de las varianzas y covarianzas de los factores . . 505
13.6. Matriz de varianzas covarianzas estimada 505
13.7. Cálculo de la matriz de covarianzas residual 506
13.8. Matriz de correlaciones residual 507
13.9. Estadístico χ^2 510
13.10. Indicadores de ajuste del modelo 511
13.11. Indicadores de ajuste del modelo 515
13.12. Índices de modificación (MI) 518
13.13. Instrumento de medida del modelo 519
13.14. Indicadores de ajuste del CFA 525
13.15. Estimación de las cargas factoriales 526
13.16. Estimación de las varianzas de factores, errores y covarianzas entre factores 527

14.1. Escalas para medir dos dimensiones de la conciencia social . . 537
14.2. Estadísticos descriptivos para la subescala "ayuda a los conocidos" 537
14.3. α de Cronbach para la subescala "ayuda a los conocidos" . . . 538
14.4. α de Cronbach para la subescala "ayuda a los conocidos" cuando se elimina cada uno de los ítems 538
14.5. α de Cronbach para las dos subescalas correctamente calculadas 539

Estadísticos e-Books & Papers

14.6. Cargas estandarizadas del CFA 541
14.7. Cálculo de la fiabilidad compuesta CR 541
14.8. Fiabilidad compuesta CR 542
14.9. Varianza extraída promedio (AVE) 545
14.10. Estimación de la correlación entre los factores 546
14.11. Criterio de la ratio HTMT 549
14.12. Estimación de las cargas factoriales 552
14.13. Cálculo de α de Cronbach, AVE y CR 554
14.14. Estimación de las correlaciones entre factores 555
14.15. Intervalos de confianza para las correlacione entre los factores 555
14.16. Criterio de Fornell y Larcker (1981) 556
14.17. Ratio HTMT 556
14.18. Fiabilidad, consistencia interna y validez convergente del instrumento de medida. 557
14.19. Validez discriminante 558

15.1. Resumen del componente estructural 571
15.2. Resumen del instrumento de medida 574
15.3. Resultados no estandarizados de la estimación 584
15.4. Resultados estandarizados de la estimación 584
15.5. Residuos estandarizados de la estimación 585
15.6. Indicadores de ajuste del modelo 586
15.7. Parámetros estimados 588
15.8. Índices de modificación 589
15.9. Ajuste del modelo 596
15.10. Estimación de la parte estructural 598
15.11. Contraste de hipótesis 599

16.1. Estructura de la base de datos 621
16.2. Análisis del tamaño de las cargas 622
16.3. Análisis del tamaño de las cargas en la segunda estimación . 624
16.4. Análisis del tamaño de las cargas 624
16.5. Correlación entre las variables latentes 625
16.6. Matriz con los ratios HTMT 626
16.7. Significatividad de las cargas factoriales obtenida mediante *boostrapping* 628
16.8. Diagnóstico de colinealidad en los indicadores del constructo formativo. VIF. 630
16.9. Decisión sobre el mantenimiento o no de los indicadores no significativos del constructo formativo 632
16.10. Cálculo de la potencia lograda en la estimación y del tamaño que hubiera sido necesario para $1 - \beta = 0{,}80$ 634
16.11. Cálculo de la Q^2 de Stone-Geisser 637
16.12. Resultados de la estimación del modelo estructural 638
16.13. Indicadores de ajuste proporcionados por `matrixpls` y por `plspm`640

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

16.14. Indicadores de ajuste para la estimación consistente proporcionados por `SmartPLS 3.0` 640
16.15. Fiabilidad, consistencia interna y validez convergente del instrumento de medida 642
16.16. Validez discriminante 643
16.17. Estimación del modelo estructural 644

Estadísticos e-Books & Papers

Estadísticos e-Books & Papers

Bibliografía

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Aldás, J. (2014). Confirmatory tetrad analysis as a tool to decide between the formative/reflective nature of constructs in marketing and management research. En Moutinho, L., Bigné, E., y Manrai, A., editores, *The Routledge Companion to the Future of Marketing*, capítulo 19, pp. 348–378. Routledge, Londres.

Aldás, J. (2016). *Modelización estructural con PLS-SEM: Constructos de segundo orden.* ADD Editorial, Madrid.

Aldás, J., Martí, J., Sanz, S., y Ruiz, C. (2013). Key factors of teenagers' mobile advertising acceptance. *Industrial Management & Data Systems*, 113(5):732–749.

Aldás, J., Sanz, S., y Ruiz, C. (2008). La influencia de la dependencia del medio en el comercio electrónico B2C. Propuesta de un modelo integrador aplicado a la intención de compra futura en Internet. *Cuadernos de Economía y Dirección de la Empresa*, 11(36):45–75.

Anderson, J. y Gerbing, D. (1988). Structural equation modeling in practice: a review and recommended two-step approach. *Psychological Bulletin*, 103(2):411–423.

Anderson, T. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics*, 34(1):122–148.

Anscombe, F. (1960). Rejection of outliers. *Technometrics*, 2(2):123–146.

Bagozzi, R. P. (1980). Performance and satisfaction in an industrial sales forece: An examination of their antecedents and simultaneity. *Journal of Marketing*, 44(2):65–77.

Bagozzi, R. P. y Yi, Y. (1988). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science*, 16(1):74–94.

Bahuley, T. (2004). Understanding statistical power in the context of applied research. *Applied Ergonomics*, 35(2):73–80.

Ball-Rokeach, S. J. (1985). The origins of individual media system dependency: A sociological framework. *Communication Research*, 12(4):485–510.

Ball-Rokeach, S. J. (1989). Media system dependency theory. En DeFleur, M. y Ball-Rokeach, S. J., editores, *Theories of Mass Communication*, pp. 297–327. Longman, Nueva York, 5ª edición.

Barnett, V. y Lewis, T. (1994). *Outliers in Statistical Data.* Wiley, Nueva York, 3ª edición.

Bartlett, M. (1950). Tests of significance in factor analysis. *British Journal of Psychology*, 3(2):77–85.

