Diseños de investigación en Psicología

Manuel Ato García
Guillermo Vallejo Seco



Descargas/material complementario
www.edicionespiramide.es

PIRÁMIDE

Diseños de investigación en Psicología

MANUEL ATO GARCÍA
CATEDRÁTICO DE METODOLOGÍA DE LAS CIENCIAS DEL COMPORTAMIENTO
EN LA UNIVERSIDAD DE MURCIA

GUILLERMO VALLEJO SECO
CATEDRÁTICO DEL ÁREA DE METODOLOGÍA DE LAS CIENCIAS
DEL COMPORTAMIENTO DE LA UNIVERSIDAD DE OVIEDO

Diseños de investigación en Psicología

EDICIONES PIRÁMIDE

COLECCIÓN «PSICOLOGÍA»

Director:
Francisco J. Labrador
Catedrático de Modificación de Conducta
de la Universidad Complutense de Madrid

Edición en versión digital

Composición y maquetación: Autores

Está prohibida la reproducción total o parcial de este libro electrónico, su transmisión, su descarga, su descompilación, su tratamiento informático, su almacenamiento o introducción en cualquier sistema de repositorio y recuperación, en cualquier forma o por cualquier medio, ya sea electrónico, mecánico, conocido o por inventar, sin el permiso expreso escrito de los titulares del copyright.

© Manuel Ato García y Guillermo Vallejo Seco, 2015
© Primera edición electrónica publicada por Ediciones Pirámide (Grupo Anaya, S. A.), 2015
Para cualquier información pueden dirigirse a piramide_legal@anaya.es
Juan Ignacio Luca de Tena, 15. 28027 Madrid
Teléfono: 91 393 89 89
www.edicionespiramide.es
ISBN digital: 978-84-368-3324-9

Índice

Prólogo **15**

1. El modelado estadístico **17**

1.1. Introducción . . . 17
1.2. El proceso de modelado estadístico . . . 18
1.2.1. Primera etapa: especificación . . . 20
1.2.2. Segunda etapa: ajuste . . . 23
1.2.3. Tercera etapa: evaluación . . . 26
1.2.4. Cuarta etapa: interpretación . . . 27
1.3. Ejemplos de modelado estadístico . . . 28
1.3.1. Modelo de regresión lineal . . . 29
1.3.2. Modelo de regresión de Poisson . . . 31
1.3.3. Modelo de regresión logística . . . 35
1.3.4. Modelo de regresión multinivel . . . 37
1.4. La estructura del modelo lineal clásico (*LM*) . . . 42
1.5. La estructura del modelo lineal generalizado (*GLM*) . . . 44
1.6. La estructura del modelo lineal mixto (*LMM*) . . . 47
1.7. La estructura del modelo de ecuaciones estructurales (*SEM*) . . . 49
1.8. Sugerencias de estudio . . . 49

2. Modelos de regresión **51**

2.1. Introducción . . . 51
2.2. Ejemplo 2.1 . . . 51
2.3. El modelo nulo . . . 55
2.4. Modelos de regresión simple . . . 56
2.4.1. Estimación de parámetros del modelo [X1] . . . 57
2.4.2. Análisis de varianza del modelo [X1] . . . 59
2.4.3. Ajuste global del modelo [X1] . . . 62
2.4.4. Ajuste condicional con el modelo [X1] . . . 64
2.5. Modelos de regresión múltiple . . . 65

©Ediciones Pirámide

2.5.1. Regresión múltiple con dos predictores . . . 65
A. Estimación de parámetros del modelo [X1,X2] . . . 66
B. Análisis de varianza del modelo [X1,X2] . . . 70
C. Ajuste global del modelo [X1,X2] . . . 71
D. Ajuste condicional con el modelo [X1,X2] . . . 72
2.5.2. Regresión múltiple con más de dos predictores . . . 73
2.6. Evaluación de un modelo de regresión . . . 79
2.7. El problema de la multicolinealidad . . . 84
2.8. Regresión con variables categóricas . . . 88
2.8.1. Codificación tipo regresión . . . 88
2.8.2. Codificación tipo ANOVA . . . 90
2.9. Selección de variables . . . 90
2.10. Modelos de mediación y moderación . . . 93
2.10.1. Modelos de mediación . . . 94
2.10.2. Modelos de moderación . . . 97
2.11. Sugerencias de estudio . . . 100

3. Modelos ANOVA con efectos fijos **103**

3.1. Introducción . . . 103
3.2. Modelos ANOVA de un factor . . . 105
3.2.1. La descomposición de la variación . . . 106
3.2.2. Ajuste global, ajuste condicional y magnitud del efecto . . . 109
3.2.3. Potencia y tamaño muestral . . . 111
3.2.4. Ejemplo 3.1 . . . 113
3.3. Diagnóstico del modelo ANOVA de un factor . . . 116
3.3.1. Supuestos del modelo . . . 116
3.3.2. Evaluación de los supuestos del modelo . . . 117
3.3.3. Los estadísticos de Welch y Brown-Forsythe . . . 119
3.3.4. Transformación de potencia para la variable de respuesta . . . 120
3.4. Análisis post-hoc . . . 121
3.4.1. Comparaciones múltiples entre medias de grupo . . . 121
3.4.2. El análisis de tendencias . . . 124
3.5. Modelos ANOVA con más de un factor . . . 128
3.5.1. Modelos ANOVA de dos factores . . . 129
3.5.2. El concepto de interacción . . . 132
3.5.3. Un ejemplo de ANOVA con dos factores fijos . . . 135
3.5.4. Generalización a modelos complejos . . . 141
3.5.5. Modelos ANOVA factorial con tres factores . . . 142
3.5.6. Ejemplo de ANOVA factorial con tres factores fijos . . . 145
3.6. Modelos ANOVA multivariante (MANOVA) . . . 149
3.7. Sugerencias de estudio . . . 153

©Ediciones Pirámide

4. Modelos ANOVA con efectos aleatorios y mixtos **155**

4.1. Introducción . . . 155
4.2. El modelo ANOVA de un factor con efectos aleatorios . . . 156
4.2.1. Diferencias entre el modelo I y el modelo II . . . 157
4.2.2. Inferencia estadística con el modelo II . . . 159
4.3. ANOVA de dos factores con modelos II y III . . . 161
4.3.1. Introducción . . . 161
4.3.2. Modelo ANOVA con efectos aleatorios . . . 162
4.3.3. Modelo ANOVA con efectos mixtos . . . 163
4.3.4. Valor esperado de las medias cuadráticas . . . 165
4.3.5. Un ejemplo de ANOVA con efectos aleatorios . . . 166
4.3.6. ANOVA con efectos mixtos . . . 172
4.4. Un algoritmo para estimar los valores esperados . . . 176
4.4.1. Introducción . . . 176
4.4.2. Construcción de las razones F en los modelos II y III . . . 181
4.4.3. La aproximación de Welch-Satterthwaite . . . 183
4.4.4. Un ejemplo con tres factores y efectos mixtos . . . 184
4.5. Sugerencias de estudio . . . 187

5. Modelos ANCOVA y regresión multinivel **189**

5.1. Introducción . . . 189
5.2. El modelo ANCOVA de un factor . . . 190
5.2.1. Submodelos derivados del modelo ANCOVA de un factor . . . 191
5.2.2. Supuestos del modelo ANCOVA . . . 193
5.2.3. El análisis de covarianza . . . 194
5.3. Aplicación del modelo ANCOVA . . . 196
5.3.1. Análisis descriptivo . . . 197
5.3.2. Análisis de covarianza mediante las etapas de modelado . . . 200
5.3.3. Análisis de covarianza mediante ajuste condicional . . . 202
5.3.4. Comparaciones entre medias del factor ajustadas . . . 203
5.3.5. Magnitud del efecto en modelos ANCOVA . . . 204
5.4. ANCOVA con pendientes no paralelas . . . 205
5.5. ANCOVA con dos covariantes . . . 210
5.6. Extensión de los modelos de regresión: la regresión multinivel . . . 213
5.6.1. El análisis tradicional y sus consecuencias . . . 215
5.6.2. El análisis multinivel . . . 217
5.6.3. Un ejemplo de análisis multinivel . . . 219
5.7. Sugerencias de estudio . . . 226

©Ediciones Pirámide

6. Introducción al diseño de investigación **227**

6.1. Conceptos básicos . . . 227
6.1.1. Los tres pilares de la investigación en Psicología . . . 227
6.1.2. Fases del proceso de investigación . . . 228
6.1.3. Estrategias para la investigación empírica en Psicología . . . 230
6.1.4. Unidad experimental, unidad de observación y unidad de muestreo . . . 231
6.1.5. La inferencia causal . . . 232
6.2. El experimento como modelo de investigación . . . 234
6.2.1. El problema del confundido . . . 236
6.2.2. Técnicas de control de variables extrañas . . . 238
6.3. Validez de la investigación . . . 242
6.3.1. Validez de la conclusión estadística . . . 243
6.3.2. Validez interna . . . 245
6.3.3. Validez de constructo . . . 247
6.3.4. Validez externa . . . 248
6.4. El diseño de experimentos . . . 249
6.4.1. Tradiciones de la investigación experimental en Psicología . . . 249
6.4.2. Las tres estructuras básicas del diseño experimental . . . 251
6.4.3. La estructura de los tratamientos . . . 251
6.4.4. La estructura del control . . . 252
6.4.5. La estructura del error . . . 254
6.4.6. La composición de un diseño de investigación . . . 255
6.5. Sugerencias de estudio . . . 256

7. Diseños intersujetos I: Aleatorización completa **259**

7.1. Diseños completamente aleatorios . . . 259
7.1.1. Algunas cuestiones sobre notación . . . 261
7.1.2. Ejemplo 7.1: DCA factorial equilibrado . . . 262
7.1.3. Ejemplo 7.2: DCA factorial no equilibrado . . . 269
7.2. Diseños jerárquicos . . . 274
7.2.1. Introducción . . . 274
7.2.2. Ejemplo 7.3: DJE con un factor y una variable anidada . . . 275
7.2.3. Ejemplo 7.4: DJE con dos factores y una variable anidada . . . 281
7.3. Ejemplos de diseños intersujetos con aleatorización completa . . . 286
7.4. Sugerencias de estudio . . . 288

8. Diseños intersujetos II: Reducción del error residual **289**

8.1. Introducción . . . 289
8.2. El diseño de bloques aleatorios . . . 289
8.2.1. Criterios de bloqueo . . . 292
8.2.2. Ventajas e inconvenientes del DBA . . . 292
8.3. El análisis estadístico del DBA . . . 293

©Ediciones Pirámide

8.3.1. *DBA* replicado . . . 296
8.3.2. *DBA* no replicado . . . 300
8.4. *DBA* y *DVC*: Dos alternativas de diseño . . . 305
8.4.1. Introducción . . . 305
8.4.2. Una comparativa entre el *DBA* y el *DVC* . . . 306
8.5. Diseño de cuadrado latino . . . 309
8.5.1. Introducción . . . 309
8.5.2. *DCL* no replicado (Ejemplo 8.4) . . . 310
8.5.3. Generalización del *DCL* . . . 315
8.5.4. *DCL* replicado con tres réplicas . . . 317
8.6. Ejemplos de diseños intersujetos con covariantes o variables de bloqueo . . . 320
8.7. Sugerencias de estudio . . . 321

9. Diseños de medidas repetidas (intrasujetos) 323

9.1. Introducción . . . 323
9.2. *DMR* con un factor intrasujeto . . . 327
9.3. Enfoques analíticos para el diseño de medidas repetidas con un factor intrasujetos . . . 330
9.3.1. Enfoque univariante clásico (ANOVA de un factor intrasujetos) . . . 330
9.3.2. Estructuras de covarianza y diseños intrasujetos . . . 334
9.3.3. El supuesto de esfericidad . . . 337
9.3.4. El enfoque multivariante (MANOVA) . . . 341
9.3.5. El enfoque mixto . . . 344
9.4. *DMR* con dos factores intrasujeto . . . 348
9.4.1. El enfoque univariante . . . 351
9.4.2. El enfoque multivariante . . . 354
9.4.3. El enfoque mixto . . . 356
9.5. El enfoque GEE . . . 359
9.6. Ejemplos publicados con *DMR* intrasujetos . . . 359
9.7. Sugerencias de estudio . . . 360

10. Diseños de medidas repetidas (mixtos) 361

10.1. Introducción . . . 361
10.2. *DMR* con un factor inter y un factor intra . . . 362
10.2.1. El enfoque univariante . . . 365
10.2.2. El enfoque multivariante . . . 369
10.2.3. El enfoque mixto y el modelado de estructuras de covarianza . . . 370
10.3. Generalización del *DMR* mixto . . . 379
10.4. *DMR* con dos factores inter y un factor intra . . . 380
10.4.1. El enfoque univariante . . . 383
10.4.2. El enfoque multivariante . . . 387
10.4.3. El enfoque mixto . . . 388
10.5. Diseños con estructura similar al *DMR* mixto . . . 392

©Ediciones Pirámide

10.5.1. El diseño de parcela dividida (DPD) . . . 392
10.5.2. El diseño conmutativo (DCO) . . . 395
10.5.3. El diseño doblemente multivariante . . . 399
10.6. Ejemplos seleccionados de la literatura . . . 399
10.7. Sugerencias de estudio . . . 401

11. Diseños cuasiexperimentales **403**

11.1. Introducción . . . 403
11.1.1. Tipología . . . 404
11.1.2. Notación . . . 405
11.2. Diseño con grupo control no equivalente . . . 406
11.2.1. Introducción . . . 406
11.2.2. Alternativas de análisis estadístico . . . 408
11.2.3. Variaciones del diseño con grupo control no equivalente . . . 415
11.2.4. Alternativas experimentales del diseño con grupo control no equivalente . . . 417
11.3. Diseño de discontinuidad de la regresión . . . 420
11.3.1. Introducción . . . 420
11.3.2. Análisis estadístico . . . 423
11.3.3. Supuestos del DDR . . . 425
11.3.4. Algunos ejemplos publicados con el DDR . . . 425
11.4. Diseños de series temporales interrumpidas . . . 426
11.4.1. Introducción . . . 426
11.4.2. Diseño de series temporales interrumpidas simple . . . 427
11.4.3. Diseño de series temporales interrumpidas con GC no equivalente . . . 437
11.4.4. Algunos ejemplos publicados con el $DSTI$ . . . 441
11.5. Sugerencias de estudio . . . 441

12. Diseños de caso único **443**

12.1. Introducción . . . 443
12.2. Características de los diseños de caso único . . . 445
12.3. Tipos de diseños de caso único . . . 446
12.3.1. Diseños intra-series de estructura simple . . . 447
12.3.2. Diseños reversibles intra-series . . . 448
12.3.3. Diseños no reversibles entre-series . . . 449
12.3.4. Diseños no reversibles de series combinadas . . . 450
12.4. Análisis de los datos en diseños experimentales de caso único . . . 453
12.4.1. Análisis visual de los resultados . . . 453
12.4.2. Análisis estadístico de los diseños de caso único . . . 457
12.4.3. Modelos de regresión en el análisis de series de tiempo . . . 458
12.4.4. Análisis multinivel de los diseños de caso único . . . 465
12.4.5. Las pruebas de aleatorización . . . 473
12.5. Consideraciones finales . . . 474

©Ediciones Pirámide

12.6. Sugerencias de estudio 474

13. Diseños no experimentales **475**

13.1. Introducción 475
13.2. Estudios comparativos 476
13.2.1. Diseños comparativos retrospectivos o *ex post facto* 477
13.2.2. Diseños comparativos transversales 481
13.2.3. Diseños prospectivos o longitudinales 484
13.2.4. Estudios evolutivos 490
13.3. Estudios predictivos y explicativos 493
13.3.1. Estudios predictivos 495
13.3.2. Estudios explicativos 497
13.5. Sugerencias de estudio 500

Referencias bibliográficas **501**

Anexo de tablas estadísticas **527**

©Ediciones Pirámide

Prólogo

Después de siete años de vigencia, intuimos que había llegado el momento de reeditar el texto original *Diseños Experimentales en Psicología* de 2007. La primera de las razones que nos animaron a tomar esta decisión tuvo su raíz en las sugerencias de cambio que muchos de nuestros alumnos y colegas de Universidad nos han hecho llegar a lo largo de estos años con el objeto de facilitar el trabajo con un texto que en general se ha considerado riguroso, pero extenso y difícil. La segunda razón fue nuestra convicción de que el material empleado en el original podía ser susceptible de mejora en este sentido para adaptarse a las demandas de estudiantes e investigadores.

