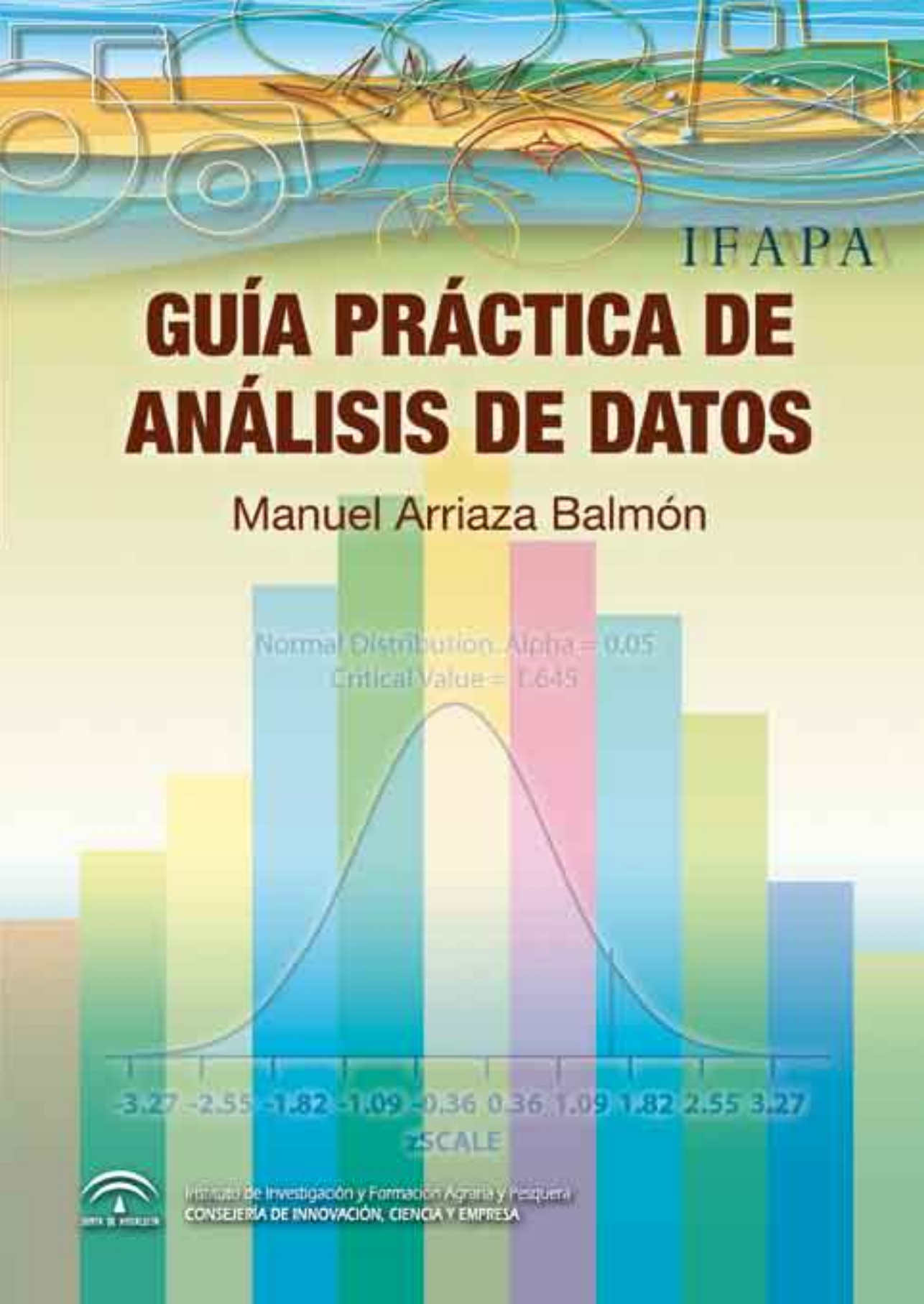
IFAPA

GUÍA PRÁCTICA DE ANÁLISIS DE DATOS

Manuel Arriaza Balmón



Instituto de Investigación y Formación Agraria y Pesquera
CONSEJERÍA DE INNOVACIÓN, CIENCIA Y EMPRESA

GUÍA PRÁCTICA DE ANÁLISIS DE DATOS

Autor:

Manuel Arriaza Balmón. IFAPA

Depósito Legal:

I.S.B.N.: 84-611-1661-5

Diseño e impresión:

Ideagonal Diseño Gráfico

A mi esposa, mi mejor hallazgo
A mis hijos, mi mayor logro

ÍNDICE

PRÓLOGO

Las ciencias sociales necesitan de sus propios laboratorios para la investigación. Estos laboratorios no son habitáculos ubicados en recintos cerrados más o menos equipados con máquinas o instrumentos de pesaje o medida como puede ocurrir en cualquiera de los conocidos laboratorios de ciencias experimentales. Los laboratorios de las ciencias sociales necesitan espacios abiertos, necesitan información básicamente primaria la cual se tiene que conseguir en el exterior de los centros de investigación, aunque también se utiliza la información secundaria, la cual se consigue en el propio centro de investigación a partir de documentos escritos, libros especializados, revistas científicas y en estos tiempos, a través de las autopistas de la información.

La información conseguida, independientemente del lugar de donde proceda o de la forma de conseguirla, hay que analizarla, depurarla y procesarla para mejorar e incrementar el nivel de conocimiento sobre la materia sobre la que se investigue. La recogida de información primaria se suele hacer mediante el procedimiento de la encuesta preguntando a las personas que poseen dicha información, la cual se recoge en un formulario o cuestionario más o menos estructurado. El tratamiento de la información conseguida se hace mediante la observación y el recuento de cada una de las respuestas obtenidas. Esto dicho así parece muy simple y fácil y realmente no es complejo, sin embargo en la práctica es tremendamente laborioso por la cantidad de información que se suele manejar en cualquier trabajo de investigación que se realice.