Estadísticos e-Books & Papers

Bartlett, M. (1951). The effect of standardization on a chi square. Approximation in factor analysis. *Biometrika*, 38(3/4):337–344.

Basilevsky, A. (1994). *Statistical factor analysis and related methods.* Wiley series in probability and statistics. Wiley, Nueva York.

Bauer, H., Barnes, S., Reichardt, T., y Neumann, M. (2005). Driving consumer acceptance of mobile marketing: a theoretical framework and empirical study. *Journal of Electronic Commerce Research*, 6(3):181–191.

Bekker, P., Merckens, A., y Wansbeek, T. (1994). *Identification, equivalent models and computer algebra.* Academic Press, San Diego, CA.

Belsey, D. (1991). *Conditioning diagnostics, collinearity and weak data in regression.* John Wiley & Sons, Nueva York.

Belsey, D., Kuh, E., y Welsch, R. (1980). *Regression diagnostics.* John Wiley & Sons, Nueva York.

Bemmaor, A. C. y Mouchoux, D. (1991). Measuring short-term effect on in-stores promotion and retail advertising on brand sales: A factorial experiment. *Journal of Marketing Research*, 28(2):202–214.

Benjamini, Y. y Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Benjamini, Y. y Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107(2):238–246.

Bentler, P. M. (1995). *EQS 6 structural equations program manual.* Multivariate Software, Encino, CA.

Bentler, P. M. y Bonett, D. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88:586–606.

Bentler, P. M. y Chou, C.-P. (1997). Practical issues in structural equation modeling. *Sociological Methods and Research*, 16(1):78–117.

Benzécri, J. (1973). *L'Analyse des Donées. Tome 1: La taxinomie. Tome 2: L'analyse des correspondances.* Dunod, París.

Berry, W. (1993). *Understanding regression assumptions.* Sage University Paper Series on Quantitative Applications in the Social Sciences 07-092. Sage, Newbury Park, CA.

Blasius, J. y Greenacre, M. J. (1994). Computation of correspondence analysis. En Greenacre, M. J. y Blasius, J., editores, *Correspondence Analysis in the Social Sciences: Recent developments and its applications*, pp. 53–78. Academic Press, San Diego, CA.

Bock, H. H. (1985). On some significance tests in cluster analysis. *Journal of Classification*, 2(1):77–108.

Bohrnstedt, G. (1976). *Evaluación de la confiabilidad y validez en la medición de actitudes.* Trillas, México.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika*, 50(2):229–242.

Box, G. (1949). A general distribution theory for a class of likelihood criteria.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Biometrika, 36(3/4):317–346.

Bray, J. y Maxwell, S. (1985). *Multivariate analysis of variance.* Sage University Paper Series on Quantitative Applications in the Social Sciences 07-054. Sage, Newbury Park, CA.

Breusch, T. y Pagan, A. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47(5):1287–1294.

Brown, M. y Forsythe, A. (1974). The small behaviour of some statistics which test the equality of several means. *Technometrics*, 16(1):129–132.

Brown, R. L. (1994). Efficacy of the indirect approach for estimation structural equation models with missing data: A comparison of five methods. *Structural Equation Modeling: A Multidisciplinary Journal*, 1(4):287–316.

Brown, T. A. (2006). *Confirmatory Factor Analysis.* The Guildford Press, New York, 1 edición.

Browne, M. y Cudek, R. (1993). Alternate ways of assessing model fit. En Bollen, K. y Long, J., editores, *Testing structural equation models*, pp. 136–162. Sage, Newbury Park, CA.

Browne, M. W. (1968). A comparison of factor analytic techniques. *Psychometrika*, 33(3):267–334.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications and programming.* Lawrence Erlbaum Associates, Mahwah, NJ.

Byrne, B. M. (2006). *Structural equation modeling with EQS: Basic concepts, applications and programming.* Lawrence Erlbaum Associates, Mahwah,NJ, 2 edición.

Calinski, T. y Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics, Theory and Methods*, 3(1):1–27.

Carroll, J. y Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart Young decomposition. *Psychometrika*, 35(3):283–319.

Cattell, R. B. (1966). The meaning and strategic use of factor analysis. En Cattell, R. B., editor, *Handbook of multivariate experimental psychology*, pp. 174–243. Rand McNally, Chicago.

Cattell, R. B. y Jaspers, J. M. F. (1967). A general plasmode for factor analytic exercises and research. *Multivariate Behavioral Research Monographs*, 3:1–212.

Cerny, C. y Kaiser, H. F. (1977). A study of a measure of sampling adequacy for factor-analytic correlation matrices. *Multivariate Behavioral Research*, 12(1):43–47.

Chakravarti, I. M., Laha, R., y Roy, J. (1967). *Handbook of methods of applied statistics*, volumen I de *Wiley series in probability and mathematical statistics.* John Wiley & Sons, Nueva York.

Charrad, M., Ghazzali, N., Boiteau, V., y Niknafs, A. (2014). Nbclust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, 61(6):1–36.

Chin, W. W. (1998). The partial least squares approach to structural equation modeling. En Marcoulides, G. A., editor, *Modern methods for business*

Estadísticos e-Books & Papers

research, pp. 295–336. Lawrence Erlbaum Associates, Mahwah,NJ.

Chin, W. W. (2000). *FAQ-Partial Least Squares and PLSGraph.* http://disc-nt.cba.uh.edu/chin/plsfaq.htm.

Chin, W. W. y Newsted, P. R. (1999). Structural equation modeling analysis with small samples using partial least squares. En Hoyle, R. H., editor, *Statistical strategies for small sample research*, pp. 307–341. Thousand Oaks, CA.

Churchill, G. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, 16(1):64–73.

Clark, L. y Watson, D. (1995). Constructing validity: basic issues in objective scale development. *Psychological Assessment*, 7(3):309–319.

Clarkson, D. B. y Jennrich, R. I. (1988). Quartic rotation criteria and algorithms. *Psychometrika*, 53(2):251–259.

Clausen, S. E. (1998). *Applied Correspondence Analysis.* Sage University Paper Series on Quantitative Applications in the Social Sciences 07-121. Sage Publications, Thousand Oaks, CA.