Pero lo que en un principio se ideó como una segunda edición, fue transformándose lentamente durante el proceso de reedición hasta convertirse en un nuevo texto con una gran parte del contenido del original para el que creemos oportuno adoptar la etiqueta de *Diseños de investigación en Psicología*, una denominación que coincide con el de la asignatura que pretende cubrir en muchas de nuestras universidades, y particularmente en las Universidades de Murcia y Oviedo. En gran medida este cambio de orientación se ha debido a la reciente disponibilidad de un sistema de clasificación de los diseños de investigación (Ato, López y Benavente, 2013). Cabe no obstante advertir al lector que en este texto se tratan en detalle los diseños de la estrategia manipulativa (experimental, cuasiexperimental y caso único) y mucho más brevemente los diseños de la estrategia asociativa (comparativos, predictivos y explicativos), pero no se tratan los diseños de la estrategia descriptiva (de encuesta y observacionales). Sin embargo, esta decisión está bien justificada porque en muchas de nuestras Universidades se imparten cursos puntuales de metodología observacional y metodología de encuesta.

Como el original de 2007, *Diseños de investigación en Psicología* mantiene la misma filosofía de trabajo, el enfoque del modelado estadístico, que articula nuestra concepción de la aplicación del análisis estadístico en el marco del diseño de investigación. El primero de los capítulos del texto ofrece una visión general de las estructuras jerárquicas que fundamentan el enfoque del modelado estadístico para la investigación en las Ciencias del Comportamiento y muchas áreas de las Ciencias Sociales y las Ciencias de la Salud, desde la estructura básica del modelo lineal clásico, a la moderna estructura de los modelos de ecuaciones estructurales.

Los cuatro capítulos siguientes constituyen un curso intermedio sobre Análisis de Datos. Desde nuestro punto de vista representan en conjunto una herramienta imprescindible para poder abordar con rigor los fundamentos del diseño experimental y no experimental. Se tratan en detalle los modelos de regresión (capítulo 2), los modelos ANOVA y MANOVA de efectos fijos (capítulo 3), los modelos ANOVA con efectos aleatorios y mixtos (capítulo 4) y los modelos ANCOVA (capítulo 5). Buena parte del contenido de esta primera parte es materia que suele impartirse en un curso previo de Análisis de Datos, aunque en este texto se orienta desde el enfoque del modelado estadístico. Pero lo más sustancial

©Ediciones Pirámide

de esta primera parte no suele ser objeto de un curso introductorio de Análisis de Datos. Cuestiones tales como el problema de la multicolinealidad, los modelos de mediación y moderación, el tratamiento de los modelos ANOVA con efectos aleatorios y mixtos, el análisis multinivel, el análisis MANOVA y el análisis de covarianza son temas que requieren cierta familiarización por parte de los estudiantes antes de acometer los contenidos propios del diseño de investigación.

Los siguientes capítulos representan la parte esencial del texto, donde después de una breve introducción al diseño de investigación (capítulo 6) se tratan los fundamentos del diseño con un marcado énfasis en el diseño experimental (capítulos 7 a 10), el diseño cuasiexperimental (capítulo 11) y el diseño de caso único (capítulo 12), y con mucho menor énfasis los diseños no experimentales, y en concreto los diseños comparativos, predictivos y explicativos (capítulo 13).

Además del cambio de orientación que justifica una nueva denominación, son muchos los cambios que hemos introducido respecto del original de 2007. En lo referente a los contenidos teóricos, las modificaciones más sustanciales se han producido, sobre todo, en la separación de los efectos fijos y los efectos aleatorios y mixtos, la introducción del análisis multivariante de varianza y el análisis multinivel, la estimación de modelos ANCOVA con pendientes no paralelas, el tratamiento de los modelos de moderación y las técnicas de control de variables extrañas, pero también se tratan a nivel introductorio otras cuestiones de interés, tales como los valores perdidos, la potencia observada con efectos aleatorios y el diseño doblemente multivariante. Respecto a cuestiones no sustantivas, lo más destacable son los numerosos ejemplos publicados de aplicaciones de diseño que se incorporan en la segunda parte del texto y los cambios que se han producido en las salidas más complejas, particularmente la de los modelos multivariantes y mixtos. Presentamos una propuesta de tabla única como resumen de varias de las salidas típicas del programa SPSS para salidas del análisis multivariante del análisis multinivel. Otro cambio destacable es la utilización del programa *ggplot* del sistema *R* para la realización de los gráficos.

Con todo, el texto ha sufrido también un recorte importante en el número de páginas, del orden de un 25 % respecto del original de 2007. En gran medida este recorte se ha debido a que hemos prescindido de los apéndices, glosario y sintaxis de *SPSS* y se ha ampliado la anchura de la caja respecto del texto original. Aunque el material del texto de 2007 puede seguir siendo utilizado, ponemos a disposición del lector, bien en la página web de Pirámide, en la URL dedicada al texto, el acceso a todos los materiales de sintaxis y a los ejercicios prácticos que desarrollamos para nuestros alumnos, así como algunos ejemplos comentados de investigaciones publicadas.

Este texto no se hubiera planteado en sus actuales coordenadas sin las sugerencias que nos han hecho llegar alumnos e investigadores que han dedicado su esfuerzo a comprender y aplicar los contenidos del texto de 2007. La demanda de atención y concentración que exige escribirlo no habría sido en absoluto posible de no haber sido por la estoica resistencia de nuestras respectivas esposas, María Ángeles y María Teresa, a prescindir de nosotros durante todo un semestre, incluyendo fechas tan señaladas como las Navidades de 2014.

Manuel Ato y Guillermo Vallejo, enero 2015.

©Ediciones Pirámide

El modelado estadístico 1

1.1. INTRODUCCIÓN

Las observaciones realizadas por investigadores de las Ciencias del Comportamiento contienen información crucial acerca del comportamiento humano y su entorno social. Recuperar esta información puede ser difícil, sobre todo si las observaciones son muy numerosas o muy complejas, pero resulta más simple cuando los expertos tienen alguna idea preconcebida sobre la información presente en los datos, o sea, acerca de su estructura. Las ideas que los investigadores tienen acerca de la estructura de los datos se denominan MODELOS. El concepto de modelo se utiliza aquí como una forma de simplificar la realidad empírica, en el mismo sentido que se emplea una maqueta a escala de una vivienda o de un automóvil, con el objeto de aproximar con la mayor precisión posible la realidad empírica que el modelo (o la maqueta) trata de capturar. Los modelos pueden clasificarse como DETERMINÍSTICOS, donde las entradas y salidas del sistema se definen con precisión, sin variación aleatoria, como sucede con muchas leyes de la Física y la Ingeniería, y modelos PROBABILÍSTICOS, donde las entradas y salidas del sistema exhiben cierta variación aleatoria, como sucede con casi todos los modelos que se emplean en Psicología, en la encrucijada entre las Ciencias de la Salud y las Ciencias Sociales.

Para tratar la variación aleatoria presente en los modelos probabilísticos se utiliza una herramienta básica, la INFERENCIA ESTADÍSTICA. El enfoque tradicional con el que se realiza la inferencia estadística en Psicología y disciplinas afines se fundamenta en el contraste de hipótesis estadísticas mediante pruebas de significación (a veces complementada con la estimación de intervalos de confianza y de tamaños del efecto) siguiendo el modo de proceder típico de la estadística clásica o fisheriana. Por esta razón este enfoque se denomina ENFOQUE DEL CONTRASTE DE HIPÓTESIS (o abreviadamente, ENFOQUE NHST, de Null Hypothesis Significance Testing). Muchas de las críticas que ha recibido este enfoque (véase Balluerka, Gómez e Hidalgo, 2005; Cohen 1994; Cumming, 2013; Judd, McClelland y Culhane, 1995; Kline, 2004; Krantz, 1999; Nickerson, 2000 y Thompson, 2002b, entre otros) argumentan que el objeto de las pruebas de significación es diferente del objeto de la inferencia científica y que se requiere complementar y ampliar el horizonte de las pruebas de significación. Lo que realmente interesa es: "dados unos datos empíricos, ¿cuál es la probabilidad de la hipótesis?", pero lo que las pruebas de significación plantean es: "asumiendo que la hipótesis nula es cierta, ¿cuál es la probabilidad de los datos empíricos?". Obviamente, la hipótesis nula no suele ser precisamente la hipótesis más informativa que puede someterse a prueba.

©Ediciones Pirámide

Un enfoque complementario al enfoque NHST se perfiló durante las décadas de los años ochenta y noventa, representando una revolución pacífica y silenciosa contra el enfoque NHST. Este nuevo enfoque se basa en el ajuste y comparación de modelos probabilísticos, y se conoce como ENFOQUE DE LA COMPARACIÓN DE MODELOS. Rodgers (2010) ha descrito su historia y sus características, en la que han tenido un protagonismo esencial, entre otros, los textos de Judd, McClelland y Ryan (2008, 1ª edición 1989) y Maxwell y Delaney (2004, 1ª edición 1990).

El eje central de este nuevo enfoque es un procedimiento integrado y coherente denominado MODELADO ESTADÍSTICO (*statistical modelling*) que, en el caso más simple, consiste en la aplicación de una serie de procesos con el objeto de conseguir una explicación apropiada del comportamiento de una *variable de respuesta* (los DATOS) a partir de una función ponderada de una o más *variables explicativas* (el MODELO). Como en todo modelo probabilístico, la explicación en general no suele ser perfecta, y por tanto es preciso evaluar la discrepancia existente entre los Datos y el Modelo (el ERROR o RESIDUAL) en términos de alguna medida diferencial al uso. El resultado es la ECUACIÓN GENERAL DEL MODELADO ESTADÍSTICO, que se expresa en los términos siguientes:

$$\begin{array}{ccccccc} Y_i & = & \mu_i & + & e_i & & \\ \boxed{\text{DATOS}} & = & \boxed{\text{MODELO}} & + & \boxed{\text{ERROR}} & & \end{array} \qquad (1.1)$$

La ecuación 1.1 establece que los datos empíricos contenidos en un vector de VALORES OBSERVADOS (Y_i), correspondiente a una variable de respuesta, son una función de un vector de VALORES ESTIMADOS (μ_i), que constituye el *componente sistemático*, representado por algún modelo teórico apropiado, y un vector RESIDUAL, que representa la discrepancia existente entre los valores observados y los valores estimados por el modelo ($e_i = Y_i - \mu_i$), y constituye el *componente aleatorio*.

En este capítulo se presenta de forma simplificada el proceso general de modelado estadístico y se introducen las estructuras básicas de los modelos más representativos de la investigación en psicología, algunos de los cuales serán objeto de estudio intensivo en los siguientes capítulos del texto. Más información sobre el proceso de modelado estadístico puede consultarse en Krzanowski (1998).

1.2. EL PROCESO DE MODELADO ESTADÍSTICO

El objeto fundamental del modelado estadístico es la búsqueda y detección del modelo más parsimonioso que sea capaz de explicar una variable de respuesta con el mínimo error posible. La secuencia de aplicación del proceso de modelado estadístico se resume en el diagrama de la figura 1.1 y contiene una serie de etapas claramente diferenciadas, al término de las cuales el investigador debe disponer de información suficiente para determinar si un modelo es aceptable o no como predicción o explicación de la/s variable/s de respuesta.

En primer lugar, es importante distinguir entre las formas más comunes que suele presentar la variable de respuesta. En el cuadro 1.1 se distinguen cinco grandes tipos de variables de respuesta, que representan un amplio espectro de opciones, junto con las distribuciones de probabilidad que asumen, las restricciones de uso que plantean y algunos ejemplos.

©Ediciones Pirámide

CUADRO 1.1
Tipos de variables de respuesta

Variable de respuesta	Distribución	Restricciones	Ejemplos comunes
Numérica	Normal	Ninguna	Puntuación global en un test
Recuento	Poisson	Valores no negativos	Nº errores cometidos en una tarea
Proporción/Tasa	Binomial	Escala de probabilidad	Proporción de aciertos
Tasa	Poisson	Escala de probabilidad	Tasa de errores por minuto
Ordinal	Multinomial	Categorías de la escala	Ítem formulado en escala Likert
Binaria	Bernouilli	Respuesta dicotómica	Ítem en escala binaria (V/F)

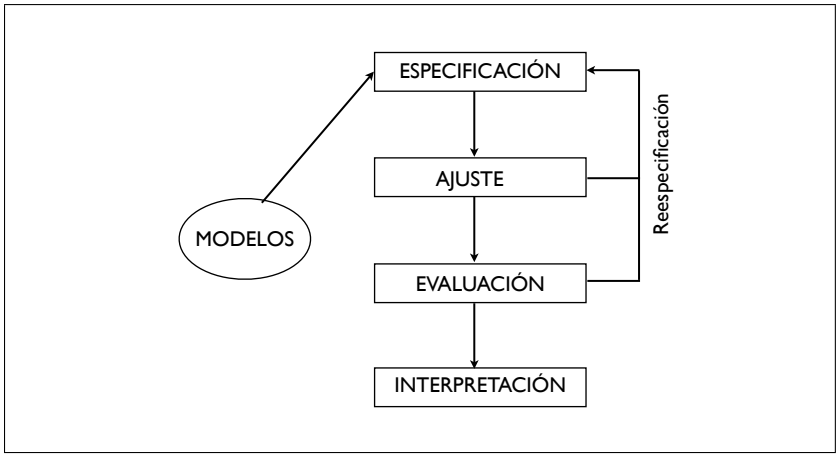

Figura 1.1. – El proceso de modelado estadístico.

El uso habitual de los modelos en el proceso de modelado estadístico requiere conocer la terminología apropiada para distinguirlos. Por orden de complejidad, puede distinguirse un MODELO PERFECTO o SATURADO (MS), que se caracteriza porque tiene tantos parámetros como observaciones, y por tanto reproduce los datos empíricos exactamente, sin ninguna simplificación, ya que valores observados y valores estimados son exactamente iguales. En este sentido, el modelo saturado es el modelo más complejo y menos parsimonioso de todos los modelos, pero en general no todos los parámetros del modelo saturado son de interés. Por el contrario, el modelo más simple (o más parsimonioso) posible es el modelo que no incluye ninguna variable explicativa y por ello se denomina MODELO NULO O BASAL ($M0$). No obstante, este modelo tiene un parámetro, que representa el valor medio común a todas las observaciones, pero usualmente no es representativo para explicar los datos empíricos. El modelo que concentra todos los parámetros de interés del investigador, excluyendo los demás, se denomina MODELO MÁXIMO. Entre el modelo máximo y el modelo nulo se encuentran muchos posibles MODELOS DE TRABAJO que pueden tener interés potencial para el investigador. El objetivo del modelado es encontrar un modelo aceptable de entre todos los modelos de trabajo posibles. En general, un modelo se considera *óptimo* cuando es parsimonioso (es decir, no incluye parámetros irrelevantes) y obtiene un ajuste aceptable (es decir, su discrepancia respecto del modelo saturado o del modelo nulo es mínima).

Cuando se comparan modelos, usualmente se utilizan dos modelos alternativos: un MODELO RESTRINGIDO, que contiene un número básico de parámetros, y un MODELO AMPLIADO, que contiene uno o más parámetros adicionales respecto del modelo restringido, con el objeto de determinar cuál de los dos obtiene mejor ajuste.

©Ediciones Pirámide

1.2.1. Primera etapa: especificación

En la primera etapa del proceso, el investigador, basándose en su propia intuición, en el conocimiento sustantivo de la temática objeto de investigación o en resultados derivados de modelos anteriores sometidos a prueba, *especifica* un modelo teórico de su interés. Especificar un modelo teórico consiste en proponer una combinación particular de un conjunto de p predictores o variables explicativas X_j, con valores X_{ij}, y de un conjunto de q parámetros β_j (para $j = 1,\ldots,q$), que puede incluir una constante opcional (β_0), con el objeto de pronosticar los valores Y_i de la variable de respuesta Y, para la que se asume una distribución determinada. Los dos ingredientes básicos de la especificación del modelo son los componentes fijo y aleatorio.

1. El COMPONENTE FIJO O SISTEMÁTICO se propone identificar las variables explicativas más relevantes con el objeto de predecir o explicar óptimamente el comportamiento de una variable de respuesta y consiste en formular una ecuación con las variables X_j y sus respectivos parámetros β_j. Los parámetros son pesos diferenciales que representan la capacidad potencial de una variable explicativa para predecir o explicar una variable de respuesta. El compuesto resultante de combinar para cada uno de los i casos los valores X_{ij} y sus parámetros β_j es un vector denominado PREDICTOR LINEAL, que se simboliza como η_i, y se define como

$$\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} \tag{1.2}$$

   El vector η_i asume que la función que relaciona las variables explicativas y los parámetros es lineal, la función más frecuentemente encontrada en investigación aplicada. Afortunadamente, algunas funciones que no tienen naturaleza estrictamente lineal pueden ser fácilmente transformadas en funciones lineales y otras muchas funciones pueden ser tratadas como intrínsecamente lineales. En escasas aplicaciones de la psicología, sin embargo, se requiere utilizar modelos intrínsecamente no lineales (véase Bates y Watts, 2008).