En este momento me viene a la memoria el programa informático en BASIC que desarrollé allá por el año 1983, cuando la informática estaba en pañales, para analizar la información que se había recogido en un proyecto que se desarrollaba por profesores del Departamento de Economía y Sociología Agrarias de la Universidad de Córdoba, en el seno del Instituto de Sociología y Estudios Campesinos, sobre el cooperativismo agrario en Andalucía. La utilidad de dicho programa fue muy importante en aquel proyecto, así como en los numerosos trabajos de investigación en los que posteriormente se utilizó. Además los primeros programas, también en BASIC que utilizábamos los doctorandos de la Escuela (ETSIA de Córdoba) en la segunda mitad de la década de los años setenta, para el ajuste de ecuaciones lineales multivariantes o multiecuacionales, que todos tratábamos de mejorar y completar para conseguir una mayor y mejor información y resultados, han pasado a la historia.

En este sentido podríamos destacar el intenso trabajo al que se sometía el primer ordenador de la Escuela, un Hewlett Packard 9830 A, de 1 Kb de

memoria, con un monitor de una línea y como hardware para el archivo de programas y datos una cinta magnética. Otro hardware que se adquirió en la Escuela posteriormente para la entrada de datos e información en general fue a través de tarjetas perforadas que tuvo una escasísima utilidad. La mayor limitación de estos equipos, aparte de la lentitud en el procesado de la información, era ocasionada por la reducida capacidad para el manejo de datos y la poca elaboración de los resultados, lo que suponía una ardua tarea en el análisis, depuración e interpretación de las salidas por impresora.

El desarrollo tan amplio y rápido que ha experimentado la informática, tanto en lo relativo al hardware como al software, ha puesto a punto en el mercado herramientas muy potentes utilizadas de igual forma en el análisis de encuestas como en el tratamiento informático de la información obtenida de las mismas. Dicho tratamiento incluye el análisis estadístico simple, o estadística descriptiva univariante, así como la obtención de modelos multivariantes en los que el comportamiento de una o varias variables (variables endógenas o dependientes) es explicado por el valor que toman otras variables de tipo técnico y/o económico (variables exógenas o independientes). En este sentido, siempre he creído en el poder explicativo de los modelos económicos, por lo que desde mis comienzos en el mundo de la investigación en economía he utilizado estas herramientas aplicándolas a diferentes campos, tanto en la determinación de modelos de oferta como de demanda en el sector agrario, así como en el campo de la valoración agraria.

En relación con el libro que el lector tiene en sus manos, tengo que decir que cuando el autor del mismo, mi amigo el Dr. Arriaza, me propuso que le escribiera el prólogo, estábamos en la Cordillera de los Andes en Venezuela, en el Estado Táchira, impartiendo un programa de doctorado. En aquel tiempo, abril de 2003, la materia que él impartía era prácticamente el contenido de este libro que gira en torno a todo el proceso necesario en un trabajo de investigación empírico en el que tenemos que empezar por ver la información disponible y la necesaria para acometer dicho trabajo de investigación. La información primaria que necesitamos se ha de conseguir mediante encuestas, las cuales hay que depurar y analizar. Posteriormente la información conseguida se tiene que procesar para obtener los resultados que nos permitan obtener conclusiones que mejoren el estado de conocimiento de la materia o el área de estudio.

Para mí es una satisfacción presentar y comentar este libro en el que podemos ver todo el proceso indicado anteriormente expuesto con claridad, de forma precisa y concisa por lo que se puede seguir perfectamente el camino que hay que recorrer desde que se plantea una investigación empírica hasta que obtenemos los resultados de la misma. Como digo, es un libro

eminentemente práctico en el que los planteamientos teóricos son los justos y necesarios para entender las aplicaciones que se pueden realizar y que se acompaña en todos los casos con ejemplos y aclaraciones que facilitan tanto la comprensión como la aplicación. La exposición de los conceptos teóricos generales se realiza en los tres primeros capítulos dedicados a la recogida de datos, a la clasificación de las variables y a la contrastación de hipótesis como elemento fundamental en el avance del conocimiento científico.

Por último, y en relación con los programas informáticos que ya citaba en párrafos anteriores, hay que destacar la utilidad del programa OS4 en el análisis y tratamiento informático de los datos. Este programa, aún siendo gratuito, tiene una gran potencia tanto en el análisis como en la evaluación de los resultados que se pueden obtener de la información recogida. En el texto se muestran de forma sencilla los requerimientos necesarios para aplicar cada una de las diferentes técnicas de análisis estadístico, así como su idoneidad de uso en cada caso, proporcionando al analista toda la información necesaria para poder tomar decisiones.

Para terminar confío y deseo que este texto no sea uno más de los existentes en el mercado, sino que sea un libro de referencia para los estudiosos e investigadores de las diferentes disciplinas relacionada con las Ciencias Sociales, convirtiéndose en una herramienta más de trabajo en el análisis de la información primaria necesaria en toda investigación.

Juan A. Cañas Madueño
Catedrático de Economía Financiera y Contabilidad
Córdoba, marzo de 2006

AGRADECIMIENTOS

En primer lugar agradezco la ayuda recibida de William G. Miller, el autor del paquete estadístico OS4, en la elaboración del presente texto. Este agradecimiento se extiende tanto a sus comentarios sobre algunos puntos del análisis estadístico como a su eficacia a la hora de introducir algunos cambios en el paquete estadístico.

A D. Juan Antonio Cañas Madueño, catedrático de Economía Financiera y Contabilidad de la Universidad de Córdoba por sus valiosos comentarios sobre una primera versión del libro y por su contribución prologándolo.

A mis compañeros del Área de Economía y Sociología Agraria del Centro 'Alameda de Obispo' de Córdoba, y en especial a D. José González Arenas, por sus comentarios y apoyo en todo momento.

Asimismo, a la institución para la que trabajo, el Instituto de Investigación y Formación Agraria y Pesquera (IFAPA) de la Consejería de Innovación, Ciencia y Empresa por la provisión de todos los medios materiales necesarios para la elaboración y publicación de esta obra.

A mi esposa y mis hijos, por todos los momentos que no he podido compartir con ellos y que dediqué a este empeño. De igual forma, a mis padres, a los que tanto les debo.

PRESENTACIÓN DEL ESTUDIO

El presente texto tiene un objetivo eminentemente práctico y está destinado a cualquier persona que, necesitando utilizar la estadística como herramienta de trabajo, tiene unos conocimientos mínimos sobre esta disciplina. Por este motivo hemos limitado los aspectos teóricos al mínimo necesario para entender la técnica estadística y las situaciones habituales en que se aplica.