Cliff, N. (1988). The eigenvalue-greater-than-one rule ant the reliability of components. *Psychological Bulletin*, 103(2):276–279.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Lawrence Erlbaum Associates, Hilsdale, NJ, 2ª edición.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1):155–159.

Cohen, J. (1994). The earth is around p<.05. *American Psychologist*, 49(12):155–159.

Cook, R. y Weisberg, S. (1982). *Residuals and influence in regression.* Chapman & Hall, Nueva York.

Cook, R. y Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika*, 70(1):1–10.

Coombs, C. H. (1964). *A theory of data.* John Wiley & Sons, Nueva York.

Cox, D. y Snell, E. (1989). *Analysis of binary data.* Chapman & Hall, Londres, 2ª edición.

Cragg, J. G. y Uhler, R. S. (1970). The demand for automobiles. *The Canadian Journal of Economics*, 3(3):386–406.

Crawley, M. J. (2013). *The R book.* Wiley, Chichester, Reino Unido, 2ª edición.

Crocker, L. y Algina, J. (1986). *Introduction to classical and modern test theory.* Harcourt Brace Jovanovich, Nueva York.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334.

Davies, D. y Bouldin, D. (1979). A cluster separation measure. *IEEE Transactions in Pattern Analysis and Machine Intelligence*, 1(2):224–227.

Davis, F. (1989). Perceived usefulness, perceived ease of use and user acceptance of information technology. *MIS Quarterly*, 13(3):319–340.

De Vellis, R. F. (1991). *Scale development, theories and applications.* Applied social research methods series 26. Sage Publications, Londres.

Diamantopoulos, A., Riefler, P., y Roth, K. (2008). Advancing formative measurement models. *Journal of Business Research*, 61(12):1203–1218.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Diamantopoulos, A. y Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research*, 38(2):269–277.

Dice, L. (1945). Measures of the amount of ecological association between species. *Ecology*, 26:297–302.

Dijkstra, T. K. y Henseler, J. (2015). Consistent and asymptotically normal PLS estimators for linear structural equations. *Computational Statistics and Data Analysis*, 81(1):10–23.

Dillon, W. R. y Goldstein, M. (1984). *Multivariate Data Analysis. Methods and applications.* John Wiley & Sons, Nueva York.

Dinno, A. (2009). Exploring the sensitivity of Horn's parallel analysis to the distributional form of simulated data. *Multivariate Behavioral Research*, 44(3):362–388.

Drolet, A. y Morrison, D. G. (2001). Do we really need multiple-item measures in service research? *Journal of Service Research*, 3(3):196–204.

Ducoffe, R. (1996). Advertising value and advertising on the web. *Journal of Advertising Research*, 36(5):21–35.

Duda, R. y Hart, P. (1973). *Pattern classification and scene analysis.* John Wiley & Sons, Nueva York.

Dunn, J. (1974). Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4(1):95–104.

Esposito-Vinzi, V., Trinchera, L., y Amato, S. (2010). PLS path modeling: from foundations to recent developments and open issues for model assessment and improvement. En Esposito-Vinzi, V., Chin, W. W., Henseler, J., y Wang, H., editores, *Handbook of partial least squares: Concepts, methods and applications and related fields*, pp. 47–82. Springer, Heidelberg, Alemania.

Everitt, B. S. (1979). A Monte Carlo investigation of the robustness of Hotelling's one and two samples T2 tests. *Journal of the American Statistical Association*, 74(365):48–51.

Fabrigar, L., Wegener, D., MacCallum, R. C., y Strahan, E. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3):272–299.

Field, A. (2005). *Discovering Statistics using SPSS.* Sage, Thousand Oaks, CA, 2ª edición.

Field, A., Miles, J., y Field, Z. (2012). *Discovering statistics using R.* Sage, Londres.

Filliben, J. J. (1975). The probability plot correlation coefficient test for normality. *Technometrics*, 17(1):111–117.

Fisher, F. (1936). The use of multiple measurements in taxonomic problems. *Journal of Human Genetics*, 7(2):179–188.

Florek, K., Lukaszewicz, J., Perkal, J., y Zubrzycki, S. (1951). Sur la liaison et la division des points d'un ensemble fini. *Colloquium Matematicae*, 2(3-4):282–285.

Forgy, E. (1965). Cluster analysis of multivariate data: Efficiency vs interpretability of classifications. *Biometrics*, 21(3):768–769.

Estadísticos e-Books & Papers

Fornell, C. y Bookstein, F. L. (1982). Two structural equation models: LISREL and PLS applied to consumer exit-voice theory. *Journal of Marketing research*, 19(4):440–452.

Fornell, C. y Cha, J. (1994). Partial Least Squares. En Bagozzi, R. P., editor, *Advanced methods of marketing research*, pp. 53–78. Basil Blackwell, Cambridge, Massachusetts.

Fornell, C. y Larcker, D. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1):39–50.

Fox, J. (2006). Structural equation modeling with the sem package in R. *Structural Equation Modeling*, 13(3):465–486.

Gefen, D., Rigdon, E., y Straub, D. W. (2011). An update and extension to SEM guidelines for administrative and social science research. *MIS Quarterly*, 35(2):iii–xiv.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350):320–328.

Glass, G. V., Peckham, P. D., y Sanders, J. R. (1972a). Consequences of failure to meet assumptions underlying the fixed effects of analysis of variance and covariance. *Educational Research*, 42(3):237–288.

Glass, G. V., Peckham, P. D., y Sanders, J. R. (1972b). Consequences of failure to meet assumptions underlying the fixed effects of analysis of variance and covariance. *Review of Educational Research*, 42(3):237–288.

Gnanadesikan, R. (1977). *Methods for statistical analysis of multivariate observations.* Wiley, Nueva York.

Gold, A., Malhotra, A., y Segars, A. (2001). Knowledge management: An organizational capabilities perspective. *Journal of Management Information Systems*, 18(1):185–214.

Goldfeld, S. M. y Quandt, R. E. (1965). Some tests for homoscedasticity. *Journal of the American Statistical Association*, 60(310):539–547.