2. El COMPONENTE ALEATORIO O DE ERROR se propone capturar la variación aleatoria individual mediante la selección de una distribución de probabilidad para la variable de respuesta, cuyos parámetros básicos, el valor esperado, $E(Y_i)$, y la varianza, $V(Y_i)$, juegan un papel primordial. En el modelo lineal clásico, el valor esperado es igual a la media de la variable de respuesta,

$$E(Y_i) = \mu_i \tag{1.3}$$

   y su VARIANZA es una función que se define como el producto

$$V(Y_i) = \phi V(\mu_i) \tag{1.4}$$

©Ediciones Pirámide

donde ϕ es el PARÁMETRO DE ESCALA O DISPERSIÓN de la distribución de probabilidad y $V(\mu_i)$ es un término que permite determinar la relación existente entre la media y la varianza, una relación que en general depende de la distribución de probabilidad seleccionada.

Puesto que el vector de valores estimados μ_i y el vector del predictor lineal η_i no necesariamente se miden en la misma escala, es preciso definir alguna transformación de la escala del predictor lineal a la escala del valor esperado, que se requiere para utilizar la ecuación 1.1. Tal vínculo se denomina FUNCIÓN DE ENLACE y se formula como

$$\eta_i = g(\mu_i) \tag{1.5}$$

siendo su función inversa

$$\mu_i = g^{-1}(\eta_i) \tag{1.6}$$

Cuando el vector del predictor lineal y el vector de valores estimados son iguales, entonces no hay necesidad de transformación. Formalmente, sin embargo, se dice que la transformación utilizada es la FUNCIÓN DE ENLACE IDENTIDAD,

$$\eta_i = g(\mu_i) = \mu_i \tag{1.7}$$

En otras ocasiones, los vectores η_i y μ_i no coinciden y entonces se requiere definir una función de enlace. Por ejemplo, en modelos que emplean variables de recuento como variables de respuesta y asumen una distribución de Poisson, la transformación entre predictor lineal y valor esperado suele ser la FUNCIÓN DE ENLACE LOGARÍTMICO,

$$\begin{aligned} \eta_i &= g(\mu_i) = \log(\mu_i) \\ \mu_i &= g^{-1}(\eta_i) = \exp(\eta_i) \end{aligned} \tag{1.8}$$

mientras que, para modelos que emplean variables categóricas binarias o dicotómicas como variables de respuesta y que asumen una distribución binomial, la transformación entre predictor lineal y valor esperado suele ser la FUNCIÓN DE ENLACE LOGIT,

$$\begin{aligned} \eta_i &= g(\mu_i) = \text{logit}\,(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) \\ \mu_i &= g^{-1}(\eta_i) = \text{logit}^{-1}(\eta_i) = \left[\frac{\exp(\eta_i)}{1+\exp(\eta_i)}\right] = \left(\frac{1}{1+\exp^{-\eta_i}}\right) \end{aligned} \tag{1.9}$$

©Ediciones Pirámide

La figura 1.2 es una representación jerárquica de las estructuras básicas que fundamentan el enfoque del modelado estadístico para la investigación en Psicología. Asumiendo cualquier distribución de la FAMILIA EXPONENCIAL DE DISTRIBUCIONES, una extensa familia que incluye las distribuciones normal, binomial, multinomial y Poisson como casos particulares, y una o más variable/s de respuesta de cualquier naturaleza (véase cuadro 1.1), esta representación distingue cinco estructuras de modelado.

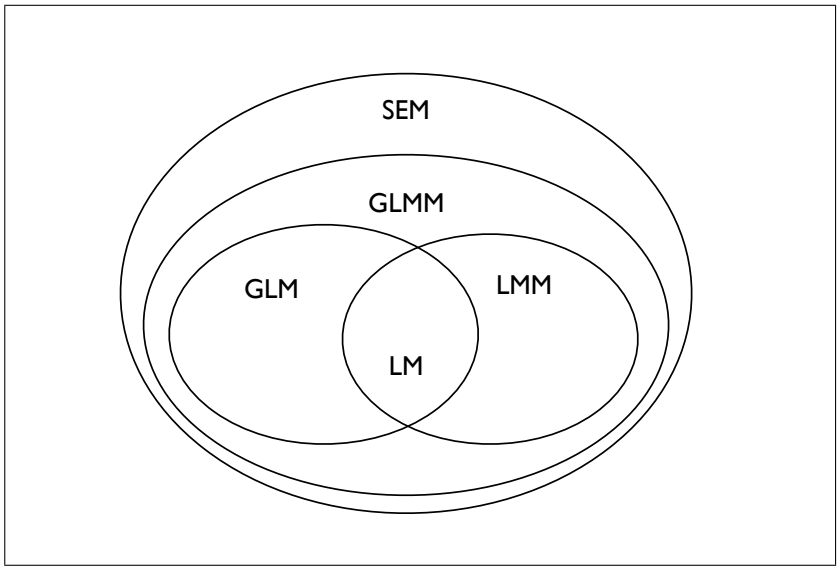

Figura 1.2. – Representación jerárquica de las estructuras básicas del modelado estadístico.

— La estructura del MODELO LINEAL CLÁSICO (o estructura *LM*, *Linear Model*), es la más básica y sus modelos se caracterizan por utilizar una variable de respuesta numérica que se asume distribuida normalmente. Los principales modelos que incluye esta estructura son los MODELOS DE REGRESIÓN, que emplean como componentes fijos/sistemáticos predictores de naturaleza numérica, los MODELOS ANOVA, que utilizan predictores categóricos, y los MODELOS ANCOVA, que usan una mezcla de predictores numéricos y categóricos. Estos tres modelos tienen en común el empleo de un único componente aleatorio. Se utilizan profusamente en la investigación psicológica y serán objeto de tratamiento en los capítulos 2 a 4.

— La estructura del MODELO LINEAL GENERALIZADO (o estructura *GLM*, *Generalized Linear Model*), se caracteriza por incluir cualquier variable de respuesta (numérica o categórica) y una distribución de la familia exponencial (véase Olsson, 2002). Aunque un caso particular de la estructura *GLM* es la estructura *LM*, que asume una distribución normal, dada su importancia en la investigación en Psicología en este texto se trata de forma separada. Todos los modelos derivados de la estructura *GLM* asumen también la existencia de un componente aleatorio único.

— La estructura del MODELO LINEAL MIXTO (o estructura *LMM*, *Linear Mixed Model*) se diferencia de la estructura *GLM* porque sus modelos permiten incluir más de un componente aleatorio, pero, como los modelos *LM*, asumen también una variable de respuesta numérica distribuida normalmente. La estructura *LM* puede ser también considerada como un caso particular de la estructura *LMM* caracterizada por incluir un único componente aleatorio.

©Ediciones Pirámide

— La estructura del MODELO LINEAL MIXTO GENERALIZADO (o estructura *GLMM*, *Generalized Linear Mixed Model*) engloba las dos estructuras básicas, *GLM* y *LMM*, permitiendo especificar modelos con dos o más componentes aleatorios, una variable de respuesta de cualquier naturaleza (numérica o categórica) y cualquier distribución de la familia exponencial (véase McCulloch, Searle y Neuhaus, 2008). Dos características peculiares de esta estructura, y de todas las subestructuras que incluye, es que sólo admite una variable de respuesta y que tanto las variables explicativas como las variables de respuesta deben ser variables observables.

— La estructura del MODELO DE ECUACIONES ESTRUCTURALES (o estructura *SEM*, *Structural Equation Model*) generaliza los modelos *GLMM* permitiendo incluir más de una variable de respuesta (y no sólo una) en un sistema de ecuaciones con variables tanto observables (manifiestas) como no observables (latentes). No es posible utilizar ecuaciones múltiples en estructuras jerárquicamente inferiores (véase Kaplan, 2009; Kline, 2011).

La gran mayoría de los modelos estadísticos que se emplean en la investigación psicológica pertenecen a alguna de las estructuras citadas. Es esencial que el investigador que trabaja con modelos estadísticos conozca bien el repertorio de estructuras de que dispone y las propiedades de los modelos que incluyen tales estructuras con el objeto de utilizarlos e interpretarlos adecuadamente.

1.2.2. Segunda etapa: ajuste

Una vez formulado un modelo estadístico, en la segunda etapa del proceso el investigador procede primero a estimar sus parámetros con un procedimiento de estimación apropiado.

— Con modelos de la estructura *LM*, que asumen una distribución normal y una respuesta numérica, el método usual de estimación de parámetros es MÍNIMOS CUADRADOS (*OLS*), que persigue como criterio objetivo minimizar la suma de los cuadrados de las diferencias entre valores observados y valores estimados, o sea, la suma de los cuadrados de los errores residuales.

— Con otros modelos de la estructura *GLM* que no asumen distribución normal ni utilizan una variable de respuesta necesariamente numérica, y también en los modelos *GLMM* y *SEM*, el método de estimación más común es MÁXIMA VEROSIMILITUD (*ML*), que persigue como criterio objetivo maximizar la FUNCIÓN DE VEROSIMILITUD, una función de las diferencias entre los valores estimados de un modelo respecto de los valores observados. Es importante notar que los objetivos de ambos métodos (*OLS* y *MV*) son similares en su finalidad, ya que el resultado de minimizar las diferencias entre valores observados y esperados es en gran medida equivalente al de maximizar una función de las diferencias entre valores observados y estimados.

Una vez estimados los parámetros de un modelo, se obtiene el correspondiente predictor lineal y después se aplica la correspondiente función de enlace para obtener el vector de valores estimados. Finalmente, se calcula una medida de la *discrepancia* entre los datos (los valores observados) y el modelo propuesto para explicarlos (los valores estimados) con la finalidad de determinar el ajuste del modelo a los datos empíricos. La medida más común de discrepancia es la DESVIANZA RESIDUAL (D_R, o simplemente D), definida como $D = -2LL$, donde LL (*LogLikehood*) es el logaritmo de (natural) de la función de verosimilitud maximizada del modelo propuesto. En general, cuanto mayor sea la desvianza peor será el ajuste del modelo. En el extremo, cuando $LL = 0$ se trata de un modelo perfecto o saturado. Toda

©Ediciones Pirámide

medida de desvianza se asocia con unos grados de libertad, que son la diferencia entre el número de observaciones (N) y el número de parámetros (q) que requiere la especificación del modelo.

Hay dos procedimientos generales de ajuste de modelos: *ajuste global* y *ajuste condicional*. Partiendo del principio de que un modelo mal ajustado no es un modelo interpretable, el AJUSTE GLOBAL se propone con el objeto de determinar si un modelo individual es una aproximación aceptable de los datos empíricos. Dado un modelo de trabajo $M1$ correctamente especificado, y siendo MS el modelo saturado o perfecto y $M0$ el modelo nulo o basal, el AJUSTE GLOBAL ABSOLUTO indaga si el modelo $M1$ incluye un número suficiente de parámetros intentando responder a la cuestión: ¿difiere $M1$ del modelo saturado MS? Por el contrario, el AJUSTE GLOBAL RELATIVO indaga si los parámetros que incluye el modelo son en su conjunto necesarios intentando responder a la cuestión: ¿difiere $M1$ del modelo nulo $M0$? No obstante, conviene tener presente que éste es todavía un tema controvertido y que hay sensibles diferencias en la forma de evaluar el ajuste global entre las diferentes estructuras de modelado.

— Si $M1$ es un modelo de la estructura LM, el proceso de ajuste implica una descomposición de la variación de la variable de respuesta (la Suma de Cuadrados Total, SC_T) en una parte explicada por el modelo (la Suma de Cuadrados del Modelo, SC_M) y otra parte no explicada (la Suma de Cuadrados Residual, SC_R), según la ecuación 1.1. La desvianza residual del modelo $M1$ es la SC Residual, $D(M1) = SC_R$, con grados de libertad residuales $gl_R = N - q$, donde N es el número de observaciones y q es el número de parámetros del modelo. El AJUSTE GLOBAL ABSOLUTO no se puede valorar para esta estructura. El AJUSTE GLOBAL RELATIVO compara la desvianza del modelo $M1$ con la desvianza del modelo nulo $M0$. El modelo nulo no tiene variables explicativas, pero emplea la misma variable de respuesta y su desvianza es $D(M0) = SC_T$, con grados de libertad $gl_T = N-1$, porque sólo requiere estimar un parámetro. El resultado del ajuste se evalúa utilizando la PRUEBA DE LA RAZÓN DE VARIANZAS con la distribución F a un nivel de significación establecido. Cuando el resultado de la prueba es significativo, el modelo se considera bien ajustado; cuando no es significativo, el modelo se considera inaceptable o incompatible con los datos empíricos.

— Si $M1$ es un modelo de la estructura GLM o LMM, el proceso de estimación obtiene el logaritmo (natural) de la función de verosimilitud maximizada (LL), y su desvianza es $D(M1) = -2LL(M1)$, con grados de libertad $gl(M1) = N - q$. El AJUSTE GLOBAL ABSOLUTO compara la desvianza del modelo $M1$ con la del modelo saturado, o sea, del modelo que tiene de desvianza $D(MS) = 0$ y grados de libertad $gl(MS) = 0$. Para ello se emplea la PRUEBA DE LA RAZÓN DE VEROSIMILITUD, $L^2(M1) = D(M1) - D(MS) = D(M1)$, con grados de libertad $gl(M1) = N - q$, o sea, simplemente la desvianza residual con sus grados de libertad. El valor de L^2 se distribuye según χ^2 y tiene excelentes propiedades asintóticas (el cociente L^2/gl tiende a uno cuando el ajuste es óptimo), de tal forma que cuando el resultado no es significativo el modelo se considera bien ajustado, mientras que cuando es significativo, se considera inaceptable o incompatible con los datos empíricos. Pero en afinidad con los modelos LM, el AJUSTE GLOBAL RELATIVO implica una descomposición de la desvianza total (D_T, con $gl_T = N-1$) en un componente explicado por el modelo (D_M, con $gl_M = q-1$) y un componente no explicado (D_R, con $gl_R = N-q$). El proceso compara la desvianza del modelo $M1$ respecto de la desvianza del modelo nulo [$D(M1) - D(M0) = D_R - D_T = D_M$] utilizando también la prueba de la razón de verosimilitud, con grados de libertad igual a la diferencia entre los grados de libertad de ambos modelos. En este caso, el ajuste se considera aceptable cuando L^2 es significativo, y no aceptable cuando L^2 no es significativo.

©Ediciones Pirámide

Una vez determinado si el modelo se ajusta o no, el investigador puede cuestionarse si los componentes especificados en un modelo son o no significativos y proceder, si cabe, a su simplificación:

— Con modelos de la estructura *LM*, este proceso tiene lugar utilizando contraste de hipótesis sobre los componentes del modelo con la PRUEBA t DE STUDENT o también con la razón F equivalente.

— Con modelos de la estructura *GLM* y *LMM*, el proceso se realiza usando contraste de hipótesis con la PRUEBA t DE STUDENT y la distribución t (o también en ocasiones con la prueba Z y la distribución normal) para los efectos fijos, y con la PRUEBA Z DE WALD con la distribución χ^2 equivalente para los efectos aleatorios.

El ajuste global permite determinar si un modelo particular se ajusta aceptablemente o no, o sea si es adecuado o inadecuado para explicar los datos observados. Pero esta conclusión no implica que sea un modelo óptimo, ni que sea el mejor modelo de todos los posibles. Pueden existir otros muchos modelos que se ajusten adecuadamente. En estos casos el investigador tiene que recurrir a la comparación de dos (o más) modelos alternativos para determinar cuál de los dos ajusta mejor.

La comparación de dos modelos solamente puede realizarse cuando se trata de MODELOS ANIDADOS, es decir, cuando uno de los modelos sea parte de otro, del que difiere en el número de parámetros. El proceso de ajuste que se emplea en estas ocasiones se denomina AJUSTE CONDICIONAL y permite concluir si un modelo más simple, el *modelo restringido*, porque cuenta con menor número de parámetros, representa una desviación significativa de un modelo más complejo, el *modelo ampliado*, porque cuenta con mayor número de parámetros. La comparación entre los dos modelos recurre también a la PRUEBA DE LA RAZÓN DE VEROSIMILITUD, que se interpreta en función del resultado de la inferencia estadística en los términos siguientes: si hay diferencias significativas entre los dos modelos, entonces el modelo más complejo (modelo ampliado) será considerado mejor que el modelo más simple (modelo restringido); en cambio, si no hay diferencias entre los dos modelos entonces el modelo restringido se considera mejor que el modelo ampliado.