La primera parte del libro trata todos los aspectos relacionados con la recogida de información a través de la encuesta y el análisis preliminar de los datos. Se explican los diferentes métodos de muestreo y se dan algunos consejos prácticos sobre el diseño del cuestionario. A continuación se detallan los diferentes tipos de variables según su escala de medida, aspecto éste fundamental a la hora de seleccionar el tipo de análisis estadístico aplicable. El siguiente capítulo desarrolla la contrastación de hipótesis estadísticas, resaltando su papel central en la interpretación de cualquier prueba estadística. La parte primera finaliza con una introducción al paquete estadístico que se ha utilizado para la elaboración de este manual.

La segunda parte del libro presenta algunas de las técnicas más comunes de análisis estadístico. De entre todos los capítulos destaca por su extensión el análisis de regresión lineal múltiple debido a la versatilidad de esta técnica y su amplio uso en el tratamiento multivariante de datos.

La elección del programa OS4 para ilustrar el desarrollo teórico del texto fue debida a tres razones: Primera, es un programa gratuito disponible en la página web del autor; segunda, no existe un programa gratuito, y en algunos casos no gratuito, tan completo como éste; y tercera, la interfaz y la presentación de los resultados son similares a otros paquetes estadísticos de amplia difusión, por lo que el aprendizaje y la interpretación del análisis estadístico con OS4 son fácilmente extrapolables a esos otros entornos.

Por último, si bien gran parte del primer capítulo está dirigido a los investigadores que trabajan en Ciencias Sociales, el resto de capítulos tienen utilidad tanto para estos investigadores como para los relacionados con las Ciencias Experimentales. En este sentido, aunque la mayoría de los ejemplos están basados en datos de tipo sociológico son fácilmente aplicables a otras situaciones. Así, el modelo estadístico que explica la disposición de un visitante a pagar o no una entrada a un parque natural en función de su nivel de ingresos y la distancia que recorre es equivalente a estudiar la recuperación o no de un paciente según la dosis de un fármaco y su edad.

LISTA DE TABLAS

LISTA DE FIGURAS

La formulación de un problema es más importante que su solución
(Albert Einstein)

CAPÍTULO 1

OBTENCIÓN DE DATOS PRIMARIOS

Capítulo 1. Obtención de datos primarios

1.1. Introducción

En las investigaciones sociológicas es habitual utilizar fuentes de datos primarias y secundarias. Brevemente, podemos caracterizarlas como sigue a continuación:

Datos secundarios. Datos recogidos para otros propósitos, por lo tanto, de rápido acceso y bajo coste. Como principal inconveniente presentan la limitación de su adecuación a las necesidades de nuestra investigación, en términos de diseño del muestreo y vigencia de los mismos.

Datos primarios. Son aquellos generados por el investigador con el propósito específico de la investigación que se está realizando. Tienen un coste, económico y de tiempo, muy superior a los datos obtenidos por fuentes secundarias.

Ante un problema de ámbito social o económico, la encuesta es uno de los métodos disponibles para la obtención de datos primarios. Debido a su coste, es aconsejable estudiar la posibilidad de la utilización de datos obtenidos en estudios similares, o la aproximación al fenómeno sociológico a través de variables que suponemos relacionadas. En este segundo caso, por ejemplo, podríamos estudiar el incremento de la renta disponible de los habitantes de un pueblo en los últimos cinco años a través del análisis de las ventas de coches de los concesionarios, la apertura de nuevos restaurantes, etc., en lugar de realizar una encuesta a dichos habitantes.

En la realización de una encuesta hay dos elementos claramente diferenciados: *El tipo de muestreo estadístico y el método de recogida de la información*. El primero hace referencia al número de sujetos que serán encuestados y cómo se seleccionarán, mientras el segundo explica el método de intercambio de información entre el encuestado y el encuestador.

1.2. Justificación del muestreo estadístico

En primer lugar cabe preguntarse ¿por qué estudiar una muestra en lugar de toda la población? Podemos apuntar las siguientes razones:

Coste. Existe una limitación de recursos, en términos de dinero, tiempo o esfuerzo. Una muestra más pequeña cuidadosamente seleccionada puede darnos una información suficientemente próxima a la realidad con el consiguiente ahorro de tiempo y de dinero. A partir de un punto, el incremento de precisión que se obtiene con una muestra mayor no compensa el incremento del coste.

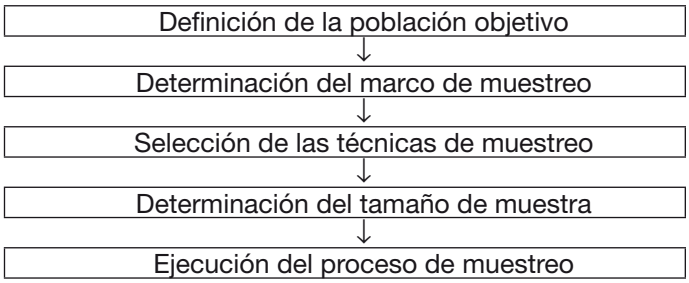
Utilidad. En algunas ocasiones el proceso de muestreo destruye el elemento, por ejemplo cuando se mide la duración de una bombilla.

Accesibilidad. Algunos elementos de la población pueden ser totalmente inaccesibles. Por ejemplo, si queremos estudiar zonas rurales de sequía puede no haber registros pluviométricos para todas las zonas en el periodo considerado, imposibilitando tener todos los datos para la población objetivo (todo el territorio).

1.3. Etapas del muestreo estadístico

En el diseño del muestreo estadístico existen cinco etapas, cada una estrechamente interrelacionada, como muestra la siguiente figura.

Figura 1.1. *Etapas en el diseño del proceso de muestreo*



Definición de la población objetivo

La población objetivo es el conjunto de elementos que poseen la información que es necesario recabar. La población objetivo determina quién puede o no ser incluido en la muestra.

Determinación del marco de muestreo

El marco de muestreo es una representación de la población objetivo. Cuando el marco de muestreo no coincide con la población objetivo se comete un error de marco de muestreo. Se expresa como: 1- ($N_{marco\ muestral}$ / $N_{población\ total}$). Por ejemplo, si queremos analizar las pautas de consumo de una población de 10.000 habitantes y utilizamos la guía telefónica que incluye a 2.400 hogares (9.600 personas, según el censo), nuestro error de marco de muestreo será del 4% (9.600/10.000=0,96 $\rightarrow$ 1-0,96=0,04).