Goldfeld, S. M. y Quandt, R. E. (1972). *Nonlinear methods in econometrics.* North Holland, Amsterdam.

Gordon, A. (1999). *Classification.* Chapman & Hall CRC, Londres, 2ª edición.

Gorsuch, R. (1983). *Factor Analysis.* Lawrence Erlbaum, Hillsdale, NJ.

Gorsuch, R. (1990). Common factor analysis versus component analysis: Some well and little known facts. *Multivariate Behavioral Research*, 25(1):33–39.

Gower, J. y Legendre, P. (1986). Metric and euclidean properties of dissimilarity coefficients. *Journal of Classification*, 3:5–48.

Grant, A., Guthrie, K., y Ball-Rokeach, S. J. (1991). Television shopping: A media system dependency perspective. *Communication Research*, 18(6):773–798.

Green, P. E., Carmone, F. J., y Smith, S. M. (1989). *Multidimensional Scaling: Concepts and Applications.* Allyn & Bacon, Boston, MA.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis.* Academic Press, Londres.

Greenacre, M. J. (1994). Correspondence analysis and its interpretation. En

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Greenacre, M. J. y Blasius, J., editores, *Correspondence Analysis in the Social Sciences: Recent developments and its applications*, pp. 3–22. Academic Press, San Diego, CA.

Greenacre, M. J. y Blasius, J. (2006). *Multiple Correspondence Analysis and Related Methods.* Chapman & Hall CRC, Boca Ratón, FL.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, 6(4):430–450.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21.

Haenlein, M. y Kaplan, A. M. (2004). A beginners guide to Partial Least Squares analysis. *Understanding Statistics*, 3(4):283–297.

Hair, J. F., Anderson, R. E., Tatham, R. L., y Black, W. C. (1998). *Multivariate Data Analysis.* Prentice Hall, Englewood Cliffs, NJ, 4ª edición.

Hair, J. F., Black, W. C., Babin, B. J., y Anderson, R. E. (2014a). *Multivariate Data Analysis.* Pearson, Harlow, UK, 7ª edición.

Hair, J. F., Hult, G., Ringle, C., y Sarstedt, M. (2014b). *A primer on Partial Least Squares Structural Equation Modeling (PLS-SEM).* Sage, Thousand Oaks: CA.

Hair, J. F., Ringle, C. M., y Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *Journal of Marketing Theory and Practice*, 19(2):139–152.

Hair, J. F., Ringle, C. M., y Sarstedt, M. (2012a). Partial least squares: The better approach to structural equation modeling? *Long Range Planning*, 45(5-6):312–319.

Hair, J. F., Ringle, C. M., y Sarstedt, M. (2013). Partial least squares structural equation modeling: Rigorous applications, better results and higher acceptance. *Long Range Planning*, 46(1-2):1–12.

Hair, J. F., Sarstedt, M., Ringle, C. M., y Mena, J. (2012b). An assessment of the use of partial least squares structural equation modeling in marketing research. *Journal of the Academy of Marketing Science*, 40(3):414–433.

Hakstian, A. R., Rogers, W. T., y Cattell, R. B. (1982). The behavior of number-of-factors rules with simulated data. *Multivariate Behavioral Research*, 17(2):193–219.

Halkidi, M., Batistakis, I., y Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2/3):107–145.

Hamilton, L. (1992). *Regressions with graphics: A second course in applied statistics.* Brooks/Cole, Monterey, CA.

Hand, D., Daly, F., Lunn, A., McConway, K., y Ostrowski, E. (1994). *A Handbook of Small Data Sets.* Chapman & Hall, Londres.

Hanley, J. A. y McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC). *Radiology*, 143(1):29–36.

Harman, H. (1976). *Modern factor analysis.* University of Chicago Press, Chicago, IL.

Hartigan, J. (1985). Statistical theory in clustering. *Journal of Classification*, 2(1):63–76.

Hartigan, J. y Wong, M. (1979). Kmeans clustering algorithm. *Applied Statis-*

Estadísticos e-Books & Papers

tics, 28(1):100–108.

Hatcher, L. (1994). *A Step by Step approach to using the SAS system for Factor Analysis and Structural Equation Modeling.* SAS Institute Inc., Cary, NC.

Hauck, W. W. y Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, 72(360a):851–853.

Hawkins, D. (1980). *Identification of outliers.* Springer, Londres.

Hayduk, L. A. y Littvay, L. (2012). Should researchers use single indicators, best indicators, or multiple indicators in structural models? *BMC Medical Research Methodology*, 12(159).

Haynes, S., Nelson, N., y Blaine, D. (1999). Psychometric issues in assessment research. En Kendall, P., Butcher, J., y Holmbeck, G., editores, *Handbook of research methods in clinical psychology*, pp. 125–154. John Wiley & Sons, Nueva York.

Haynes, S., Richard, D., y Kubany, E. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, 7(3):238–247.

Henseler, J., Hubona, G., y Ray, P. (2016). Using PLS path modeling in new technology research: Updated guidelines. *Industrial Management & Data Systems*, 116(1):2–20.

Henseler, J., Ringle, C. M., y Sarstedt, M. (2014). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, 43(1):115–135.

Henseler, J., Ringle, C. M., y Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing. *Advances in International Marketing*, 20:277–319.

Henseler, J. y Sarstedt, M. (2013). Goodness-of-fit indices for partial least squares path modelling. *Computational Statistics*, 28(2):565–580.

Henze, N. y Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Communications in Statistics, Theory and Methods*, 19(10):3595–3617.

Hlavac, M. (2015). stargazer: Well-formated regression and summary statistics tables. R package version 5.2. *http://CRAN.R-project.org/package=stargazer.*

Hoaglin, D. C. y Welsch, R. E. (1978). The hat matrix in regression and ANOVA. *The American Statistician*, 32(1):17–22.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–803.

Hoffman, D. L. y Franke, G. R. (1986). Correspondence analysis: Graphical representation of categorical data in marketing research. *Journal of Marketing Research*, 23(3):213–227.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75:383–386.