Puesto que el ajuste condicional mediante la comparación de modelos es un proceso tedioso, puede servir de ayuda el empleo de algún CRITERIO DE INFORMACIÓN, como el CRITERIO AIC (*Akaike Information Criterion*) de Akaike o el CRITERIO BIC (*Bayesian Information Criterion*) de Schwartz. Ambos criterios constan de dos términos: uno mide la bondad de ajuste (desvianza residual, D) y el otro penaliza el valor del ajuste en función de la complejidad del modelo. Estos criterios se interpretan con el formato de *mejor cuanto más pequeño*, y se definen (Burnham y Anderson, 2002) mediante

$$AIC = D + 2q \tag{1.10}$$

$$BIC = D + q\log(N) \tag{1.11}$$

donde D es la desvianza residual del modelo, $\log(N)$ el logaritmo natural del número de casos u observaciones y q es el número de parámetros empleado por el modelo. En general, dados dos modelos bien ajustados, suele ser mejor aquel que tiene el más bajo índice BIC o AIC. Aunque su aplicación no requiere que los modelos sean anidados, es importante recalcar que estos criterios son sólo aproximaciones y no deben nunca emplearse en lugar de la prueba formal de ajuste condicional, que es la que permite la decisión en última instancia.

©Ediciones Pirámide

1.2.3. Tercera etapa: evaluación

En la tercera etapa del proceso, y una vez aceptado un modelo como mejor representación de los datos observados, debe decidirse si tal modelo es válido practicando una evaluación concienzuda de sus propiedades. Los modelos derivados del *LM* plantean ciertas condiciones sobre la estructura de los datos, los denominados SUPUESTOS DE LA ESTRUCTURA *LM*, que con bastante frecuencia no se cumplen en la práctica.

1. NORMALIDAD de la distribución de la variable de respuesta.
2. HOMOCEDASTICIDAD, HOMOGENEIDAD DE LAS VARIANZAS o VARIANZA CONSTANTE de la variable de respuesta para cada uno de los valores del/los predictor/es.
3. INDEPENDENCIA de las observaciones de la variable de respuesta.
4. LINEALIDAD, o relación lineal entre variable de respuesta y cada una de las variables explicativas.

Existen muchos procedimientos estadísticos y gráficos para evaluar el cumplimiento de los supuestos. Sin embargo, resulta también interesante conocer las razones que han podido provocar tales problemas. Por lo general suelen ser errores que tuvieron lugar durante el proceso de recogida de datos o durante el proceso de especificación de modelos. Los más comunes son los siguientes:

— ERROR DE MEDIDA en la variable de respuesta y/o en las variables explicativas. Las medidas con baja fiabilidad son más propensas a presentar este tipo de error.

— ERROR DE ESPECIFICACIÓN, que ocurre cuando se especifica un modelo que incumple alguno de los supuestos que debe satisfacer. La consecuencia directa del error de especificación es alguna forma de SESGO, que por lo general produce una sobreestimación o subestimación de la relación entre variables. El error de especificación puede tomar varias formas posibles dependiendo de que sea afectado el componente sistemático, el componente aleatorio o la función de enlace del modelo o se registren observaciones extremas.

  - El ERROR DE ESPECIFICACIÓN DEL COMPONENTE SISTEMÁTICO suele presentarse cuando:
    - ⋆ se utiliza una forma funcional inapropiada para la relación entre variables explicativas y respuesta;
    - ⋆ se omite alguna variable explicativa importante;
    - ⋆ se incluye alguna variable explicativa irrelevante;
    - ⋆ se selecciona un rango inapropiado de valores para la/s variable/s explicativas.
  - El ERROR DE ESPECIFICACIÓN DEL COMPONENTE ALEATORIO ocurre con frecuencia cuando se asume una distribución incorrecta para la variable de respuesta.
  - El ERROR DE ESPECIFICACIÓN DE LA FUNCIÓN DE ENLACE ocurre con modelos *GLM* y *GLMM* cuando se selecciona una función de enlace inadecuada para transformar los valores del predictor lineal en valores ajustados.

— La presencia de OBSERVACIONES EXTREMAS ocurre cuando se registran OBSERVACIONES ATÍPICAS (*outliers*), o sea, puntuaciones que se desvían de la nube de puntos del diagrama de dispersión pero no tienen influencia en el ajuste, y/u OBSERVACIONES INFLUYENTES (*influentials*), que son puntuaciones atípicas que producen cambios drásticos en la estimación de los parámetros y tienen una influencia determinante en el ajuste del modelo.

©Ediciones Pirámide

Afortunadamente existen sofisticados procedimientos para detectar y corregir la mayoría de estos errores. Algunos de ellos pueden corregirse de forma simple mediante alguna transformación directa de la variable de respuesta. La *familia de transformaciones Box-Cox* permite seleccionar una transformación apropiada para corregir violaciones de la normalidad y de la homogeneidad de las varianzas. En el caso de que se viole uno o más de los supuestos anteriores y no sea posible ninguna transformación, los PROCEDIMIENTOS NO PARAMÉTRICOS O DE DISTRIBUCIÓN LIBRE pueden ser una alternativa interesante (véase Corder y Foreman, 2009 ó Nussbaum, 2015).

1.2.4. Cuarta etapa: interpretación

Finalmente, si el modelo no presenta problemas que recomienden su rechazo, en la cuarta etapa del proceso el investigador puede interpretar el modelo, comprobando previamente que cumple con los tres criterios básicos que articulan todo el proceso de modelado, a saber:

— El criterio lógico es la PARSIMONIA: el modelo aceptado debe ser suficientemente simple para describir con relativa aproximación el complejo mecanismo subyacente que produjo la respuesta. Los modelos simples aportan además la ventaja de que tienen un mayor grado de falsabilidad.

— El criterio estadístico es la BONDAD DE AJUSTE: para que resulte óptimo, el modelo aceptado debe diferir significativamente del modelo nulo (modelos *LM*) o no desviarse demasiado del modelo saturado (modelos *GLM*) y no debe haber ningún otro modelo más apropiado con igual, o menor, número de parámetros (ajuste condicional).

— El criterio sustantivo es la INTEGRACIÓN TEÓRICA: el modelo aceptado debe tener significado desde el punto de vista de la teoría que lo generó y ser fácilmente integrable en su red conceptual.

Finalmente queda a juicio del investigador la adopción de alguna de tres decisiones posibles:

— Si el ajuste no es adecuado, o en su evaluación se detectan problemas importantes, el investigador debería rechazarlo definitivamente y retornar al comienzo del proceso en busca de un modelo más conveniente.

— Si el ajuste es apropiado, pero el modelo es complejo, lo acepta temporalmente y retorna al comienzo del proceso en busca de un modelo más parsimonioso.

— Si el ajuste es apropiado, y el modelo es parsimonioso, lo acepta de forma tentativa y lo interpreta e integra con la teoría sustantiva desde la que se generó.

En contraposición con el enfoque clásico del contraste de hipótesis, que considera solamente la aceptación o el rechazo de una determinada hipótesis nula, el enfoque de la comparación de modelos proporciona en su lugar el VALOR DE PROBABILIDAD asociado con la hipótesis, es decir, la probabilidad de observar un valor tan extremo o más extremo en la dirección del rechazo como el valor observado. Este valor de probabilidad puede también conceptualizarse como el nivel en el que la hipótesis sería rechazada y, por tanto, nos permite aplicar los criterios anteriormente citados proporcionando más información que el simple rechazo o aceptación de una hipótesis.

©Ediciones Pirámide

1.3. EJEMPLOS DE MODELADO ESTADÍSTICO

Para ilustrar el enfoque del modelado estadístico, y particularmente el proceso de ajuste, con alguno de los modelos más representativos tratados en este texto, el cuadro 1.2 presenta los datos empíricos correspondientes a una muestra de 20 jubilados (en adelante, Ejemplo 1.1) en los que se pretende determinar si las puntuaciones en un subtest de WAIS (que denotaremos como *wais* o variable X, con valores X_i) pueden servir para predecir ciertos comportamientos de interés, que se toman como variable de respuesta (variable Y, con valores Y_i). Utilizamos tres variables de respuesta diferentes para ejemplificar los modelos: el deterioro cognitivo (*dcog* o variable Y1, con valores Y_{i1}), la presencia o no de síntomas seniles (*ssen* o variable Y2, con valores Y_{i2}) y el número de errores cometido en una tarea cognitiva (*nerr* o variable Y3, con valores Y_{i3}).

CUADRO 1.2
Datos empíricos del ejemplo 1.1

Caso	wais X_i	dcog Y_{i1}	ssen Y_{i2}	nerr Y_{i3}
1	9	4	0	0
2	11	6	0	1
3	12	6	0	0
4	8	5	0	0
5	10	6	0	2
6	7	3	0	0
7	9	6	0	1
8	11	7	0	3
9	10	5	0	5
10	8	6	0	0
11	14	5	1	2
12	10	6	1	0
13	11	7	1	1
14	16	9	1	0
15	12	8	1	0
16	17	6	1	2
17	14	10	1	4
18	15	7	1	4
19	13	8	1	3
20	12	9	1	5

— La variable *dcog* ($Y1$) es una variable numérica que se distribuye de forma aproximadamente normal (con valores que se mueven libremente en el rango -∞/+∞ y por tanto pueden ser positivos o negativos). Para esta variable utilizaremos un MODELO DE REGRESIÓN LINEAL.

— La variable *ssen* ($Y2$) es una variable binaria o dicotómica que sigue una distribución binomial (solamente puede tomar valores 0-1), para la cual emplearemos un MODELO DE REGRESIÓN LOGÍSTICA.

— La variable *nerr* ($Y3$) es una variable de recuento que sigue una distribución de Poisson (con valores en el rango 0/+∞, y por tanto todos los valores son positivos). Para esta variable emplearemos un MODELO DE REGRESIÓN DE POISSON.

©Ediciones Pirámide

1.3.1. Modelo de regresión lineal

Ilustramos en primer lugar el proceso de ajuste especificando un MODELO DE REGRESIÓN LINEAL, un modelo de la estructura *LM*. Como variable de respuesta empleamos la variable *dcog*, que asumimos que se distribuye normalmente, y como predictor la variable *wais*.

La salida 1.1 presenta los estimadores por mínimos cuadrados ordinarios del modelo de regresión lineal resultante. La ECUACIÓN PREDICTIVA representa los estimadores de los parámetros en forma de ecuación: $\hat{\mu}_i = 2.209 + 0.370 X_i$, donde $\hat{\beta}_0 = 2.209$ es la CONSTANTE o INTERCEPCIÓN, que se define como el valor de *dcog* cuando *wais* = 0, y $\hat{\beta}_1 = 0.370$ es la PENDIENTE de la recta de regresión que se representa, junto con las bandas de confianza del 95 %, en la figura 1.3.

SALIDA 1.1

Estimadores de los parámetros del modelo de regresión lineal

Variable dependiente: *dcog*

		Error			Límites IC 95 %	
Parámetro	***B***	**típico**	***t***	***P*** > |***t***|	**Inferior**	**Superior**
Intersección	2.209	1.416	1.560	.136	-0.767	5.184
wais	0.370	0.120	3.075	.006	0.117	0.623

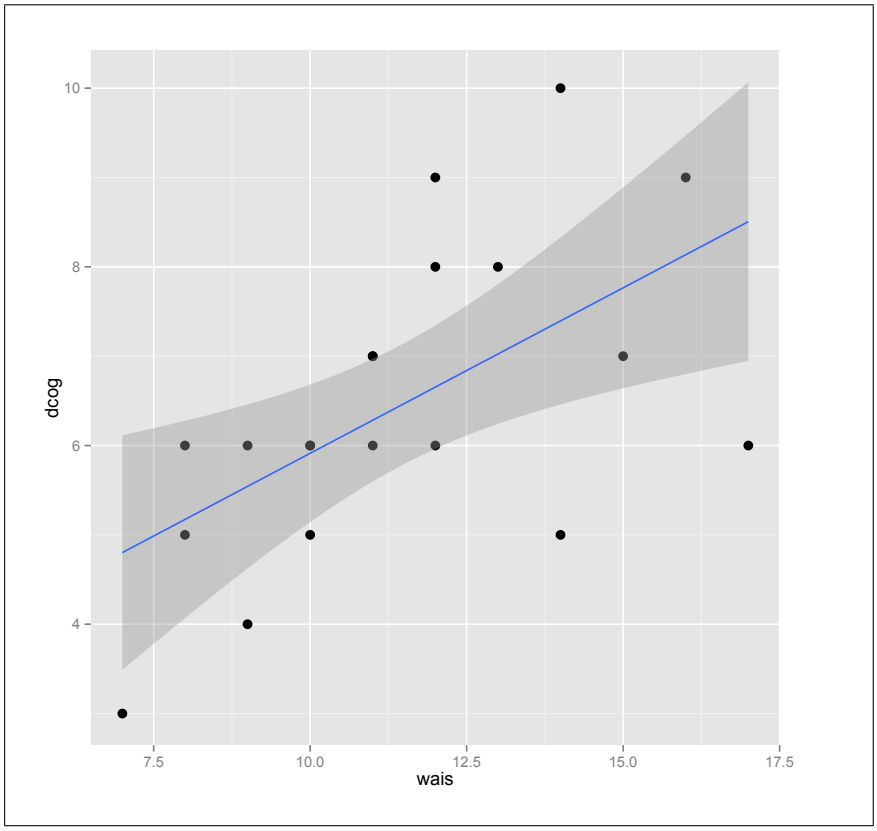

Figura 1.3. – Bandas de confianza del 95 % en torno a la función de regresión lineal.

©Ediciones Pirámide

El cuadro 1.3 resume de forma detallada, además de los vectores del predictor y de la respuesta, el vector de la media (o sea, el valor de la respuesta cuando el predictor es igual a cero), el vector de valores estimados y las tres fuentes de variación de la ecuación general del modelado: $SC_T = SC_M + SC_R$. El proceso de cálculo para cada fuente de variación se detalla en la parte final del cuadro 1.3.

CUADRO 1.3

Detalle del proceso de ajuste del modelo de regresión lineal clásico

	wais	**Observado**	**Media**	**Estimado**	**Fuentes de variación**			**Grados de**
Caso	X_i	Y_i	$\hat{\mu}$	$\hat{\mu}_i$	$(Y_i - \hat{\mu})$	$(\hat{\mu}_i - \hat{\mu})$	$(Y_i - \hat{\mu}_i)$	**libertad**
1	9	4	6.45	5.54	−2.45	−0.91	−1.54	1
2	11	6	6.45	6.28	−0.45	−0.17	−0.28	1
3	12	6	6.45	6.65	−0.45	0.20	−0.65	1
4	8	5	6.45	5.17	−1.45	−1.28	−0.17	1
5	10	6	6.45	5.91	−0.45	−0.54	0.09	1
6	7	3	6.45	4.80	−3.45	−1.65	−1.80	1
7	9	6	6.45	5.54	−0.45	−0.91	0.46	1
8	11	7	6.45	6.28	0.55	−0.17	0.72	1
9	10	5	6.45	5.91	−1.45	−0.54	−0.91	1
10	8	6	6.45	5.17	−0.45	−1.28	0.83	1
11	14	5	6.45	7.39	−1.45	0.94	−2.39	1
12	10	6	6.45	5.91	−0.45	−0.54	0.09	1
13	11	7	6.45	6.28	0.55	−0.17	0.72	1
14	16	9	6.45	8.14	2.55	1.69	0.86	1
15	12	8	6.45	6.65	1.55	0.20	1.35	1
16	17	6	6.45	8.51	−0.45	2.06	−2.51	1
17	14	10	6.45	7.39	3.55	0.94	2.61	1
18	15	7	6.45	7.76	0.55	1.31	−0.76	1
19	13	8	6.45	7.02	1.55	0.57	0.98	1
20	12	9	6.45	6.65	2.55	0.20	2.35	1
[1]: *SC* media			832.050					1
[2]: *SC* v. observados		889.000						20
[3]: *SC* v. estimados				851.663				2
		SC Total: [2] − [1]			56.950			19
		SC Modelo: [3] − [1]				19.613		1
		SC Residual [2] − [3]					37.337	18

La salida 1.2 muestra la descomposición de la variación del modelo de regresión lineal con el predictor *wais* para explicar la variable de respuesta *dcog*, que define la variación total a explicar (*SC* Total = 56.950) y asigna una parte al componente explicado (*SC* Modelo= 19.613) y el resto al componente no explicado (*SC* Residual= 37.337). El único componente con el que se ha especificado el modelo se representa también en la salida y aporta la misma información que el modelo. El contraste de hipótesis con la razón *F* del ANOVA permite concluir que el modelo (o similarmente, el predictor *wais*) es estadísticamente significativo, y por tanto difiere del modelo nulo: $F(1, 18) = 9.455; P = .007$. Recuérdese que *F* es una razón entre medias cuadráticas: $F = 19.613/2.074 = 9.455$. Y puesto que $R^2 = SC_M/SC_R$, una prueba *F* equivalente para valorar el ajuste global relativo utiliza el coeficiente de determinación R^2 para indagar si R^2 difiere significativamente de cero. En consecuencia, el ajuste global relativo del modelo de regresión lineal se considera aceptable.