Técnicas de muestreo

Las diferentes técnicas de muestreo pueden clasificarse en probabilísticas y no probabilísticas. Esta distinción hace referencia a la posibilidad de saber o no *a priori* cuál es la probabilidad de cada elemento de la población de ser seleccionado en la muestra[1]. Podemos clasificar estas técnicas en dos tipos:

Técnicas no probabilísticas:

1. Muestreo de conveniencia: Los encuestados son seleccionados porque estaban en el lugar preciso en el momento adecuado. Ejemplo: Se entrevista a todos los asistentes a un curso.
2. Muestreo de juicio: Los encuestados son seleccionados siguiendo el criterio del investigador, basándose en su conocimiento de la población objetivo. En esta ocasión el investigador selecciona a encuestados que, a su juicio, representan todo el espectro de la población objetivo.

3. Muestreo encadenado: Se selecciona a una muestra que sirve como punto de partida para otra muestra. Se utiliza cuando, por la naturaleza delicada de la pregunta o la dificultad de encontrar a los encuestados, es necesario que el encuestado nos dirija a otro. Un ejemplo típico de la utilidad de este método sería la investigación sobre hábitos de conducta moralmente no aceptados por la sociedad.

4. Muestreo por cuotas: En este caso el investigador tiene una información más detallada de la distribución de la población según algunas variables que están relacionadas con la variable a estudiar. De acuerdo con estas variables se divide la población en estratos y se entrevista un número determinado en cada estrato. Así, si suponemos que en un 75 por ciento de las ocasiones es la mujer quien decide la ropa infantil, sería conveniente incluir en nuestra muestra una cuota de mujeres similar. Si además pensamos que la edad es también importante, podemos clasificar la población por sexo y edad, reflejando en cada celda (por ejemplo mujer menor de 30 años) su importancia relativa.

Técnicas probabilísticas[2]:

1. Muestreo aleatorio simple (MAS) Cada elemento de la población y cada posible muestra de un tamaño *n* tienen una probabilidad conocida e igual de ser seleccionados. Esta técnica requiere tener un listado exhaustivo de todos los elementos de la población objetivo. Un caso típico sería la utilización de la guía telefónica. En este caso, sin embargo, habría que analizar si la representatividad de la muestra se ve afectada por la imposibilidad de seleccionar a aquellos ciudadanos sin número de teléfono (en un estudio sobre el nivel de pobreza claramente sí, en otro sobre el consumo de bienes de lujo no).

2. Muestreo sistemático: Se elige un primer elemento aleatoriamente y a continuación todos los siguientes cada *n* posiciones. Así, si por ejemplo queremos seleccionar 20 individuos de entre 100, el primer paso es seleccionar un número aleatorio entre 1 y 5 (ya que 100/20=5), digamos el 3; a continuación seleccionamos los individuos 3, 8, 13, 18,... 93, 98. Esto implica la ordenación de todos los elementos de la población, si bien el criterio de ordenación no debe guardar ninguna relación con el fenómeno sociológico a estudiar. Así, por ejemplo, en un estudio sobre hábitos de consumo podemos ordenar los elementos del marco de muestreo por orden alfabético, pero no por el número del DNI debido a su relación con la edad del sujeto, la cual, a su vez, está relacionada con los hábitos de consumo.

3. Muestreo estratificado: Requiere al menos dos etapas. En una primera etapa la población objetivo se divide en estratos según las variables que se consideran relacionadas con el fenómeno sociológico a estudiar. Por ejemplo, queremos estudiar la importancia de la caza en una provincia: Suponemos que el tamaño del municipio tiene una gran influencia en esta variable, así, dividimos la muestra como sigue: (a) municipios de menos de 5.000 hab., (b) entre 5.000 y 15.000, y (c) más de 15.000.

La segunda fase consiste en seleccionar aleatoriamente una muestra dentro de cada estrato cuyo número puede ser:

a. Igual para todos los estratos.
b. Proporcional al número de casos de ese estrato respecto a la población total. Ésta es la opción más frecuente cuando todos los estratos tienen una variabilidad similar.
c. Proporcional al número de casos y la desviación típica de ese estrato respecto a la población total. Cuando un estrato tiene una variabilidad muy superior al resto debemos incrementar el número de casos en ese estrato por encima de su peso proporcional. Teniendo en cuenta que en la práctica no dispondremos de la desviación típica de cada estrato de la población podemos estimarla a partir de los primeros casos de nuestra encuesta y corregir así el número final de casos en cada estrato.

El procedimiento de selección de los elementos de cada estrato puede ser aleatorio simple o cualquier otro de los anteriores.

4. Muestreo por conglomerados (*clusters en inglés*): (a) Monoetápico: La población es dividida en conglomerados (barrios, manzanas), seleccionando un grupo de ellos con probabilidad proporcional a su importancia. Una vez se tienen estos conglomerados se encuesta a todos los elementos del conglomerado (todos los vecinos de las manzanas *B, H* y *G*); (b) Bietápico: Igual que el anterior pero en lugar de encuestar a todos los elementos del conglomerado se selecciona una muestra.

Según el objetivo de la investigación y la importancia de la representatividad de la muestra con respecto a la población de la que procede, y por tanto, el nivel de recursos disponibles, optaremos por una técnica u otra. En la Tabla 1.1 se resumen las ventajas e inconvenientes de cada una de ellas.

Tabla 1.1. *Ventajas e inconvenientes de cada técnica de muestreo*

	Técnica	Ventajas	Inconvenientes
Téc. no probabilísticas	Muestreo de conveniencia	El de menor coste y tiempo de ejecución	Muestra no representativa
	Muestreo de juicio	Bajo coste y poco tiempo	Subjetividad
	Muestreo de cuotas	La muestra puede controlarse para ciertas características	No se asegura la representatividad Difícil elegir las variables que dividen la población en estratos
	Muestreo encadenado	Puede estudiar características que se consideran poco habituales	Requiere mucho tiempo
Técnicas probabilísticas	Muestreo aleatorio simple (MAS)	Fácil de entender, los resultados son proyectables a toda la población	Difícil tener el marco de muestreo Cara y de baja precisión No se asegura representatividad
	Muestreo sistemático	Puede incrementar representatividad Más fácil que el MAS	Puede reducir representatividad si un tipo de elementos se concentra en una parte del listado
	Muestreo estratificado	Incluye todas las subpoblaciones Puede incrementar la precisión	Difícil elegir las variables que dividen la población en estratos Cara
	Muestreo por conglomerados	Fácil de llevar a cabo Barata	Imprecisa; difícil de computar e interpretar los resultados

Fuente: Malhotra y Birks (1999, p. 363).