Hopkins, J. y Clay, P. (1963). Some empirical distributions of bivariante T2

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

and homoscedasticity criterion M under unequal variance and leptokurtosis. *Journal of the American Statistical Association*, 58(304):1048–1053.

Horn, J. (1965). A rationale and a test for the number of factors in factor analysis. *Psychometrika*, 30:179–185.

Hosmer, D. W. y Lemeshow, S. (1980). A goodness-of-fit test for multiple logistic regression model. *Communications in Statistics*, A10:1043–1069.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441.

Howell, D. (2006). *Statistical methods for psychology.* Thomson, Belmont, CA, 6ª edición.

Hu, L.-T. y Bentler, P. M. (1995). Evaluation model fit. En Hoyle, R. H., editor, *Structural Equation Modeling: Concepts, issues and applications*, pp. 76–99. Sage, Thousand Oaks: CA.

Hu, L.-T. y Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensivity to underparameterized model misspecification. *Psychological Methods*, 3(4):424–453.

Hu, L.-T. y Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1):1–55.

Hu, L.-T., Bentler, P. M., y Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, 11(2):351–362.

Huber, E. y Stephens, J. D. (1993). Political parties and public pensions: A quantitative analyisis. *Acta Sociologica*, 36(4):309–325.

Hubert, L. y Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

Ito, K. (1962). A comparison of the powers of two MANOVA tests. *Biometrika*, 49(3/4):455–462.

Jaccard, J. y Wan, C. (1989). *LISREL approaches to interaction effects in multiple regression.* Sage, Thousand Oaks, CA.

Jaccard, P. (1901). Etude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 37(547-549).

Jarque, C. y Bera, A. (1980). Efficient test for normality, homoscedasticity and serial independence of residuals. *Economic Letters*, 6(3):255–259.

Jarvis, C. B., MacKenzie, S. B., y Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research*, 30(2):199–218.

Johnson, D. E. (1998). *Applied multivariate methods for data analysts.* Brooks/Cole, Nueva York.

Johnson, R. A. y Wichern, D. W. (1998). *Applied multivariate statistical analysis.* Prentice Hall, Upper Saddle River, NJ.

Joreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, 32(4):443–482.

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika*, 57(2):239–251.

Estadísticos e-Books & Papers

Jöreskog, K. G. y Sorborm, D. (1996). *LISREL 8 user's reference guide.* Scientific Software, Chicago, IL.

Joreskog, K. G. y Wold, H. (1982). The ML and PLS techniques for modeling with latent variables: Historical and comparative aspects. En Joreskog, K. G. y Wold, H., editores, *Systems under indirect observation: Part I*, pp. 263–270. North Holland, Ámsterdam.

Judd, C. M., McClelland, G. H., y Ryan, C. S. (2009). *Data analysis: A model comparison approach.* Routledge, Nueva York, 2ª edición.

Kabacoff, R. I. (2015). *R in action. Data analysis and graphics with R.* Manning, Shelter Island, NY.

Kachigan, S. K. (1991). *Multivariate Statistical Analysis.* Radius Press, Nueva York, 2ª edición.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20(4):141–151.

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika*, 35(4):401–415.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, 39(1):31–36.

Kaiser, H. F. y Rice, J. (1974). Little Jiffy, Mark IV. *Educational and Psychological Measurement*, 34(1):111–117.

Karjaluoto, H., Lehto, H., Leppaniemi, M., y Jayawardhena, C. (2008). Exploring gender influence on customer's intention to engage permission-based mobile marketing. *Electronic Markets*, 18(3):242–259.

Kirk, R. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, 56(5):746–759.

Kruskal, J. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27.

Kruskal, J. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29(2):115–129.

Kruskal, J. y Wish, M. (1978). *Multidimensional Scaling.* Sage University Paper Series on Quantitative Applications in the Social Sciences 07-001. Sage, Beverly Hills, CA.

Kruskal, W. y Wallis, W. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621.

Lauter, J. (1978). Sample size requirements for the T2 test of MANOVA (tables for one-way classification). *Biometrical Journal*, 20(4):389–406.

Lawley, D. (1956). Tests of significance for the latent roots of covariance and correlation matrix. *Biometrika*, 43(1/2):128–136.

Lawley, D. y Maxwell, A. (1971). *Factor analysis as a statistical method.* American Elsevier Publishing Co., Nueva York, 2ª edición.

Lê, S., Josse, J., y Husson, F. (2008). FactoMinerR: An R package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18.

Lebart, L., Morineau, A., y Piron, M. (2000). *Statistique Exploratoire Multidimensionnelle.* Dunod, Paris.

Lebart, L., Morineau, A., y Warwick, K. M. (1984). *Multivariate descriptive*

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

statistical analysis: Correspondence analysis and related techniques for large matrices. John Wiley & Sons, Nueva York.

Lenth, R. (2001). Some practicals guidelines for effective sample size determination. *American Statistician*, 55(3):187–193.

Levene, H. (1960). Robust tests for equality of variances. En Olkin, I., Ghurye, S. B., Hoeffding, W., Madow, W. G., y Mann, H. B., editores, *Contributions to probability and statistics: Essays in honor of Harold Hotelling*, volumen 2 de *Stanford Studies in Mathematics and Statistics*, pp. 278–292. Stanford University Press, Stanford, CA.

Linn, R. (1968). A Monte Carlo approach to the number of factors problem. *Psychometrika*, 33(1):37–71.

Little, R. J. y Rubin, D. B. (1987). *Statistical analysis with missing data.* Wiley, Nueva York.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28:128–137.

Lohmoller, J. B. (1987). *LVPLS program manual, version 1.8.* Universitat Koln, Colonia, Alemania.

Lohmoller, J. B. (1989). *Latent variable path modeling with partial least squares.* Physica, Heidelberg, Alemania.

Lohnes, P. R. (1961). Test space and discriminant space classification models and related significance tests. *Educational and Psychological Measurement*, 21(3):559–574.

Long, J. (1983a). *Confirmatory Factor Analysis.* Número 07-033 en Quantitative Applications in the Social Sciences. Sage, Newbury Park, CA.

Long, J. (1983b). *Covariance Structure Models: An introduction to Lisrel.* Beverly Hills, CA.