©Ediciones Pirámide

SALIDA 1.2
Análisis de varianza del modelo de regresión lineal

Variable dependiente: *dcog*

Fuentes	***SC* tipo III**	***gl***	***MC***	***F***	***P* > *F***
Modelo	19.613	1	19.613	9.455[a]	.007
wais	19.613	1	19.613	9.455[a]	.007
Residual	37.337	18	2.074		
Total	56.950	19	(2.997)		

[a] R cuadrado = 0.344, R cuadrado ajustado = 0.308

Cuando hay varios modelos bien ajustados, el ajuste condicional compara dos o más modelos para determinar cuál de ellos es mejor. Puesto que para un modelo con un predictor no hay más que dos modelos posibles, el modelo con el predictor *wais* (modelo ampliado) y el modelo nulo (modelo reducido), el ajuste condicional con el modelo de regresión lineal proporciona también el mismo contraste de hipótesis que se obtuvo con el ajuste global. La salida 1.3 resume no obstante el procedimiento de ajuste condicional comparando las SC_R y los grados de libertad del modelo de regresión lineal y del modelo nulo. Observe que el resultado del ajuste condicional es similar al de la salida 1.2 para el ajuste global, es decir, $F(1,18) = 9.455; P = .007$. Pero en este caso el resultado no se valora solamente en términos de significación estadística, sino que concluye con la selección de uno de los dos modelos que se comparan. Puesto que hay diferencias significativas entre los dos modelos, la decisión más prudente es seleccionar el modelo ampliado y descartar el modelo reducido.

SALIDA 1.3
Ajuste condicional de modelos

Modelo 1: *dcog* = 1 (nulo)
Modelo 2: *dcog* = 1 + *wais* (regresión lineal simple)

Modelo	SC_R	gl_R	ΔSC_R	Δgl_R	***F***	***P* > *F***
1	56.950	19				
2	37.337	18	19.613	1	9.455	0.007

1.3.2. Modelo de regresión de Poisson

Algunos detalles de la salida de modelos de la estructura *GLM* difieren de los de la estructura *LM*, aunque el concepto del ajuste es el mismo para todos los modelos. Para ilustrar algunas de las diferencias utilizamos primero un modelo de regresión de Poisson. Como variable de respuesta utilizamos ahora *nerr* (véase cuadro 1.2), una variable de recuento con valores limitados dentro del rango $0/+\infty$ que contabiliza el número de errores cometidos por cada uno de los participantes en una tarea cognitiva, y como predictor *wais*. Para esta variable de respuesta no suele servir la distribución normal. La distribución de Poisson con enlace logarítmico se adapta bien en casi todos los casos. El modelo resultante es por esta razón un MODELO DE REGRESIÓN DE POISSON.

La salida 1.4 muestra los estimadores de los parámetros del modelo de regresión de Poisson utilizando máxima verosimilitud (*ML*). La interpretación de la salida 1.4 es muy similar a la de la salida 1.1 del modelo de regresión lineal, aunque observe que se recurre a la distribución normal estándar en lugar de la distribución t de Student. La ecuación predictiva es $\hat{\mu}_i = -1.102 + 0.134X_i$, donde μ_i son los valores esperados que se obtienen aplicando la función inversa de la función de enlace: $\mu_i = \exp(\eta_i)$. La constante o intercepción ($\hat{\beta}_0 = -1.102$) es, en unidades logarítmicas, el valor de *nerr* cuando *wais* es cero. La

©Ediciones Pirámide

pendiente de la función de regresión muestra un incremento positivo de $\hat{\beta}_1 = 0.134$ unidades logarítmicas en la variable de respuesta por unidad de aumento en el predictor. La figura 1.4 presenta la función de regresión resultante con bandas de confianza del 95 %, que aparecen sombreadas en gris y demuestran que el intervalo de confianza no es simétrico y que la predicción se deteriora considerablemente con valores de *wais* >14 puntos.

SALIDA 1.4

Estimadores de los parámetros del modelo de regresión de Poisson

Variable dependiente: *nerr*

		Error			Límites IC 95 %	
Parámetro	**B**	**típico**	***z***	***P*** > |***z***|	**Inferior**	**Superior**
Intersección	−1.102	0.801	−1.375	0.169	−2.673	0.468
wais	0.134	0.063	2.136	0.033	0.011	0.257

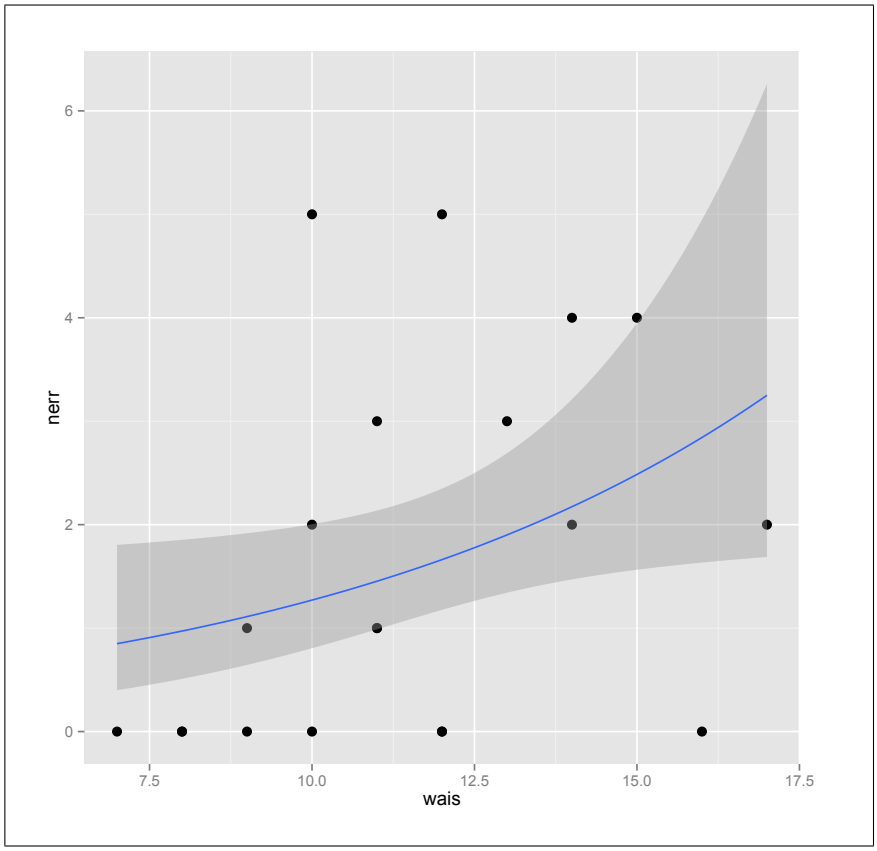

Figura 1.4. – Bandas de confianza del 95 % en torno a la función de regresión de Poisson.

La salida 1.5 muestra la descomposición de la variación del modelo de regresión de Poisson. Algunos aspectos de esta descomposición que conviene detallar son los siguientes:

©Ediciones Pirámide

— En primer lugar, observe que lo que se descompone en este modelo no son sumas de cuadrados (como sucede en modelos de la estructura *LM*) sino componentes de la desvianza, por cuya razón se utiliza el término ANÁLISIS DE DESVIANZA (en lugar de análisis de varianza), un concepto específico de modelos que utilizan estimación por máxima verosimilitud. Sin embargo, el argumento conceptual es similar. La DESVIANZA TOTAL es la desvianza residual del modelo nulo, y representa la variación que es objeto de explicación, $D_T = 42.820$, con 19 grados de libertad. La DESVIANZA RESIDUAL es la porción de la desvianza total que no explica el modelo, $D_R = 38.318$, con 18 grados de libertad, y es la desvianza que aporta el proceso de ajuste: $D_R = -2LL$. La diferencia entre los dos es la DESVIANZA DEL MODELO, $D_M = 42.819 - 38.318 = 4.502$, con $19 - 18 = 1$ grados de libertad.

SALIDA 1.5

Análisis de desvianza del modelo de regresión de Poisson

Variable dependiente: *nerr*

Fuentes	**Desvianza**	gl	χ^2	$P > \chi^2$
Modelo	4.502	1	4.502	.034
Residual	38.318	18		
Total	42.820	19		

— En segundo lugar, el ajuste global absoluto del modelo de regresión de Poisson se valora comparando su desvianza residual con la desvianza del modelo saturado (o sea, del modelo que tiene desvianza y grados de libertad igual a 0). La regla general para considerar un modelo como globalmente ajustado en este caso es que su desvianza (residual) no difiera mucho de sus grados de libertad (residuales). La prueba de la razón de verosimilitud se evalúa con la distribución χ^2. Siendo $D_R = 38.318$ y $gl_R = 18$, la prueba empírica es $\chi^2(18) = 38.318; P = 0.004$. El ajuste global absoluto se considera aceptable cuando el valor de P no sea significativo, y no aceptable cuando sea significativo. El resultado obtenido apunta que el ajuste absoluto del modelo de regresión de Poisson no es satisfactorio, y sugiere por tanto que los predictores utilizados no son suficientes.

— En tercer lugar, el ajuste global relativo cuestiona si el modelo propuesto difiere significativamente del modelo nulo. Esta valoración es similar a la que hemos tratado para modelos *LM*, pero se realiza con la prueba de la razón de verosimilitud y la distribución χ^2. Para ello se emplea la desvianza del modelo con sus grados de libertad. En concreto, $\chi^2(1) = 4.502; P = .034$ apunta que el ajuste relativo es aceptable y que el predictor *wais* es un buen predictor de la variable de respuesta.

— Finalmente, el ajuste global concierne a un conjunto de predictores. Es necesario comprobar individualmente el efecto de cada uno de los predictores. Puesto que no hay más que un predictor en este ejemplo, este paso es realmente innecesario. Las dos pruebas más comunes para esta finalidad son la prueba z, que asume una distribución normal y se muestra en la salida 1.4, $P(z > |2.136|) = .033$, y la prueba de Wald, que asume una distribución χ^2 y es esencialmente el cuadrado del valor de z de la salida 1.4, con un grado de libertad, $\chi^2(1) = 2.136^2 \approx 4.564; P = .033$. Observe que, aunque no producen exactamente el mismo resultado, con ambas pruebas se puede concluir que el predictor *wais* es un predictor aceptable para explicar la variable de respuesta.

El ajuste condicional para un modelo con un predictor es también innecesario, porque sólo hay dos modelos posibles a valorar (modelo nulo y modelo de regresión de Poisson). El resultado es similar al de

©Ediciones Pirámide

la salida 1.5, pero conviene resaltar algunas diferencias respecto de los modelos de la estructura *LM*, en particular la utilización de la distribución χ^2 en lugar de la distribución *F*. La salida 1.6 muestra el ajuste condicional de los dos modelos, permitiendo de nuevo concluir que hay diferencias significativas entre los dos y por tanto es mejor el modelo de regresión de Poisson.

SALIDA 1.6
Ajuste condicional de modelos

Modelo 1: *nerr* = 1
Modelo 2: *nerr* = 1 + *wais*

Modelo	Desvianza	gl	$\|\Delta D\|$	$\|\Delta gl\|$	$P > \chi^2$
1	42.820	19			
2	38.318	18	4.502	1	0.007

Aunque es un tema controvertido, es posible también calcular medidas pseudo-R^2 equivalentes al coeficiente de determinación R^2 de modelos *LM* como medidas de ajuste global. Una de las más comunes es $\tilde{R}^2$, propuesta por McFadden (1974), que se obtiene siguiendo la misma lógica que en los modelos de la estructura *LM* con las desvianzas del modelo de regresión de Poisson y la del modelo nulo:

$$\tilde{R}^2_M = \frac{D_0 - D_1}{D_0} = 1 - \frac{D_1}{D_0} \tag{1.12}$$

Para el modelo de regresión de Poisson, $\tilde{R}^2_M = 1 - (38.318/42.819) = .105$, que explica el 10.5 % de la desvianza. McFadden recomendó valores $R^2_M > .25$ para aceptar un modelo como globalmente ajustado. La interpretación de las medidas pseudo-R^2 no debe realizarse con un único modelo, sino en el contexto de varios modelos anidados, y debe aplicarse con suma precaución.

El cuadro 1.4 presenta la descomposición de la variación en el modelo de regresión de Poisson. Obsérvense las diferencias en formato respecto del cuadro 1.3, en consonancia con las diferencias que se comentaron anteriormente para la salida 1.6. Para interpretarlo adecuadamente, conviene tener presente algunas precisiones:

— En primer lugar, mientras que la variable de respuesta *nerr* sólo puede tomar valores dentro del rango $0/+\infty$, el predictor lineal presenta algunos valores negativos. Es preciso hacer uso de la inversa de la función de enlace, $\mu_i = \exp(\eta_i)$, para transformar los valores del predictor lineal en valores estimados. Así, utilizando la función de enlace inversa para el primer caso se obtiene el valor ajustado $\mu_1 = \exp(0.105) = 1.111$. Nótese que la columna de los valores estimados no presenta ningún valor fuera de rango, o sea, ningún valor inferior a cero.

— En segundo lugar, la descomposición de la variación se establece siempre tomando como modelo de partida el modelo nulo, cuya desvianza (la desvianza total, $D_T = 42.820$) se divide en un componente explicado por el modelo (la desvianza del modelo de regresión de Poisson simple, $D_M = 4.502$) y otro componente no explicado (la desvianza residual, $D_R = 38.318$).

— En tercer lugar, es importante comprobar que en este caso el componente residual no es exactamente la diferencia entre valores observados y valores estimados por el modelo. En los modelos de la estructura *GLM* pueden emplearse varios tipos de residuales. El que se utiliza aquí es un tipo de residual, denominado RESIDUAL DE DESVIANZA, que tiene una función similar al residual

©Ediciones Pirámide

clásico, pero no se define como la diferencia entre valores observados y valores estimados, sino que se utiliza una función más compleja (véase Olsson, 2002).

CUADRO 1.4
Detalle del proceso de ajuste del modelo de regresión de Poisson

Caso	*wais*	*nerr*	Predictor lineal	Valores estimados	Residual de desvianza	Grados de libertad
1	9	0	0.105	1.111	−1.491	1
2	11	1	0.374	1.453	−0.399	1
3	12	0	0.508	1.662	−1.823	1
4	8	0	−0.029	0.972	−1.394	1
5	10	2	0.240	1.271	0.596	1
6	7	0	−0.162	0.850	−1.304	1
7	9	1	0.105	1.111	−0.107	1
8	11	3	0.374	1.453	1.120	1
9	10	5	0.240	1.271	2.498	1
10	8	0	−0.029	0.972	−1.394	1
11	14	2	0.776	2.173	−0.119	1
12	10	0	0.240	1.271	−1.594	1
13	11	1	0.374	1.453	−0.399	1
14	16	0	1.045	2.842	−2.384	1
15	12	0	0.508	1.662	−1.823	1
16	17	2	1.179	3.251	−0.747	1
17	14	4	0.776	2.173	1.108	1
18	15	4	0.911	2.486	0.882	1
19	13	3	0.642	1.901	0.735	1
20	12	5	0.508	1.662	2.083	1
Desvianza total		42.820				19
Desvianza residual					38.318	18
Desvianza modelo				4.502		1

1.3.3. Modelo de regresión logística

El modelo de regresión logística es otro modelo de la estructura *GLM* que presenta sus propias peculiaridades. La variable de respuesta es ahora una variable binaria (*ssen*), que evalúa la presencia *versus* ausencia de síntomas seniles. Para esta variable de respuesta no sirve la distribución normal ni la distribución de Poisson, pero se adapta muy bien la distribución binomial con una de dos funciones de enlace muy similares: o bien el *enlace logit*, o bien el *enlace probit*. La salida 1.7 presenta los estimadores de los parámetros del modelo de regresión logística.