Determinación del tamaño de la muestra

La determinación del tamaño de la muestra es, posiblemente, uno de los aspectos del diseño muestral más complejos y, a la vez, más demandados por los investigadores. Para abordar el problema es necesario conocer el nivel de precisión, el nivel de confianza y el grado de variabilidad del atributo que se mide.

El nivel de precisión. También conocido como error relativo de la estimación, indica la amplitud del intervalo de nuestra estimación dentro del cual se encuentra con una determinada probabilidad el valor real (el que obtendríamos si tuviéramos datos de toda la población en lugar de una muestra).

Ejemplo. Supongamos que queremos estimar la altura media de los alumnos de un colegio, para lo cual recogemos los datos de altura de una clase con 36 alumnos. La media de esos 36 datos es 178 cm y su desviación típica 9 cm. Con una probabilidad del 95% la verdadera media, la que resultaría de medir a todos los alumnos del colegio, se encuentra dentro del intervalo:

$\bar{x} \pm t \frac{s}{\sqrt{n}}$ (se utiliza *t* en lugar de *z* cuando el tamaño de la muestra es inferior a 100)

donde $\bar{x}$ y *s* representan respectivamente la media y la desviación típica de la muestra, y *t* es el valor de la distribución t-*Student* que deja a ambos lados de la distribución un área de 2,5% (lo que equivale a un nivel de confianza del 95%). El cociente $s/\sqrt{n}$ se conoce como error estándar o error típico (valor SE en la terminología anglosajona). El producto del error estándar por el valor del estadístico (*t* o *z* según el tamaño de la muestra) es el error absoluto de la estimación. Sustituyendo los valores de nuestro ejemplo tenemos:

$$178 \pm 2{,}03 \frac{9}{\sqrt{36}} = 178 \pm 3{,}045$$

Es decir, la verdadera media se encuentra dentro del intervalo de confianza que tiene por límite inferior 175,0 (la diferencia 178-3,045) y por límite superior 181,0 (la suma 178+3,045) con un 95% de probabilidad. Los errores cometidos en nuestra estimación son:

- El error absoluto de la estimación, es decir, la amplitud del intervalo de confianza, es igual a 3,045.

- El error relativo de la estimación, o nivel de precisión, es igual a 3,045/178=0,0171, o lo que es lo mismo, un error relativo del 1,71%.

Como es natural, cuanto menor es el error relativo de la estimación mayor es su nivel de precisión. Así, por ejemplo, una estimación con un error relativo del 3% es más precisa que otra con un error del 5%. Si quisiéramos aumentar el nivel de precisión sería necesario recurrir a una muestra de mayor tamaño ya que incrementando *n* disminuye el error estándar y, consiguientemente, el error absoluto de la estimación.

Nivel de confianza. Nos informa de la probabilidad de que el valor real se encuentre dentro del intervalo de confianza, que generalmente se sitúa en el 95%.

Ejemplo. Siguiendo con el ejemplo anterior, la probabilidad de que la altura media real de los alumnos del colegio se encuentre entre 175 y 181 cm es del 95%.

Un nivel de confianza mayor (por ejemplo del 95% al 99%) implica un intervalo de confianza más amplio ya que para incrementar la probabilidad de encontrar el verdadero valor dentro de éste tendremos que ampliar su rango, lo que a su vez implica menor precisión en la estimación. Siguiendo con nuestro ejemplo, la media de la altura de los alumnos del colegio se sitúa entre 175 y 181 cm con una probabilidad del 95%, sin embargo este intervalo se amplía entre 174 y 182 cm [3] para un nivel de confianza del 99%.

Grado de variabilidad. Cuanto mayor sea la heterogeneidad de una población mayor será el tamaño de la muestra para conseguir un mismo nivel de precisión. En un caso extremo, en una población completamente homogénea, por tanto con una incidencia del fenómeno estudiado del 100%, el tamaño de la muestra se reduciría a 1 caso. El mayor grado de heterogeneidad se da cuando la característica estudiada está presente en el 50% de la población, por lo que es el valor que se usa para tener una mayor seguridad en la determinación del tamaño muestral.

Una vez explicados brevemente estos conceptos, el investigador puede seguir alguna de las estrategias que a continuación se describen para la determinación del tamaño de la muestra.

Utilización de tablas predeterminadas

Existen tablas que nos indican el tamaño mínimo de una muestra para un determinado nivel de precisión, nivel de confianza y variabilidad[4]. Es importante recordar que el número indicado hace referencia a las respuestas obtenidas. Así, en el caso de una encuesta por correo, cuya tasa media de respuestas se sitúa entre el 5 y el 10 por ciento, necesitaremos enviar más de 1000 cartas para asegurarnos conseguir nuestro objetivo de 100 elementos.

La utilización de estas tablas implica que la variable objeto de estudio tiene un formato dicotómico (es decir, variables con sólo dos valores, por ejemplo, sí/no, hombre/mujer, etc.). A continuación incluimos en la Tabla 1.2 el tamaño de la muestra según el tamaño de la población y el nivel de precisión. Se ha fijado un nivel de confianza del 95% y una variabilidad máxima (p=0,5) por ser los valores considerados en la gran mayoría de estudios.

Tabla 1.2. *Tamaño de la muestra según el tamaño de la población y nivel de precisión*

(implica la medición de un atributo dicotómico, por ejemplo respuestas sí/no)

Tamaño de la población	±3%	±5%	±7%	±10%
100	a	81	67	51
150	a	110	86	61
200	a	134	101	67
250	a	154	112	72
300	a	172	121	76
400	a	201	135	81
500	a	222	145	83
600	a	240	152	86
700	a	255	158	88
800	a	267	163	89
900	a	277	166	90
1.000	a	286	169	91
2.000	714	333	185	95
3.000	811	353	191	97
4.000	870	364	194	98
5.000	909	370	196	98
6.000	938	375	197	98
7.000	959	378	198	99
8.000	976	381	199	99
10.000	1.000	385	200	99
15.000	1.034	390	201	99
20.000	1.053	392	204	100
50.000	1.087	397	204	100
100.000	1.099	398	204	100
>100.000	1.111	400	204	100

Fuente: Israel (1992).