Looney, S. W. y Gulledge, T. R. (1985). Use of correlation coefficient with normal probability plots. *The American Statistician*, 39(1):75–79.

Lunney, G. (1970). Using analysis of variance with dichotomous dependent variable: an empirical study. *Journal of Educational Measurement*, 7(4):263–269.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. En Le Cam, L. y Neyman, J., editores, *Proceedings of the Fifth Berkeley Symposium on mathematical statistics and probability*, volumen 1, pp. 281–297, Berkeley, CA. University of California Press.

Maddala, G. (1983). *Limited dependent and qualitative variables in Econometrics.* Cambridge University Press, Cambridge, MA.

Mardia, K. (1980). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3):519–530.

Marsaglia, G. y Marsaglia, J. (2004). Evaluating the Anderson-Darling distribution. *Journal of Statistical Software*, 9(2):1–5.

Marsh, H. W., Balla, J. R., y McDonald, R. P. (1988). Goodness-of-fit indices in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, 103(3):391–410.

McCallum, R., Roznowski, M., y Necowitz, L. (1982). Model modifications

Estadísticos e-Books & Papers

in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, 111:490–504.

McDonald, R. (1985). *Factor analysis and related techniques.* Lawrence Erlbaum, Hillsdale, NJ.

McDonald, R. P. (1982). A note on the investigation of local and global indentifiability. *Psychometrika*, 47(1):101–103.

McDonald, R. P. (1996). Path analysis with composite variables. *Multivariate Behavioral Research*, 31(2):239–270.

McFadden, D. (1979). Quantitative methods for analysing travel behavior of individuals: Some recent developments. En Hensher, D. y Stopher, P., editores, *Behavioural travel modelling*, pp. 279–318. Croom Helm, Londres.

Menard, S. (1995). *Applied logistic regression analysis.* Sage University Paper Series on Quantitative Applications in the Social Sciences 07-106. Sage, Thousand Oaks, CA.

Merisavo, M., Kajalo, S., Karjaluoto, H., Virtanen, V., Salmenkivi, S., Raulas, M., y Leppaniemi, M. (2007). An empirical study of the drivers of consumer acceptance of mobile advertising. *Journal of Interactive Advertising*, 7(2):41–50.

Milligan, G. y Cooper, M. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179.

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, 45(3):325–342.

Mulaik, S. (1972). *The foundation of factor analysis.* McGraw Hill, Nueva York.

Nagelkerke, E. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692.

Nenadic, O. y Greenacre, M. J. (2007). Correspondence analysis in R, with two and three-dimensional graphics: The ca package. *Journal of Statistical Software*, 20(3):1–13.

Netemeyer, R., Pullig, C., y Bearden, W. O. (2002). Observations on some key psychometric properties of paper-and-pencil measures. En Woodside, A. G. y Moore, E., editores, *Essays by distinguished marketing scholars of the Society for Marketing Advances*, pp. 115–138. Jai Press, Ámsterdam.

Newton, R. R. y Rudestam, K. E. (2013). *Your statistical consultant: Answers to your data analysis questions.* Sage, Thousand Oaks, CA, 2ª edición.

Norusis, M. J. (2008). *SPSS Statistics 17.0 Statistical Procedures Companion.* Prentice-Hall, Inc., Upper Saddle River, NJ.

Nunnally, J. y Bernstein, I. (1994). *Psychometric Theory.* McGraw-Hill, Nueva York, 3ª edición.

Ochiai, A. (1957). Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. *Bulletin of the Japanese Society of Scientific Fisheries*, 22:526–530.

Olson, C. L. (1974). Comparative robustness of six texts in multivariante analysis of variance. *Journal of the American Statistical Association*, 69(348):894–908.

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Olson, C. L. (1976). On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, 83(4):579–586.

Olson, C. L. (1979). Practical considerations in choosing a MANOVA test statistic: A rejoinder to stevens. *Psychological Bulletin*, 86(6):1350–1352.

O'Rourke, N. y Hatcher, L. (2013). *A Step by Step approach to using the SAS system for Factor Analysis and Structural Equation Modeling.* SAS Institute Inc., Cary, NC, 2ª edición.

Osborne, J. W. y Banjanovic, E. S. (2016). *Exploratory factor analysis with SAS.* SAS Institute Inc., Cary, NC.

Pedhazur, E. (1982). *Multiple regression in behavioural research.* Holt, Rinehart and Winston, Inc, Nueva York.

Pérez, C. (2004). *Técnicas de Análisis Multivariante de Datos.* Pearson Educación, Madrid.

Pillai, K. C. S. y Jayachandian, K. (1967). Power comparisons of tests of two multivariate hypotheses based on four criteria. *Biometrika*, 54(1/2):195–210.

Punj, G. y Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research*, 20(2):134–148.

Ramsay, J. (1977). Maximum likelihood estimation in multidimensional scaling. *Psychometrika*, 42(2):241–266.

Ramsay, J. (1982). Some statistical approaches to multidimensional scaling data. *Journal of the Royal Statistical Society. Series A (General)*, 145(3):285–312.

Rasmussen, J. L. (1988). Evaluating outlier identification tests: Mahalanobis D Squared and Comrey D. *Multivariate Behavioral Research*, 23(2):189–202.

Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research*, 32(4):329–353.

Raykov, T. (2001). Bias of Cronbach's alpha for mixed congeneric measures with correlated errors. *Applied Psychological Measurement*, 25(1):69–76.

Raykov, T. (2004). Behavioral scale reliability and measurement invariance evaluation using latent variable modelling. *Behavior Therapy*, 35:299–331.

Reinartz, W., Haenlein, M., y Henseler, J. (2009). An empirical comparison of the efficacy of covariance-based and variance-based SEM. *International Journal of Research in Marketing*, 26(4):332–344.

Ringle, C. M., Wende, S., y Becker, J.-M. (2015). *SmartPLS 3.0.* SmartPLS, Bönningstedt.

Ringle, C. M., Wende, S., y Will, A. (2005). *SmartPLS 2.0.* www.smartpls.de. [Computer Software], Hamburgo.