SALIDA 1.7
Estimadores de los parámetros del modelo de regresión logística

Variable dependiente: *ssen*

Parámetro	*B*	Error Típico	*z*	*P* > |*z*|	Límites IC 95 %	
					Inferior	Superior
Intersección	−14.597	6.816	−2.142	.035	−33.699	−5.096
wais	1.302	0.611	2.132	.033	0.455	3.012

©Ediciones Pirámide

La diferencia en los estimadores de los parametros entre los modelos de la estructura *GLM* y los de la estructura *LM* es que se emplea la distribución normal en lugar de la distribución t de Student. La ecuación predictiva es $\hat{\mu}_i = -14.497 + 1.302X_i$, donde la constante o intercepción es, en unidades logit, el valor de *ssen* cuando *wais* = 0 y la pendiente es el incremento en el logit de *ssen* por unidad de incremento en el predictor *wais*. La figura 1.5 muestra la función de regresión logística con sus bandas de confianza al 95 % sombreadas en gris. Nótese el deterioro de la predicción en casi todo el rango de valores del predictor.

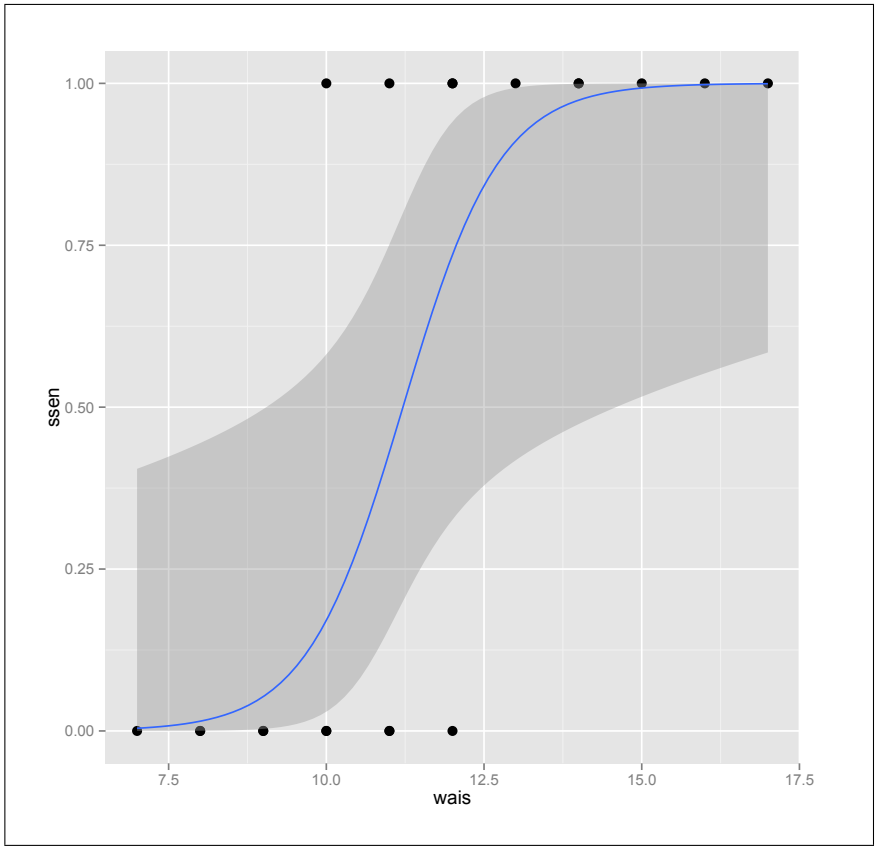

Figura 1.5. – Bandas de confianza del 95 % en torno a la función de regresión logística.

La salida 1.8 presenta, como la salida 1.5, la descomposición de la desviana total en dos componentes: la desviana no explicada por el modelo y la desviana explicada por el modelo. La desviana total (o desviana residual del modelo nulo) es $D_T = 27.726$, con $gl_T = 19$ grados de libertad, y la desviana residual del modelo de regresión logística con el predictor *wais* es $D_R = 12.709$, con $gl_R = 18$ grados de libertad. La desviana explicada por el modelo es la diferencia $D_M = D_T - D_R = 15.017$, con $gl_M = 1$. El ajuste global absoluto del modelo compara el modelo de regresión logística con el modelo saturado, utilizando la desviana residual con sus grados de libertad y la distribución χ^2 para la inferencia. En este caso, $\chi^2(18) = 12.709; P = .809$ apunta que el modelo ajusta bien, pero para muchos autores (p. ej., Agresti, 2013), con modelos de regresión logística el ajuste global absoluto sólo es válido con datos

©Ediciones Pirámide

agrupados y los datos que se emplean en este ejemplo son no agrupados. El ajuste global relativo nos permite concluir que el modelo de regresión logística difiere del modelo nulo: $\chi^2(1) = 15.017; P = .026$.

SALIDA 1.8

Análisis de desvianza del modelo de regresión logística

Variable dependiente: *ssen*

Fuentes	**Desvianza**	gl	χ^2	$P > \chi^2$
Modelo	15.017	1	15.017	.026
Residual	12.709	18		
Total	27.726	19		

CUADRO 1.5

Detalle del proceso de ajuste del modelo de regresión logística

caso	***wais*** X_i	***ssen*** Y_{i3}	**Predictor lineal**	**Valores estimados**	**Residual de desvianza**	**Grados de libertad**
1	9	0	−2.854	0.054	−0.330	1
2	11	0	−0.287	0.429	−1.061	1
3	12	0	0.997	0.730	−1.631	1
4	8	0	−4.138	0.016	−0.174	1
5	10	0	−1.570	0.172	−0.612	1
6	7	0	−5.421	0.004	−0.091	1
7	9	0	−2.854	0.054	−0.330	1
8	11	0	−0.287	0.429	−1.061	1
9	10	0	−1.570	0.172	−0.612	1
10	8	0	−4.138	0.016	−0.174	1
11	14	1	3.564	0.972	0.229	1
12	10	1	−1.570	0.172	1.881	1
13	11	1	−0.287	0.429	1.298	1
14	16	1	6.131	0.998	0.063	1
15	12	1	0.997	0.730	0.784	1
16	17	1	7.414	0.999	0.033	1
17	14	1	4.847	0.992	0.229	1
18	15	1	4.847	0.992	0.120	1
19	13	1	2.280	0.907	0.432	1
20	12	1	0.997	0.730	0.784	1
Desvianza total		27.726				20
Desvianza residual					12.709	18
Desvianza modelo				15.017		2

Como medida relativa de ajuste global, con cierta precaución puede aplicarse también una medida pseudo-R^2, por ejemplo, el coeficiente $\tilde{R}^2$ de McFadden. De la salida 1.8 se toman la desvianza del modelo nulo y la desvianza del modelo de regresión logística y se aplica: $\tilde{R}^2 = (27.726 - 12.709)/27.726 = 0.542$, apuntando que más del 54 % de la desvianza está explicada por el modelo con el predictor *wais*. En este caso parece que el ajuste global del modelo es aceptable.

1.3.4. Modelo de regresión multinivel

Aunque los modelos que consideramos a continuación son bastante más complejos que los modelos tratados hasta aquí, dada la relevancia que están alcanzando en la investigación psicológica ilustramos

©Ediciones Pirámide

también el proceso de ajuste de modelos simples de la estructura *LMM*. Los modelos ajustados anteriormente (regresión lineal simple de la estructura *LM* y regresión de Poisson y regresión logística de la estructura *GLM*) se caracterizan porque sólo tienen un componente de error residual, que es concretamente la desvianza residual de las salidas 1.2, 1.5 y 1.8. Incorporar otro (u otros) componente/s aleatorio/s cambiaría la estructura de los modelos *LM* y *GLM*, que asumen un único componente aleatorio. En tales casos el proceso de ajuste debe recurrir a estructuras de modelado más complejas como las estructuras *LMM* o *GLMM*.

El modelo de regresión lineal simple, que se ajustó anteriormente en la sección 1.2.1, utilizó 20 jubilados de una muestra más amplia, se propuso para explicar la variable de respuesta *dcog* en función de *wais* y obtuvo un ajuste global aceptable con $F(1,18) = 9.455, P = .007$, siendo su ecuación predictiva $\hat{Y}_i = 2.209 + 0.370X_i$. Sin embargo, el análisis sólo se practicó con una submuestra perteneciente a una muestra mayor. Más información sobre la muestra global utilizada revelaría que los 20 sujetos que se emplearon en el ajuste del modelo de regresión lineal tenían en común el pertenecer a un rango de edad igual o menor de 64 años, donde el deterioro cognitivo se asume leve o moderado. Pero la muestra total de unidades incluía en realidad 120 jubilados, divididos en 6 submuestras o grupos de 20 en función de su edad, de los que sólo se incluyó para ajustar el modelo de regresión lineal la primera submuestra (grupo 1). Los restantes grupos tenían rangos de edad que oscilaban entre 65 y 67 años (grupo 2), entre 68 y 70 años (grupo 3), entre 71 y 73 años (grupo 4), entre 74 y 76 años (grupo 5) y más de 76 años (grupo 6). Se asume que el deterioro cognitivo es también función de los grupos de edad, pero esta variable, que puede considerarse aleatoria porque recorre todo el rango de valores de edad que interesa a la investigación, no se incluyó en el análisis.

El investigador podría haber utilizado toda la información de los grupos y haber estimado un modelo de regresión con la muestra total de 120 sujetos, combinando todos los participantes de todos los grupos de edad. El resultado que habría obtenido, $F(1,118) = 61.072; P = .000$, le habría llevado también a considerar que la subescala del *wais* es un buen predictor del deterioro cognitivo. La ecuación predictiva en este caso sería $\hat{Y}_i = 2.919 + 0.440X_i$, y es muy similar a la obtenida para el grupo 1 (véase salida 1.1). Pero este análisis no toma en cuenta la variación debida a los grupos de edad y asume que los parámetros de intercepción y pendiente del correspondiente modelo de regresión lineal para cada nivel de grupo son iguales.

Para resolver la duda acerca de si es razonable asumir que los modelos de regresión lineal para cada grupo sean iguales, el investigador podría también realizar sendos análisis de regresión lineal simple separados con cada una de las submuestras y determinar si los resultados son iguales en el conjunto de los casos o si difieren entre grupos. El cuadro 1.6 muestra los resultados obtenidos con los seis análisis de regresión lineal simple separados para cada grupo, donde se intuye que las intercepciones aumentan considerablemente del primero al último grupo, pero las pendientes son muy similares y significativas en los cuatro primeros intervalos de edad, y no significativas y sensiblemente menores en los dos últimos intervalos.

La figura 1.6 muestra un gráfico con las rectas de regresión separadas para cada grupo de edad. El investigador puede concluir que el resultado del modelo de regresión lineal que se trató en la sección 1.2.1 sólo es válido para el intervalo de edad al que pertenece la submuestra o grupo de jubilados, pero no para la muestra global. Lo que los datos apuntan es que, dependiendo del grupo, el pronóstico de deterioro cognitivo a partir de la subescala del *wais* cambia. Es mejor en los grupos 1 a 4 y peor en los grupos 5 y 6. Por tanto, podría tener interés para el investigador tener en cuenta la variable de grupo, lo que plantea un nuevo problema. Mientras que en el análisis del modelo de regresión simple, sin tener en cuenta la variable de grupo, las unidades en las que se registra la variable de respuesta son jubilados, te-

©Ediciones Pirámide

niendo en cuenta la variable de grupo deben definirse dos tipos diferentes de unidades, la que se refiere a los jubilados (de cualquiera de los seis grupos) y la que se refiere al grupo en su conjunto. Los jubilados se consideran unidades de primer nivel, mientras que los grupos son unidades de segundo nivel. Con esta estructura jerárquica, el análisis que se requiere es el ANÁLISIS MULTINIVEL, que se ha desarrollado específicamente para el tratamiento de datos agrupados (*clustered*), emplea estimación por máxima verosimilitud y analiza por separado efectos fijos y efectos aleatorios. Y puesto que el modelo define más de un efecto aleatorio, pertenece a la estructura del modelo lineal mixto (*LMM*). El enfoque analítico usado con modelos de la estructura *LMM* es el ENFOQUE MIXTO.

CUADRO 1.6

Modelos de regresión separados para cada submuestra

Grupo	Intercepción	Pendiente	R^2	*P*
1: < 65	2.209	0.370	0.344	.007
2: 65 – 67	2.675	0.381	0.279	.017
3: 68 – 70	4.127	0.301	0.218	.038
4: 71 – 73	4.265	0.323	0.320	.009
5: 74 – 76	9.177	0.155	0.073	.249
6: > 76	11.036	0.138	0.039	.401
Global	2.919	0.440	0.341	.000

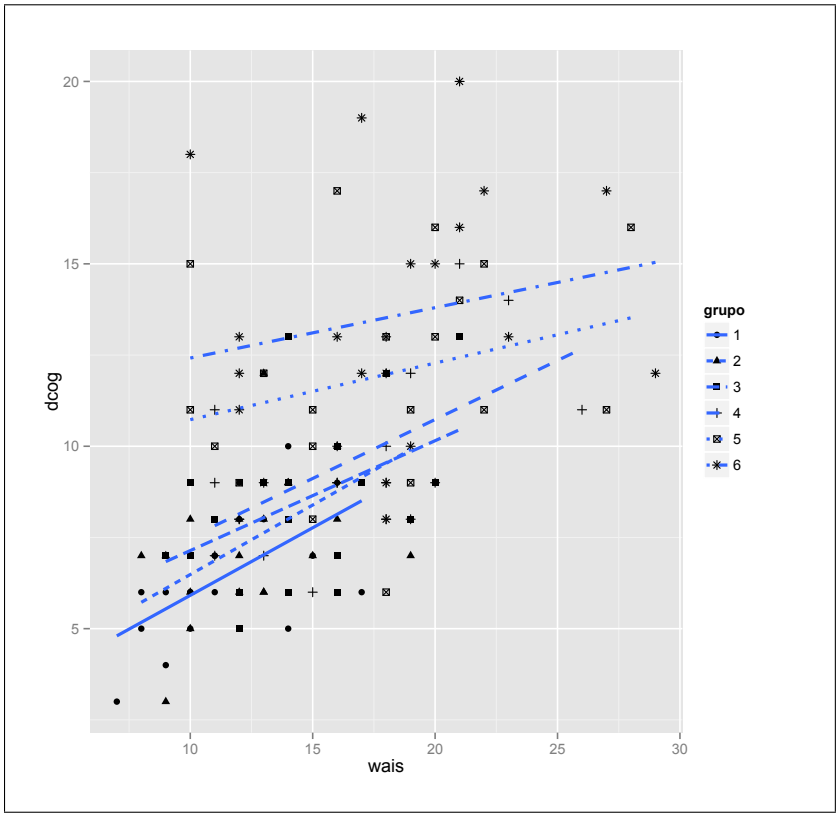

Figura 1.6. – Rectas de regresión separadas para cada grupo.

©Ediciones Pirámide

El primer paso para realizar un análisis multinivel consiste en dividir la variación de la variable de respuesta en dos componentes aleatorios: un componente de variación intragrupo que refleja la homogeneidad interna de cada grupo, y un componente de variación intergrupo que refleja la homogeneidad entre grupos. Si existiera poca variabilidad intergrupo (p. ej., menor del 5 %), el análisis multinivel no tendría ningún interés, y en su lugar se podría realizar un análisis de regresión convencional con la muestra completa de 120 casos.

La aplicación de análisis multinivel con la muestra completa de 120 casos requiere en primer lugar definir un *modelo nulo*, que con estructuras *LMM* es también un modelo que no tiene ninguna variable explicativa, pero que tiene $q = 3$ parámetros (la constante y los dos componentes de variación inter e intragrupo) en lugar de uno, como sucede en los modelos *LM*. La salida 1.9 resume los resultados del ajuste del modelo nulo con el enfoque mixto. Observe en primer lugar que el formato de salida es muy diferente de los formatos que hemos tratado hasta aquí, y distingue una primera parte informativa referente al ajuste del modelo (en particular, la desvianza y el número de parámetros del modelo), una segunda parte que contiene los efectos aleatorios del modelo y la correlación intraclase (o sea, la razón entre la variación intragrupo y la suma de la variación inter e intragrupo) y una tercera parte que contiene los efectos fijos del modelo, que en el modelo nulo solamente incluye la intercepción y usualmente no es objeto de inferencia.