[a] El supuesto de distribución normal de la población no puede asumirse por tanto la muestra debe ser igual al total de la población (Yamane, 1967).

Ejemplo. De una población de 10.000 individuos obtengo una muestra de 350 casos. En dicha muestra la media de la variable analizada alcanza un valor del 60% (por ejemplo el 60% de la muestra votaría por el partido X). Por tanto, ¿Cuál sería el nivel de precisión de esta estimación?

Respuesta: Con una muestra de 350 casos proveniente de una población de 10.000, el nivel de precisión es del 7%, sin llegar al 5% ya que para ello necesitaríamos una muestra de 385 individuos.

¿Qué intervalo de esta variable tiene una probabilidad del 95% de contener la media de la población?

Respuesta: El valor de la media de la muestra es del 60%, por tanto el valor real, el que se obtendría si pudiéramos analizar toda la población, se sitúa en el 95% de los casos en un intervalo de amplitud 4,2 (el 7% de 60), es decir, comprendido entre 55,8 y 64,2.

Utilización de fórmulas para el cálculo del tamaño muestral

Para el cálculo del tamaño de la muestra que aparece en la tabla anterior se ha utilizado la siguiente fórmula:

$$n = \frac{N}{1 + N * e^2}$$

Ejemplo. El tamaño de la muestra para una población de 2.000 individuos con un error del 5% se calcula:

$$n = \frac{2.000}{1 + 2.000 * 0{,}05^2} = 333$$

Como hemos indicado anteriormente, esta fórmula (o la tabla) se puede utilizar cuando la variable que estamos analizando sólo toma dos valores (dicotómica), es decir, el encuestado presenta o no el atributo que se mide (por ejemplo una respuesta sí/no).

En el caso de querer estimar un valor medio de una variables métricas (por ejemplo la altura o el consumo medio por hogar) es necesario incluir en la fórmula la varianza de la población. En estos casos, el tamaño de la muestra se calcula como sigue:

$$n = \frac{z^2 \sigma^2}{e^2}$$

z= valor de la distribución normal (igual a 1,96 para un nivel de confianza del 95%).
σ^2= varianza de la población.
e= error de muestreo (en valor absoluto, no en tanto por uno)

Para la estimación de la varianza de la población, dato que generalmente se desconoce, se pueden seguir las siguientes estrategias:

1.- Utilizar la estimación de la varianza de la población de otros estudios similares.

2- Utilizar la cuasivarianza de la muestra piloto como estimación de la varianza de la población. Teniendo en cuenta que siempre es recomendable comprobar la validez del cuestionario en una muestra reducida, podemos calcular la cuasivarianza muestral como sigue:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

Ejemplo. Calcular el tamaño muestral para estimar el número medio de horas que los niños de una ciudad pasan frente a la televisión los sábados, a partir de los siguientes datos:

En una encuesta piloto a 18 niños de una clase se obtuvieron los siguientes datos

(minutos): 60, 60, 90, 90, 120, 120, 120, 150, 150, 180, 210, 210, 210, 240, 240, 270, 270 y 300

s^2 = cuasivarianza muestral=5.697 (aplicando la fórmula anterior, donde 172)

Nivel de confianza =95 por ciento (por tanto z =1,96).

Margen de error =15 minutos

$$n = \frac{1{,}96^2 * 5.697}{15^2} = 97{,}3 \cong 98$$

3.- También es posible estimar la varianza de la población dividiendo por 6 el rango de la variable en la muestra. Esta estimación se apoya en el hecho de que el rango en una distribución normal es igual a ± 3 veces la desviación típica de dicha distribución.

Si la muestra supone el 10% o más de la población es necesario corregir el tamaño muestral aplicando el factor de corrección de población finita.

Ejemplo. Siguiendo con el ejemplo anterior, si el número total de niños del colegio fuera de 800 la muestra inicial superaría el 10% de dicha población (98/800=0,12), por lo tanto, el tamaño corregido de la muestra debería calcularse como sigue a continuación:

$$n_c = \frac{nN}{N+n-1} = \frac{98*800}{800+98-1} = 87{,}4 \cong 88$$

Como podemos ver, el tamaño de la muestra se ha reducido de 98 a 88 alumnos.

A partir de la nueva varianza y media muestral de los 88 alumnos podemos plantearnos el problema inverso, es decir, determinar el error muestral de la nueva muestra.

Ejemplo. Imaginemos que la media de los 88 alumnos arroja un cifra de 170 minutos (frente a la media de la encuesta piloto de 172) y una varianza de 5.800 (frente a 5.697). Por tanto, tenemos que:

$$88 = \frac{1{,}96^2 * 5.800}{e^2} \rightarrow$$ e =15,9 16 ≅minutos

El error de muestreo es superior al margen inicial de 15 minutos porque nos hemos quedado cortos en la estimación de la varianza de la población (en la encuesta piloto era igual a 5.697 y en la encuesta principal es 5.800).

Por tanto, el número de minutos que los niños pasan viendo la televisión se sitúa entre 170±16, es decir, entre 154 y 186 minutos. El valor real se localiza dentro de este intervalo con una probabilidad del 95%.

En el caso de calcular el tamaño muestral para estimar una proporción la fórmula a utilizar es:

$$n = \frac{\pi(1-\pi)z^2}{e^2}$$

z= valor de la distribución normal (igual a 1,96 para un nivel de confianza del 95%).
π= proporción obtenida en otro estudio o en la encuesta piloto.
e= error muestral (en este caso, en tanto por uno)

Ejemplo. Se desea calcular el tamaño de la muestra para estimar la proporción de mujeres que trabajan diariamente 10 horas o más con un nivel de confianza del 95% (por tanto z=1,96) y un error muestral del 10%.

z= 1,96

e= 0,1

π= en un estudio piloto esta proporción fue de 0,30

$$n = \frac{0{,}30*(1-0{,}30)*1{,}96^2}{0{,}1^2} = 80{,}7 \cong 81$$

Al igual que antes, si la muestra supone el 10% o más de la población es necesario corregir el tamaño muestral aplicando el factor de corrección de población finita.