Robinson, J., Shaver, P., y Wrightsman, L. (1991). Criteria for scale selection and evaluation. En Robinson, J., Shaver, P., y Wrightsman, L., editores, *Measures of personality and social psychological attitudes*, pp. 1–15. Academic Press, San Diego, CA.

Rogers, D. J. y Tanimoto, T. T. (1960). An elementary mathematical theory of classificacion and prediction. *Science*, 132(3434):1115–1118.

Estadísticos e-Books & Papers

Ronsenberg, M. (1965). *Society and the adolescent self image.* Princeton University Press, Princeton, NJ.

Rossiter, J. R. (2002). The C-OAR-SE procedure for scale development in marketing. *International Journal of Research in Marketing*, 19(4):305–335.

Royston, J. (1982). An extension of Saphiro and Wilk's W test for normality to large samples. *Applied Statistics*, 31(2):115–124.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rummel, R. J. (1970). *Applied factor analysis.* Northwestern University Press, Evanston, IL.

Sarabia, F. J. y Sánchez, M. (1999). Validez y fiabilidad de escalas. En Sarabia, F. J., editor, *Metodología para la investigación en marketing y dirección de empresas.* Pirámide, Madrid.

Sarle, W. (1983). SAS technical report A-108, Cubic Clustering Criterion. Technical report, SAS Institute Inc, Cary, NC.

Satorra, A. y Bentler, P. M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. *Proceedings of the American Statistical Association*, pp. 308–313.

Scariano, S. M. y Davenport, J. M. (1987). The effects of violation of the independence assumptions in the one way ANOVA. *The American Statistician*, 41(2):123–129.

Schiffman, S. S., Reynolds, M. L., y Young, F. W. (1981). *Introduction to Multidimensional Scaling.* Academic Press, Orlando, FL.

Schumacker, R. y Lomax, R. (1996). *A beginner's guide to structural equation modeling.* Erlbaum, Mahwah,NJ, 1 edición.

Schwager, S. J. y Margolin, B. H. (1982). Detection of multivariate outliers. *The Annals of Statistics*, 10(3):943–954.

Shapiro, S. S. y Wilk, M. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

Sharma, S. (1996). *Applied Multivariate Techniques.* John Wiley & Sons, Hoboken, NJ, 1ª edición.

Shepard, R. N. (1962). The analysis of proximities: multidimensional scaling with an unknow distance function. *Psychometrika*, 27(2):125–140.

Shepard, R. N. (1972). A taxonomy of some principal types of data and of multidimensional methods for their analysis. En Shepard, R. N., Romney, A. K., y Nerlove, S. B., editores, *Multidimensional scaling: Theory and Application in Behavioral Sciencies*, volumen 1. Seminar Press, Nueva York.

Shimp, T. A. y Sharma, S. (1987). Consumer ethnocentrism: construction and validation of the CETSCALE. *Journal of Marketing Research*, 24(3):280–289.

Sokal, R. y Michener, C. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, 38:1409–1438.

Sokal, R. y Sneath, P. (1963). *Principles of numerical taxonomy.* W.H. Freeman, San Francisco, CA.

Sorbom, D. (1989). Model modification. *Psychometrika*, 54(3):371–384.

Sorenson, T. (1948). A method of establishing groups of equal amplitude in

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

plant sociology based on similarity of species content. *Kongelige Danske Videnskabernes Selskab, Biologiske Skrifter*, 5:1–34.

Spearman, C. (1904). A general intelligence objectivity objectively determined and measured. *American Journal of Psychology*, 15(2):201–293.

Stefanksy, W. (1971). Rejecting outliers by maximum normed residual. *The Annals of Mathematical Statistics*, 42(1):35–45.

Steiger, J. y Lind, J. (1980). Statistically based tests for the number of common factors. En *Proceedings of the Psychometric Society*, Iowa City, IA. Psychometric Society.

Stephens, M. (1974). EDF statistics for goodnes of fit and some comparison. *Journal of the American Statistical Association*, 69(347):730–737.

Stevens, J. P. (1979). Comment on Olson: Choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, 86(2):355–360.

Stevens, J. P. (1996). *Applied multivariate statistics for the social sciences.* Lawrence Erlbaum Associates, Mahwah, NJ.

Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences.* Routledge, Nueva York, 5 ª edición.

Stevens, S. (1946). On the theory of scales of measurement. *Science*, 103(2684):677–680.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147.

Strahan, R. y Gerbasi, K. C. (1972). Short, homogenous version of the Marlowe-Crowne social desirability scale. *Journal of Clinical Psychology*, 28(2):191–193.

Tabachnick, B. G. y Fidell, L. S. (2001). *Using multivariate statistics.* Allyn & Bacon, Boston, MA, 4ª edición.

Takane, Y. (1982). Maximum likelihood additivity analysis. *Psychometrika*, 17(3):225–241.

Takane, Y., Young, F. W., y De Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, 42(1):7–67.

Tanaka, J. (1987). How big is big enough? sample size and goodness of fit in structural equation models with latent variables. *Child Development*, 58(1):134–146.

Tatsuoka, M. (1971). *Multivariate analysis: Techniques for educational and psychological research.* Wiley, Nueva York.

Taylor, S. y Todd, P. (1995). Understanding information technology usage: a test of competing models. *Information Systems Research*, 6(2):144–176.

Teetor, P. (2011). *R Graphics Cookbook.* O'Reilly Media Inc, Sebastopol, CA.

Tenenhaus, M., Amato, S., y Vinzi, V. E. (2004). A global goodness-of-fit index for PLS structural equation modelling. En *Proceedings of the XLII SIS scientific meeting*, pp. 739–742.

Tenenhaus, M., Vinzi, V. E., Chatelin, Y. M., y Lauro, C. (2005). PLS path modeling. *Computational Statistics & Data Analysis*, 48(1):159–205.

Estadísticos e-Books & Papers

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* American Psychological Association, Washington, DC.

Tomarken, A. y Serlin, R. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin*, 99(1):90–99.