SALIDA 1.9

Ajuste del modelo nulo (modelo 1) con estimación REML

Estructura de covarianza	***Loglik***	**Desvianza**	**Nº parámetros (*q*)**	***AIC***	***BIC***
Componentes de varianza	−290.323	580.646	3	584.646	590.204
Efectos aleatorios					
Parámetros	**Estimador**	**E. típico**	***Z* Wald**	$P > \chi^2$	**C.Intraclase**
Intercepción (edad)	6.944	4.597	1.151	.131	0.517
Residual	6.493	0.860	7.550	.000	
Efectos fijos					
Fuentes de variación	**Estimador**	**E. típico**	**Grados libertad**	***t***	$P > \|t\|$
Intercepción	9.592	1.101	114	8.715	.000

Los elementos esenciales de la salida del modelo nulo son, en primer lugar, su ajuste global. La desvianza (residual) del modelo nulo, con 6 (total de grupos) − 3 (parámetros utilizados) = 3 grados de libertad residuales es $D(3) = 580.646$, lo que revela que discrepa sensiblemente del modelo saturado y por tanto se requiere que el investigador continúe con el modelado de los datos. En segundo lugar, la razón entre la variación intergrupos y la variación total (o sea, la correlación intraclase) es $\hat{\rho}_I = 0.517$, lo que indica que casi un 51.7 % de la variación de la variable de respuesta se debe a diferencias entre grupos y que el restante 48.3 % es variación residual no explicada. En tercer lugar, en un modelo nulo no es usual evaluar la significación del único componente fijo, la constante o intercepción, que se refiere a la media de la variable de respuesta ($\hat{\beta}_0 = \overline{Y} = 9.592$). Pero una proporción tan importante de variación debida a diferencias entre grupos (más del 50 %) justifica ampliamente el empleo del análisis multinivel.

Una vez definido el modelo nulo, y observando que su ajuste global no es aceptable, el investigador puede considerar modelos más complejos. A título ilustrativo, por ejemplo, puede especificar un modelo introduciendo como efecto fijo la variable explicativa *wais*. Es posible ajustar un modelo que contemple variación en las intercepciones de las rectas de regresión separadas, o bien variación en las pendientes, o ambas (véase figura 1.6). Asumiendo que no habrá diferencias entre las pendientes (o sea, asumiendo

©Ediciones Pirámide

que hay una pendiente fija común) puede concentrar su atención en evaluar si las intercepciones difieren entre grupos. El modelo se denomina *modelo con intercepción aleatoria* (*random intercept model*) y su ajuste se presenta en la salida 1.10.

SALIDA 1.10

Ajuste del modelo de intercepción aleatoria (modelo 2) con estimación REML

Estructura de covarianza	***Loglik***	**Desvianza**	**Nº parámetros (*q*)**	***AIC***	***BIC***
Componentes de varianza	−282.402	564.803	4	568.803	574.344
Efectos aleatorios					
Parámetros	**Estimador**	**E. típico**	***Z* Wald**	$P > \chi^2$	**C.Intraclase**
Intercepción (edad)	3.911	2.714	1.441	.230	0.409
Residual	5.639	0.750	7.518	.006	
Efectos fijos					
Fuentes de variación	**Estimador**	**E. típico**	**Grados libertad**	***t***	$P > \|t\|$
Intercepción	5.714	1.177	17.241	4.856	.000
wais	0.256	0.055	117.251	4.682	.000

Las cuestiones esenciales a considerar en esta salida son, en primer lugar, la desvianza del modelo, que con 6 (número de casos/grupos) − 4 (número de parámetros usados) = 2 grados de libertad residuales, es $D(2) = 564.803$, lo que representa una leve mejora respecto del modelo nulo pero todavía se desvía mucho del modelo saturado. En segundo lugar, la correlación intraclase (la razón entre los dos efectos aleatorios del modelo, la variación intergrupos y la variación total) es de 0.409, o sea, casi un 41 % de la variación de la variable de respuesta se debe a variación de nivel 2 entre grupos y el restante 59 % es variación residual de nivel 1. En tercer lugar, la interpretación de los efectos fijos del modelo se aplica al conjunto de los jubilados participantes. Respecto a la constante o intercepción, el deterioro cognitivo observado cuando *wais* = 0 es de 5.714 unidades, y respecto a la pendiente, el incremento que se produce en deterioro cognitivo por incremento en *wais* es de $\hat{\beta}_1 = 0.256$ unidades, un incremento estadísticamente significativo: $t(117.251) = 4.682; P = .000$. Observe que los grados de libertad de la prueba t no es, como es usual, un número entero (117.251).

En comparación con el modelo nulo, con la introducción del predictor *wais* la variación residual de nivel 1 (intragrupo) ha descendido de 6.493 a 5.639, una reducción mayor del 13 %: (6.493−5.639)/6.493 = 0.132, lo que apunta que las puntuaciones obtenidas en la subescala del *wais* por los jubilados explica más del 13 % de la variabilidad intragrupo. Además, el predictor *wais* afecta también a la variación de nivel 2 (intergrupos), que tiene una reducción de casi un 44 %: (6.944 − 3.911)/6.944 = 0.437, lo que sugiere que las puntuaciones intragrupo en *wais* por los jubilados explican un 44 % de la variabilidad entre grupos en deterioro cognitivo.

Para dirimir si el modelo con intercepción aleatoria es mejor que el modelo nulo debe utilizarse el ajuste condicional de modelos. La salida 1.11 representa el ajuste condicional de dos modelos que difieren por la inclusión del predictor *wais*, el primero donde D se refiere a la desvianza residual y q se refiere al número total de parámetros empleado en los dos modelos. El ajuste condicional utiliza la prueba de la razón de verosimilitud y obtiene un valor de $L^2 = 19.613$. El valor de probabilidad de la distribución de χ^2 que le corresponde es: $\chi^2(1) = 19.613; P < .001$. Y puesto que se encuentran diferencias significativas entre ambos modelos, el investigador debe seleccionar el segundo modelo, o sea, modelo con intercepción aleatoria, como modelo más apropiado. A la misma conclusión conduce la comparación de los criterios de información AIC y BIC entre los dos modelos.

©Ediciones Pirámide

SALIDA 1.11

Ajuste condicional de los modelos nulo (1) y con intercepción aleatoria (2)

Modelo 1: *dcog* = 1 + (1 | *grupo*)

Modelo 2: *dcog* = 1 + *wais* + (1 | *grupo*)

Modelo	D	q	$\|\Delta D\|$	$\|\Delta q\|$	$P > \chi^2$
1	580.646	3			
2	564.803	4	15.843	1	0.000

1.4. LA ESTRUCTURA DEL MODELO LINEAL CLÁSICO (*LM*)

El Modelo Lineal Clásico (*LM*) es el instrumento analítico básico de la investigación en las Ciencias del Comportamiento, que integra procedimientos familiares para el investigador como el análisis de regresión, el análisis de varianza (ANOVA), el análisis de covarianza (ANCOVA) y, por generalización, muchos procedimientos multivariantes (MANOVA, MANCOVA y Análisis Discriminante, entre otros).

En su formulación univariante, la ecuación fundamental del *LM* es

$$Y_i = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j + e_i \tag{1.13}$$

donde Y_i ($i = 1,\ldots,N$) son N valores de la variable de respuesta, β_0 es un parámetro constante para todo i, $X_{ij} = (X_{i1},\ldots,X_{ij},\ldots,X_{ip})$ son los valores de p variables explicativas (covariantes y/o factores), cada una de las cuales tiene N elementos, y $\beta_j = \beta_1,\ldots,\beta_p$ es un conjunto de p coeficientes de regresión. En esta ecuación, el predictor lineal ($\eta_i = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j$) y el valor esperado ($\mu_i$) se miden en la misma escala de medida. Nótese que, en esta formulación, los datos son la variable de respuesta (Y_i), el modelo es el vector de valores estimados, que es igual al predictor lineal ($\mu_i = \eta_i$), y la discrepancia entre datos y modelo es el error residual, $e_i = Y_i - \mu_i$. Desarrollando la ecuación 1.13 para las N observaciones

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_i \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} X_{10} & X_{11} & \ldots & X_{1p} \\ X_{20} & X_{21} & \ldots & X_{2p} \\ \vdots & \ddots & \ddots & \vdots \\ X_{i0} & X_{i1} & \ldots & X_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ X_{N0} & X_{N1} & \ldots & X_{Np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_i \\ \vdots \\ e_N \end{pmatrix} \tag{1.14}$$

y escribiendo de forma compacta la ecuación (1.14) con notación matricial, resulta

$$\mathbf{Y}_{(N\times 1)} = \mathbf{X}_{(N\times q)}\boldsymbol{\beta}_{(q\times 1)} + \boldsymbol{e}_{(N\times 1)} \tag{1.15}$$

©Ediciones Pirámide

En esta ecuación, $\boldsymbol{Y}$ es el vector de la variable de respuesta, de orden $(N \times 1)$, donde cada caso corresponde a una observación, $\boldsymbol{X}$ es la matriz del modelo (o matriz de diseño) conocida, de orden $(N \times q)$, donde $q = p + 1$ (o sea, el número de parámetros es igual al número de predictores más la constante), cuyas filas son casos y cuyas columnas son variables, $\boldsymbol{\beta}$ es el vector de parámetros desconocidos del modelo, de orden $(q \times 1)$ y $\boldsymbol{e}$ es el vector de errores residuales, de orden $(N \times 1)$.

En su formulación multivariante, dadas m medidas diversas de una variable de respuesta (p. ej., ítems de un test, medidas repetidas de las mismas unidades experimentales, etc.) o bien m variables de respuesta diferentes, la estructura multivariante *LM* se formula mediante

$$\begin{pmatrix} Y_{11} & Y_{12} & \dots & Y_{1m} \\ Y_{21} & Y_{22} & \dots & Y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{i1} & Y_{i2} & \dots & Y_{im} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \dots & Y_{Nm} \end{pmatrix} = \begin{pmatrix} X_{10} & X_{11} & \dots & X_{1p} \\ X_{20} & X_{21} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{i0} & X_{i1} & \dots & X_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n0} & X_{n1} & \dots & X_{Np} \end{pmatrix} \begin{pmatrix} \beta_{10} & \beta_{11} & \dots & \beta_{1m} \\ \beta_{20} & \beta_{21} & \dots & \beta_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p0} & \beta_{p1} & \dots & \beta_{pm} \end{pmatrix} + \begin{pmatrix} e_{11} & e_{12} & \dots & e_{1m} \\ e_{21} & e_{22} & \dots & e_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ e_{i1} & e_{i2} & \dots & e_{im} \\ \vdots & \vdots & \ddots & \vdots \\ e_{n1} & e_{n2} & \dots & e_{Nm} \end{pmatrix} \tag{1.16}$$

y con notación matricial

$$\mathbf{Y}_{(N\times m)} = \mathbf{X}_{(N\times q)}\mathbf{B}_{(q\times m)} + \mathbf{E}_{(N\times m)} \tag{1.17}$$

donde **Y** es la matriz de medidas de la variable de respuesta, de orden $(N \times m)$, **X** es la matriz del modelo, de orden $(N \times q)$, **B** es la matriz de parámetros, de orden $(q \times m)$, y **E** es la matriz de errores residuales, de orden $(N \times m)$. Es importante notar que la matriz **X** es la misma para las dos formulaciones.

Las principales características de la estructura *LM* en su formulación univariante son las siguientes:

— La naturaleza del vector de la variable respuesta **Y** es numérica.

— La naturaleza de la matriz de variables explicativas **X** es mixta porque puede incluir variables numéricas y/o variables categóricas. Si todas las variables explicativas son numéricas (sólo predictores o covariantes), el resultado es un MODELO DE REGRESIÓN. Si todas las variables explicativas son categóricas (sólo factores), el resultado es un MODELO ANOVA. Si hay alguna mezcolanza de covariantes y factores, el resultado es un MODELO ANCOVA. Pero, en cualquier de los casos, la matriz del modelo **X** se considera una matriz de valores fijos que no cambian en posteriores repeticiones de la investigación. Como consecuencia, el vector $\boldsymbol{\beta}$ es un vector de parámetros de efectos fijos.

— Los valores estimados (μ_i) y el predictor lineal (η_i) se miden en la misma escala de medida, y por tanto la función de enlace es la identidad.

— Los supuestos básicos que deben cumplir los modelos de esta estructura son:

1. Se asume una distribución normal de la variable de respuesta (*supuesto de normalidad*).

©Ediciones Pirámide

2. Se supone constante la varianza de la variable de respuesta para cada uno de los valores de las covariantes y/o factores (*supuesto de homocedasticidad*).
3. Las observaciones Y_i se registran en diferentes unidades experimentales y por tanto se asumen mutuamente independientes (*supuesto de independencia*).
4. Se asume una relación lineal entre variables explicativas y variable de respuesta (*supuesto de linealidad*).

Muchos modelos estadísticos no presentan graves problemas cuando se viola de forma moderada el supuesto de normalidad. Pero todo lo contrario sucede cuando se violan los supuestos de independencia, homocedasticidad o linealidad. La violación de los supuestos de independencia y homocedasticidad requieren un tratamiento diferenciado. Los modelos de la estructura *LMM* y el modelo *GEE* de la estructura *GLM* se han desarrollado precisamente con el objeto de tratar violaciones de estos dos supuestos. La violación del supuesto de linealidad representa un error de especificación que requiere emplear modelos no lineales (véase Bates y Watts, 2008).

Una forma compacta de resumir estos supuestos suele presentarse argumentando que los errores residuales se distribuyen normal e independientemente con varianza constante, $e_i \sim NID(0, \sigma_e^2)$, donde σ_e^2 es el parámetro de escala que se asume constante para todas las observaciones. Al suponer constante la varianza, el interés se concentra en el modelado de las medias de la variable de respuesta utilizando los parámetros de efectos fijos $\boldsymbol{\beta}$.

— El procedimiento más común de estimación de los parámetros β_j de un *LM* es MÍNIMOS CUADRADOS ORDINARIOS, y su objetivo empírico es la minimización de la suma cuadrática de los errores residuales $e_i = Y_i - \hat{\mu}_i$, o sea, que $\sum e_i^2 \longrightarrow$ mínimo.

— La interpretación de los parámetros de un modelo especificado por el investigador se determina utilizando la distribución t de Student, que se comporta aceptablemente incluso con muestras pequeñas.

— El ajuste global absoluto de un modelo especificado por el investigador no puede valorarse. El ajuste global relativo se determina comparándolo con el modelo nulo, utilizando para ello la razón F de varianzas.

Dedicamos un tratamiento más detallado de algunos de los modelos de la estructura *LM* en los capítulos 2 a 5. Una excelente exposición del modelo lineal clásico desde la perspectiva de la comparación de modelos puede consultarse en el texto de Judd *et al.* (2008) y en español en el primer volumen del texto del Grupo ModEst (2005a).

1.5. LA ESTRUCTURA DEL MODELO LINEAL GENERALIZADO (*GLM*)

El *GLM* es en esencia una reformulación más amplia del *LM*, al que se han incorporado como casos especiales todos los procedimientos del *LM* para variables de respuesta numéricas (regresión, ANOVA y ANCOVA) y procedimientos específicos para variables de respuesta categóricas tales como el análisis loglineal, el análisis logit, la regresión logística y la regresión de Poisson. Lo que todos estos modelos añadidos tienen en común es el requisito de linealidad y un método común para estimar sus parámetros, que a diferencia de los modelos *LM* es MÁXIMA VEROSIMILITUD (*ML*).

La ecuación fundamental del *GLM* es la misma que para el *LM*. Con notación matricial:

©Ediciones Pirámide

$$\mathbf{Y}_{(n\times 1)} = \mathbf{X}_{(n\times q)}\boldsymbol{\beta}_{(q\times 1)} + \mathbf{e}_{(n\times 1)} \tag{1.18}$$

donde **Y** son los datos, $\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\mu}$ es el modelo y **e** es el error residual. Hay dos diferencias esenciales entre el *LM* y el *GLM*:

— La utilización de una función de enlace diferenciada, para transformar la escala del predictor lineal, donde se realiza la computación estadística, a la escala del valor esperado, donde se calcula la discrepancia entre modelo y datos empíricos,

$$\eta_i = g(\mu_i) \tag{1.19}$$

— La definición de una función de varianza que captura cómo la varianza de la variable de respuesta depende de la media,

$$V(Y_i) = \phi V(\mu_i) \tag{1.20}$$

donde ϕ es un parámetro de variación constante.

Teniendo en cuenta que la formulación de modelos *GLM* incluye los modelos *LM* como un caso particular (donde la función de enlace es la identidad), las características específicas del *GLM* son las siguientes:

— La naturaleza del vector de la variable de respuesta **Y** puede ser numérica o categórica. La generalización del *LM* al *GLM* es precisamente la inclusión de variables de respuesta categóricas.

— La naturaleza de la matriz del modelo **X** puede ser también variable. Puede contener variables explicativas numéricas, categóricas o una mezcolanza de las dos.