Representatividad de la muestra en muestreos estratificados o por cuotas

Siendo el muestreo estratificado y el muestreo por cuotas dos de los procesos de muestreo más utilizados en las Ciencias Sociales, sólo diferenciados por la forma en que se seleccionan los elementos en cada estrato/cuota, resulta conveniente determinar el grado de validez de la muestra en el caso de que coincidan las frecuencias observadas de la muestra con las frecuencias teóricas de la población objetivo. Para esta validación podemos utilizar la prueba de asociación de Chi-cuadrado. Veamos un ejemplo:

Supongamos que deseamos estudiar la disposición a pagar una entrada a un espacio natural protegido en una provincia. Según el juicio del investigador, basado en estudios previos en otras provincias, esta disposición está relacionada con la edad del entrevistado y el tamaño de la ciudad en la que reside. Utilizando los datos del padrón municipal de la provincia, y teniendo en cuenta que hemos fijado el tamaño de la muestra en 400 individuos, las frecuencias teóricas que debería tener nuestra muestra serían:

Frecuencias esperadas	<5.000 hab.	5.000 – 15.000	>15.000 hab.
Menos de 35 años	13	15	45
35-50 años	11	19	48
51-65 años	22	35	53
Más de 65 años	29	41	69

Sin embargo, debido a ciertas dificultades en el proceso de muestreo[5] estratificado/ por cuotas, las frecuencias obtenidas en cada estrato/cuota de la muestra no son las señaladas anteriormente sino las que se muestran a continuación:

Frecuencias observadas	<5.000 hab.	5.000 – 15.000	>15.000 hab.
Menos de 35 años	10	12	49
35-50 años	13	23	48
51-65 años	19	34	54
Más de 65 años	29	38	71

Para comprobar si la muestra puede considerarse como estadísticamente representativa utilizamos el estadístico χ^2 (Chi-cuadrado), cuya formulación es: $\chi^2 = \sum [(E_i - O_i)^2 / E_i]$, donde E_i son las frecuencias esperadas y O_i las observadas. Secuencialmente tenemos:

Diferencias: $(E_i - O_i)$	<5.000 hab.	5.000 – 15.000	>15.000 hab.
Menos de 35 años	3	3	-4
35-50 años	-2	-4	0
51-65 años	3	1	-1
Más de 65 años	0	3	-2

Cuadrados: $(E_i - O_i)^2$	<5.000 hab.	5.000 – 15.000	>15.000 hab.
Menos de 35 años	9	9	16
35-50 años	4	16	0
51-65 años	9	1	1
Más de 65 años	0	9	4

Cuad. / Esp.: $(E_i - O_i)^2 / E_i$	<5.000 hab.	5.000 – 15.000	>15.000 hab.
Menos de 35 años	0,69	0,60	0,36
35-50 años	0,36	0,84	0,00
51-65 años	0,41	0,03	0,02
Más de 65 años	0,00	0,22	0,06

La suma de la última tabla (0,69+0,60+...0,22+0,06) es igual a 3,59. El valor crítico de una Chi-cuadrado de (3-1)*(4-1) -ya que tenemos 3 categorías en el tamaño de la ciudad y 4 niveles de edad- para un nivel de significación del 5% es igual a 12,59 (ver Anexo 2). Teniendo en cuenta que el valor calculado del estadístico (3,59) es inferior al valor crítico (12,59) no rechazamos la hipótesis nula de no diferencia entre los dos grupos (la población y la muestra), por lo que aceptamos la muestra como una representación estadísticamente válida de la población objetivo.

1.4. Métodos de recogida de información

Cuando la información que se necesita no está disponible en fuentes secundarias (archivos, otros trabajos, etc.) y se descarta la observación directa sin interacción con los individuos (por ejemplo mediante cámaras ocultas en los centros comerciales) tenemos que recurrir a la encuesta. En la fase anterior hemos decidido la técnica de muestreo y el tamaño de la muestra. Ahora es necesario determinar cómo vamos a recoger la información. Tras la elaboración de un cuestionario, en la mayoría de los estudios se opta por una de las siguientes opciones[6]:

Entrevista cara a cara

Ventajas: Reduce la posibilidad de mal interpretación de las preguntas y aumenta el número de respuestas.

Inconvenientes: Es el método más caro. Se requiere una formación de los encuestadores para evitar que orienten al encuestado hacia una respuesta. Este sesgo puede ocurrir voluntaria o involuntariamente. Otra fuente de error es la llamada "respuesta socialmente correcta", en la cual, el entrevistado responde lo que el encuestador espera o lo que socialmente se acepta como normal.

Encuestas por correo

Ventajas: Es el método más barato. Permite al encuestado pensar las respuestas y buscar información si fuera necesario. Garantiza el anonimato.

Inconvenientes: Baja tasa de respuestas (en general, inferior al 10%). Dificultad en determinar si los que no responden tienen una característica común que afecta a la variable que estamos estudiando. Necesidad de un listado. El tiempo de respuesta puede ser muy largo (incluso de varios meses si se envía una carta recordatoria para incrementar la tasa de respuestas).

Encuestas por teléfono

Ventajas: No es demasiado caro. Es el método más rápido.

Inconvenientes: Puede no cubrir toda la población objetivo. Por ejemplo, cometeríamos un grave error utilizando este método para medir la pobreza extrema debido a que, probablemente, gran parte de estos individuos no posean teléfono. De igual forma, muchos nuevos hogares hoy en día no disponen de teléfono fijo sino de uno o varios teléfonos móviles por lo que no podrían ser seleccionados en la muestra a partir de, por ejemplo, un directorio telefónico. Siguiendo con este problema, imaginemos que queremos analizar las pautas de consumo de la población más joven a través de una encuesta por teléfono. Si tenemos pruebas (o al menos indicios razonables) de que los jóvenes que tienen teléfono fijo no se diferencian en sus pautas de consumo de los que sólo tienen teléfonos móviles, no cometeríamos ningún error al no incluir éstos últimos en la muestra. Sin embargo, ¿no es posible que los no incluidos, los jóvenes sin teléfono fijo, no lo tengan porque pasan menos tiempo en casa y esto influyera en sus pautas de consumo (por ejemplo consumiendo más platos precocinados)?