Toothaker, L. (1993). *Multiple comparison procedures.* Sage University Paper Series on Quantitative Applications in the Social Sciences 07-089. Sage, Newbury Park, CA.

Torgeson, W. S. (1952). Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419.

Trochim, W. M. (1989). An introduction to concept maping for planning and evaluation. *Evaluation and Program Planning*, 12(1):1–16.

Tsang, M., Ho, S.-C., y Liang, T.-P. (2004). Consumer attitudes toward mobile advertising: an empirical study. *International Journal of Electronic Commerce*, 8(3):65–78.

Tucker, L. R., Koopman, R., y Linn, R. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika*, 34(4):421–459.

Tucker, L. R. y Lewis, C. (1973). A reliability coefficient for maximum likelihood factors analysis. *Psychometrika*, 38(1):1–10.

Ullman, J. (2001). Structural equation modeling. En Tabachnick, B. y Fidell, L., editores, *Using multivariate statistics*, pp. 653–771. Allyn & Bacon, Boston, MA.

Velleman, P. F. y Wilkinson, L. (1993). Nominal, ordinal, interval and ratio typologies are misleading. *The American Statistician*, 47(1):65–72.

Ward, J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

Welch, B. (1951). On the comparison of several mean values: an alternative approach. *Biometrika*, 38(3/4):330–336.

Weller, S. C. y Romney, A. K. (1990). *Metric scaling: Correspondence analysis.* Sage University Paper Series on Quantitative Applications in the Social Sciences 07-075. Sage Publications.

Werts, C. E., Linn, R. L., y Joreskog, K. G. (1974). Intraclass reliability estimates: Testing structural equations assumptions. *Educational and Psychological Measurement*, 34(1):25–33.

Wheaton, B. (1987). Assessment of fit in overidentified models with latent variables. *Sociological Methods & Research*, 16(1):118–154.

Wheaton, B., Muthén, B. O., Alwin, D., y Summers, G. (1977). Assessing reliability and stability in panel models. En Heise, D., editor, *Sociological Methodology*, pp. 84–136. Jossey-Bass, San Francisco, CA.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838.

Widaman, K. (1993). Common factor analysis versus principal components analysis: Differential bias in representing model parameters? *Multivariate*

 © Ediciones Paraninfo

Estadísticos e-Books & Papers

Behavioral Research, 28(3):263–311.

Wilcox, R. (2003). Multiple comparison based on a modified one-step m-estimator. *Journal of Applied Statistics*, 30(10):1231–1241.

Wilcox, R. (2005). *Introduction to robust estimation and hypothesis testing.* Elsevier, Burlington, MA, 2 edición.

Wilk, M., Shapiro, S. S., y Chen, H. (1968). A comparative study of various tests of normality. *Journal of the American Statistical Association*, 63(324):1343–1372.

Wish, M. y Carroll, J. (1974). Applications of individual differences scaling to studies of human perception and judgement. En Carterette, E. y Friedman, M., editores, *Handbook of perception.* Seminar Press, Nueva York.

Wold, H. (1973). Non-linear partial least squares (NIPALS) modelling. some current developments. En Krishnaiah, P. R., editor, *Multivariate Analysis*, volumen III. Academic Press, Nueva York.

Wold, H. (1982). Soft modeling: the basic design and some extensions. En Wold, H. y Joreskog, K. G., editores, *Systems under indirect observation.*, volumen 2. North Holland, Ámsterdam.

Young, F. W. y Hamer, R. M. (1987). *Multidimensional scaling: History, Theory and Applications.* Lawrence Erlbaum Associates, Hillsdale, NJ.

Zimmerman, D. W. (1994). A note on the influence of outliers on parametric and nonparametric tests. *The Journal of General Psychology*, 121(4):391–401.

Estadísticos e-Books & Papers

Estadísticos e-Books & Papers

Estadísticos e-Books & Papers

Análisis multivariante aplicado con R

El principal objetivo de esta obra es proporcionar al lector una visión rigurosa y a la vez muy aplicada de las herramientas estadísticas de análisis multivariante. Las herramientas desarrolladas cubren un espectro muy amplio de lectores potenciales: desde **estudiantes de grado o máster** que preparan asignaturas de estadística, investigación de mercados o métodos cuantitativos aplicados a la economía, la dirección de empresas, la sociología o la psicología, hasta **investigadores** de esos mismos campos que desean estar al día de los últimos avances en modelos de ecuaciones estructurales o PLS-SEM. Todas estas herramientas se desarrollan utilizando el ***software* libre R**.

El enfoque del manual combina la rigurosidad con la aplicabilidad práctica a partir del desarrollo de **más de 40 casos** resueltos y multitud de ejemplos que permiten entender la lógica de la técnica de análisis de datos y cómo aplicarla fácilmente mediante R. Asimismo, la web del manual permite al usuario acceder a todas las **bases de datos** que soportan esos casos, así como a la **sintaxis** que permite su resolución mediante R.

Además de un capítulo dedicado a la preparación de los datos (análisis de valores perdidos, casos atípicos y comprobación de las propiedades de normalidad, homocedasticidad, linealidad e independencia de las observaciones), el resto de temas abordan el análisis de conglomerados, escalamiento multidimensional, análisis de correspondencias, análisis de la varianza, análisis multivariante de la varianza, regresión lineal múltiple, análisis discriminante, regresión logística, análisis de componentes principales, análisis factorial, análisis factorial confirmatorio, validación de los instrumentos de medida, modelos de ecuaciones estructurales y PLS-SEM.

Autores

Joaquín Aldás es catedrático de Comercialización e Investigación de Mercados en la Facultat d'Economia de la Universitat de València y profesor investigador del Instituto Valenciano de Investigaciones Económicas (Ivie). Su campo de especialización son los métodos cuantitativos de investigación en marketing.

Ezequiel Uriel es catedrático emérito de Fundamentos del Análisis Económico de la Universitat de València y profesor investigador del Instituto Valenciano de Investigaciones Económicas. Sus áreas de especialización son mercado de trabajo, sistemas de información estadística y técnicas de predicción.

ISBN: 978-84-283-2969-9

9 788428 329699

Paraninfo
www.paraninfo.es