— Valor esperado (μ_i) y predictor lineal (η_i) no se miden en la misma escala, y por tanto se precisa de una función de enlace para transformar los valores del predictor lineal en valores estimados. La naturaleza de la matriz **X** y de la función de enlace determinan los procedimientos estadísticos que son aplicables. En el cuadro 1.7 se resumen los procedimientos más usuales.

— La generalización desde el *LM* al *GLM* permite supuestos mucho menos restrictivos que los del *LM*. En concreto:

1. No se asume normal la distribución de la variable de respuesta, porque depende del modelo especificado y de la función de enlace utilizada. El *GLM* puede utilizar cualquier distribución de la familia exponencial y no solamente la distribución normal.
2. No se asume constante la varianza de la variable de respuesta para cada uno de los valores de los predictores/covariantes, porque depende también de la distribución y la función de enlace.

©Ediciones Pirámide

3. El supuesto de linealidad se asume para la escala del predictor lineal, pero no para la escala del valor esperado.
4. El supuesto de independencia, en cambio, es un requisito fundamental para todos los procedimientos del *GLM*. Sin embargo, existe una extensión de la estructura *GLM* apropiada para el análisis de datos que no cumplen el supuesto de independencia (p. ej., datos longitudinales y jerárquicos). Tal extensión fue originalmente propuesta por Liang y Zeger (1986) y Zeger, Liang y Albert (1988). Los modelos típicos de esta extensión de la estructura *GLM* se conocen como modelos *GEE* (*Generalizing Estimating Equations*). Más detalles sobre las características de los modelos *GEE*, junto con una aplicación práctica de los mismos, puede consultarse en Ghisletta y Spini (2004).

— Para los procedimientos que generalizan el *LM*, el método general de estimación de parámetros es MÁXIMA VEROSIMILITUD o alguna de sus variantes (p. ej., *máxima cuasi-verosimilitud* para el modelo *GEE*).

El cuadro 1.7 presenta una clasificación de los principales modelos lineales generalizados (adaptada de Grupo ModEst, 2005b). Los criterios de clasificación que definen las columnas del cuadro 1.7 se especifican a continuación (véase también cuadro 1.1):

— *Naturaleza de la variable de respuesta*. Varios tipos de variable de respuesta son posibles en un *GLM*:

- VARIABLE NUMÉRICA (cuantitativa), que puede adoptar un número infinito de valores positivos y/o negativos. Se registra para cada unidad de observación.
- VARIABLE CATEGÓRICA, una variable que sólo puede adoptar un número limitado de valores positivos y se mide en escala nominal u ordinal. Se distinguen entre sí en función del número de categorías en VARIABLES BINARIAS o DICOTÓMICAS (si solamente tienen dos categorías) y VARIABLES POLITÓMICAS (si tienen más de dos categorías). Este tipo de variables puede utilizarse de forma no agrupada (para cada unidad de observación) o bien, si todas las variables explicativas son categóricas, de forma agrupada (para conjuntos de unidades de observación) y organizadas bajo la forma de *tablas de contingencia*.
- VARIABLE DE RECUENTO, una variable que sólo puede adoptar un número finito de valores positivos y representa el número de veces que un evento ocurre. Se registra para cada unidad de observación.
- VARIABLE DE FRECUENCIA, una variable que puede adoptar un número finito de valores positivos, se registra agrupando unidades de observación y se representa bajo la forma de tablas de contingencia.

— *Componente sistemático*. Se refiere al predictor lineal, que es una función de las variables explicativas, y su estructura depende de la naturaleza de las variables explicativas, que puede ser numérica, si todas las variables explicativas tienen naturaleza cuantitativa; categórica, si todas tienen naturaleza categórica (nominal u ordinal), o mixta, si en el predictor lineal se combinan variables explicativas continuas y categóricas.

— *Componente aleatorio*. Hay un componente único, y se refiere a la distribución que se asume para el componente de error, que puede ser cualquier miembro de la familia exponencial, siendo las más comunes las distribuciones normal, binomial, multinomial y Poisson.

©Ediciones Pirámide

— *Función de enlace*. Se refiere a la función de transformación de predictor lineal a valores estimados. Las más comunes funciones de enlace de los *GLM* son:

- el ENLACE IDENTIDAD: predictor lineal y valor esperado se expresan en la misma unidad de medida;
- el ENLACE LOGARÍTMICO: el predictor lineal es el logaritmo del valor esperado, y
- el ENLACE LOGIT: el predictor lineal es el logaritmo de una razón de probabilidades (*odds*) del valor esperado.

CUADRO 1.7
Principales procedimientos del MLG

Naturaleza de la variable respuesta (rango)	**Componente sistemático**	**Componente aleatorio**	**Función de enlace**	**Modelo aplicable**
numérica cuantitativa $(-\infty/+\infty)$	numérica	Normal	identidad	Regresión lineal/GEE
	categórica	Normal	identidad	ANOVA
	mixta	Normal	identidad	ANCOVA
categórica binaria (0/1)				
* no agrupada	mixta	Binomial(1)	logit	Regresión logística/GEE
* agrupada	categórica	Binomial(n)	logit	Análisis logit
categórica politómica (0/*n*)				
* no agrupada	mixta	Multinomial	logit	Regresión logística/GEE
* agrupada	categórica	Multinomial	logit	Análisis logit
recuento $(0/+\infty)$	mixta	Poisson	logarítmica	Regresión de Poisson/GEE
frecuencia (0/100)	categórica	Poisson	logarítmica	Análisis loglineal

1.6. LA ESTRUCTURA DEL MODELO LINEAL MIXTO (*LMM*)

Todos los modelos incluidos en la estructura *GLM* contienen sólo un componente aleatorio, representado por los errores residuales e_i, que se distribuyen con varianza constante según $e_i \sim NID(0,\sigma_e^2)$. Los supuestos de homocedasticidad e independencia son en particular requisitos altamente restrictivos que hacen inaplicable los modelos de la estructura *GLM* (con la excepción del modelo *GEE*, como tratamos más arriba) en muchas situaciones de investigación comunes en psicología, sobre todo en estudios longitudinales y en estudios multinivel que implican diferentes niveles de agregación (*clustered data*).

La estructura del Modelo Lineal Mixto (*LMM*) es una extensión de la estructura *LM* propuesta por Laird y Ware (1982) que se formula, respecto de la ecuación 1.15, mediante

$$\mathbf{Y}_{(N\times 1)} = \mathbf{X}_{(N\times q)}\boldsymbol{\beta}_{(q\times 1)} + \mathbf{Z}_{(N\times t)}\mathbf{u}_{(t\times 1)} + \mathbf{e}_{(N\times 1)} \tag{1.21}$$

donde $\boldsymbol{\beta}$ es un vector de efectos fijos, $\mathbf{u}$ es un vector de efectos aleatorios, $\mathbf{X}$ es una matriz de efectos fijos constantes del modelo y $\mathbf{Z}$ es una matriz de efectos aleatorios, que es en general un subconjunto de la matriz de efectos fijos, ya que $t \leq q$. Los errores aleatorios se distribuyen según $e_i \sim N(0,\sigma_e^2)$, aunque ya no se requiere que la varianza σ_e^2 sea constante u homocedástica, y los efectos aleatorios se distribuyen también de forma normal con $u_i \sim N(0,\mathbf{G})$, siendo $\mathbf{G}$ la matriz de covarianza de los efectos aleatorios. En

©Ediciones Pirámide

consecuencia, la extensión que la estructura *LMM* supone respecto de la estructura *LM* es simplemente la inclusión de más efectos aleatorios en el modelo. Por lo demás, **Y** es el vector que representa los datos empíricos, $\mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} = \boldsymbol{\eta} = \boldsymbol{\mu}$ es el predictor lineal (que incluye los efectos fijos y efectos aleatorios del modelo) y **e** es el error residual, que representa la discrepancia entre datos empíricos y modelo.

Esta estructura es especialmente idónea para modelar situaciones de investigación donde la varianza no es constante y/o las observaciones no son independientes, aunque se asume con cierta flexibilidad que son poblaciones normales. Situaciones donde se utilizan como unidades experimentales individuos que participan de características comunes (p. ej., miembros de una misma familia al investigar la interacción diádica madre-hijo), observaciones jerárquicamente dependientes (p. ej., sujetos dentro de aulas, aulas dentro de colegios y colegios dentro de distritos escolares) o en general estudios longitudinales en los que se registran medidas repetidas de los mismos sujetos en diferentes momentos en el tiempo, son aplicaciones potenciales de los modelos de la estructura *LMM* (Hedeker, 2005). Muchas investigaciones de la psicología aplicada utilizan alguna de estas situaciones.

Sin embargo, el procedimiento de estimación de los parámetros $\boldsymbol{\beta}$ y **u** no puede ser el procedimiento utilizado en la estructura *LM*, o sea, mínimos cuadrados ordinarios. En su lugar es preciso emplear alguna forma de estimación por máxima verosimilitud, cuya complejidad hace inviable que pueda practicarse con una sencilla calculadora de escritorio, como sería posible con la estimación por mínimos cuadrados ordinarios. Afortunadamente, los paquetes estadísticos profesionales han introducido entre sus programas la estimación de este tipo de modelos, por lo que en los capítulos que utilizan modelos *LMM* nos centraremos en la interpretación y pasaremos por alto los detalles computacionales.

1.7. LA ESTRUCTURA DEL MODELO DE ECUACIONES ESTRUCTURALES (*SEM*)

La ecuación general del modelado: DATOS = MODELO + ERROR (ecuación 1.1) requiere una ecuación única, mediante la cual se define una variable dependiente (los DATOS) y se especifica un MODELO compuesto de un conjunto de variables independientes junto con sus correspondientes parámetros.

Muchas situaciones de investigación requieren emplear un sistema de ecuaciones para capturar de forma más precisa las relaciones entre las variables y para profundizar en su estructura causal. Usualmente tal sistema de ecuaciones se representa mediante un diagrama donde todas las variables (que se representan con rectángulos) se conectan con flechas o rutas (*paths*) con otras. Las variables que no reciben ninguna flecha se denominan en este contexto *variables exógenas* y las que reciben flechas de otras variables se denominan *variables endógenas*. Los errores del modelo se representan con círculos. El modelo explicativo que se representa en un diagrama se denomina MODELO ESTRUCTURAL o MODELO PATH (*path model*). El análisis que se practica en estos casos se conoce como ANÁLISIS DE RUTAS o ANÁLISIS PATH *path analysis*) y requiere definir un sistema de ecuaciones de regresión que refleje las relaciones entre las variables manifiestas. Más detalles sobre los componentes de un modelo estructural y el análisis *path* pueden consultarse a un nivel introductorio en Pedhazur (1997) y Lei y Wu (2007).

Un ejemplo de modelo *path* se representa en la figura 13.7 (capítulo 13), y muestra las relaciones entre cuatro variables observables, de las que una es exógena (*estresor conglomerado*) y las otras tres son endógenas (*apoyos*, *sentido de la coherencia* y *nivel de estrés*).

Un modelo *path* es por tanto una generalización de los modelos de regresión de la estructura *GLM* (fundamentalmente, regresión normal, regresión logística y regresión de Poisson) y solamente utilizan VARIABLES MANIFIESTAS u OBSERVABLES, cada una de las cuales se encierra en un rectángulo en el diagra-

©Ediciones Pirámide

ma path. Pero también es posible emplear VARIABLES LATENTES NO OBSERVABLES, que son constructos que se definen a partir de un conjunto de variables manifiestas (también llamadas en este contexto *indicadores*). Para distinguirlos de las variables manifiestas suelen representarse con círculos en un diagrama path. Se denomina MODELO DE MEDIDA a la representación de las variables latentes en función de un conjunto de variables manifiestas o indicadores.

Los modelos *SEM* (*Structural Equation Model*) se definen en este contexto como la combinación particular de un modelo de medida y un modelo estructural. Un ejemplo de modelo *SEM* se representa en la figura 13.8 (capítulo 13) y consta de un modelo estructural que expresa las relaciones entre cuatro constructos o variables latentes (los círculos del diagrama), cada uno de los cuales se definen con un conjunto de indicadores, donde *violencia* se formaliza con los indicadores 'física', 'psicológica' y 'sexual' (expresados con rectángulos en el diagrama), *estrés postraumático* con los indicadores 'estrés1' y 'estrés2', y *desconexión y rechazo* con los indicadores 'rechazo1' y 'rechazo2' y *autonomía deteriorada* con los indicadores 'autonomía1' y 'autonomía2'.

El proceso de ajuste de un modelo *SEM* se realiza de forma similar a los modelos tratados anteriormente, aunque tiene sus propias peculiaridades. Básicamente, con las correlaciones entre las variables (los DATOS) se obtiene una matriz de covarianza muestral. A partir de esta matriz y del modelo *SEM* propuesto para explicar los datos empíricos se calcula una matriz de covarianza estimada por el modelo. Si la matriz de covarianza estimada no se desvía sensiblemente de la matriz de covarianza muestral el modelo se ajustará aceptablemente; en caso contrario, el modelo no se ajustará. El ajuste global se practica, como es usual, con la prueba de la razón de verosimilitud L^2, distribuida según χ^2, aunque se acompaña también de otros indicadores de ajuste alternativos para capturar la desviación del modelo propuesto respecto de los datos empíricos que se propone explicar. La prueba de la razón de verosimilitud no resultó significativa ni en el modelo *path* de la Figura 13.7, $\chi^2(1) = 0.114; P = .74$, ni en el modelo *SEM* de la figura 13.8, $\chi^2(21) = 31.62; P = .06$, y por tanto se asume un ajuste global absoluto satisfactorio.

1.8. SUGERENCIAS DE ESTUDIO

Asumimos en el lector una información básica sobre Estadística Descriptiva e Inferencial. Los textos de Botella, Suero y Ximénez (2012), Howell (2013), Agresti y Franklin (2013) o preferentemente Kaplan (2011) pueden servir para consolidar tal información.

La filosofía general y la historia del modelado estadístico y el proceso de revolución silenciosa contra el *NHST* puede leerse en un trabajo de Rodgers (2010). Claves en esta historia han sido las monografías de Judd *et al.* (2008), Maxwell y Delaney (2004) y Kaplan (2011). Un texto moderno sobre el modelado y las estructuras que incluye es el de Tarling (2009). En español pueden consultarse dos asequibles monografías del Grupo ModEst (2005a, 2005b).

Para los modelos de la estructura *LM* hay una ingente cantidad de referencias con orientación psicológica, la mayoría especializadas en modelos de regresión, modelos ANOVA y modelos ANCOVA que trataremos con más detalle en los próximos capítulos.

Una auténtica revolución estadística produjo en la década de los 70 del siglo pasado la formulación del MODELO LINEAL GENERALIZADO (estructura *GLM*) propuesta por Nelder y Wedderburn (1972). La "biblia" de los modelos lineales generalizados es el texto de McCullagh y Nelder (1989). Textos más introductorios son Collett (2002), Dobson y Barnett (2002), Hutcheson y Sofroniou (1999) y Olsson (2008). Más completo es el texto de Myers, Montgomery, Vining y Robinson (2010), aunque los ejemplos son

©Ediciones Pirámide

ajenos a la Psicología. Una introducción básica en español se ofrece en el segundo volumen del Grupo ModEst (2005b).

Introducciones asequibles a los MODELOS *GEE* de la estructura *GLM* son Ballinger (2004) y Ghisletta y Spini (2004). Pueden encontrarse algunas aplicaciones con este tipo de modelos en la investigación psicológica, particularmente en Psicología del desarrollo.

Todavía no son muy abundantes las referencias existentes sobre los MODELOS LINEALES MIXTOS (estructuras *LMM* y *GLMM*), y casi todas ellas se asocian a algún (o algunos) programa/s de software. Además del trabajo clásico de Laird y Ware (1982), son especialmente recomendables el texto de Pinheiro y Bates (2000) para *S* y *R*, el de Littell, Milliken, Stroup, Wolfinger y Shabenberger (2006) para el programa *SAS*, el de Heck, Thomas y Tabata (2010) para *SPSS*, y los dos volúmenes de Rabe-Hesketh y Skrondal (2012a,b) para *STATA*. Todos los programas estadísticos que ajustan modelos *LMM* se tratan brevemente en West y Galecki (2011) y con mayor profundidad en West, Welch y Galecki (2015).

Más abundantes son las referencias sobre los modelos de la estructura *SEM*, que hoy por hoy son los más demandados. Una de las referencias más recomendadas sobre algunos de los modelos más comunes empleados en la literatura psicológica que incluye esta estructura es el texto de Kline (2011).

©Ediciones Pirámide