Otro inconveniente de la encuesta por teléfono es el riesgo de interpretación errónea de las preguntas, lo que limita el cuestionario tanto en su dificultad como en su duración. Asimismo, al igual que en las entrevistas cara a cara, el encuestado puede sentirse inclinado a responder "correctamente" por sentirse identificado.

Tabla 1.3. *Comparación de los distintos métodos de recogida de información*

Característica	Cara a cara	Correo	Teléf.	E-mail
Probabilidad de localizar individuo seleccionado	Media	Alta	Alta	Alta
Tasa de respuestas	Alta	Baja	Alta	Media
Posibilidad de un cuestionario largo	Alta	Media	Baja	Baja
Posibilidad de preguntas complejas	Alta	Media	Baja	Media
Éxito con preguntas abiertas	Alto	Bajo	Alto	Medio
Éxito con preguntas tediosas o aburridas	Alto	Bajo	Medio	Bajo
Probabilidad de evitar preguntas sin respuestas	Alta	Media	Alta	Media
Probabilidad de evitar respuestas "socialmente deseables"	Baja	Alta	Media	Baja
Probabilidad de evitar la distorsión del entrevistador	Baja	-	Media	Media
Probabilidad de encontrar entrevistador cualificado	Baja	-	Alta	-
Probabilidad de realizar consultas cuando son necesarias	Media	Alta	Baja	Alta
Probabilidad de rapidez de realización	Baja	Baja	Alta	Alta
Probabilidad de mantener los costes bajos	Baja	Alta	Media	Alta

Fuente: Adaptado de Dillman (1978) y extendido para el correo electrónico.

1.5. Diseño de un cuestionario

Aspectos previos

Antes de diseñar el cuestionario es fundamental tener muy claro qué se pretende medir. En algunas ocasiones será posible medir la variable sobre la que queremos información directamente, por ejemplo, medir la altura media de una población. En otros casos tendremos que fabricarnos un índice que sirva de medida, por ejemplo, medir la satisfacción de un cliente. En cualquier caso, el proceso de diseño de un cuestionario es decisivo en el éxito de nuestro estudio. Nuestros resultados serán tan buenos como la calidad de los datos (este será el límite máximo).

En el diseño de un cuestionario es necesario anticipar cualquier situación o respuesta, no dejando duda al encuestador sobre cómo transferir la información del encuestado al papel (en el caso de entrevista cara a cara o por teléfono), y no dejando la posibilidad de más de una interpretación de la pregunta (en el caso de la encuesta por correo). Como estrategia general del diseño del cuestionario, especialmente en las encuestas por correo, se suele comenzar con preguntas fáciles que no comprometen al encuestado.

Una práctica aconsejable, y obligada en estudios más rigurosos, es la realización de una encuesta piloto para comprobar la validez del cuestionario. El tamaño de esta encuesta piloto debería ser el 5-10 por ciento del tamaño de la encuesta principal.

Redacción de las preguntas

Si existe alguna posibilidad, por pequeña que sea, de interpretar una pregunta en más de un sentido, tenga la completa seguridad de que un porcentaje de los encuestados responderá a la pregunta equivocada. Para reducir los posibles problemas en la elaboración de un cuestionario, a continuación señalamos algunos aspectos que deberían tenerse en cuenta[7]:

(a) Ambigüedad. El vocabulario utilizado debe guardar una relación con el nivel cultural del encuestado. Deben ser claras sin posibilidad de duda. Por ejemplo: "¿Conoce usted la marca A y B? □ Sí □ No". Si hay individuos que conocen A pero no B en estricta lógica habría que responder "No", sin embargo un porcentaje indeterminado optará con seguridad por la respuesta "Sí". Por tanto, para reducir el error de interpretación, habría que formular dos preguntas por separado. Otro ejemplo: "Usted es consumidor habitual de alcohol: □ Sí □ No". ¿Es *habitual* en el sentido de frecuencia, de cantidad, o de ambos?. Claramente puede mejorarse esta pregunta indicando alguna cantidad semanal o varios intervalos.

(b) Neutralidad. Las preguntas deben ser neutrales, sin inclinar hacia un sentido u otro.

(c) Recelo Se deben evitar preguntas que pongan a la defensiva al encuestado. Son típicas las preguntas relativas a los niveles de ingreso y nivel cultural. Como estrategia general para reducir el rechazo del encuestado se suelen incluir intervalos amplios (muchos encuestados no declararían que su nómina mensual es Y pero quizás no les importaría indicar que se sitúa en un intervalo entre X y Z). Por este motivo, este tipo de preguntas es aconsejable dejarlas para el final del cuestionario.

(d) Privacidad. En el caso de preguntas sensibles (por ejemplo, sobre actividades socialmente rechazadas como el juego, la prostitución, etc.), puede buscarse un indicador relacionado con la variable original que no presente problemas de respuesta.

(e) Exclusividad. Las categorías posibles en la respuesta deben ser mutuamente excluyentes. Por ejemplo, en la pregunta "Edad: □ <35 □ 35-50 □ 50-65 □ >65" ¿dónde se sitúa alguien con exactamente 50 años?

(f) Exhaustividad. Las categorías posibles en la respuesta deben incluir cualquier opción que el encuestado hubiera deseado marcar. Un ejemplo claro consiste en dejar la posibilidad al encuestado de señalar "No sé", "No contesta", "No me importa", etc. Asimismo, es conveniente separar estas posibilidades en lugar de agruparlas en la tradicional respuesta "No sabe / No contesta" (ya que la motivación, y por tanto su interpretación, es distinta).

(g) Ordenación de preferencias No se debe pedir al encuestado que ordene más de 4 ó 5 alternativas. Si se incrementa este número corremos un riesgo serio de que se ordenen de cualquier manera para pasar a la siguiente pregunta.

Preguntas con intervalos numéricos

Este tipo de pregunta es útil para obtener información sobre cuestiones de las cuales el encuestado prefiere no dar una respuesta exacta (un ejemplo típico es el nivel de ingresos) o no es capaz de dar una cifra exacta. En estos casos se le pide que marque el intervalo que mejor le define. A la hora de crear estos intervalos se nos plantean tres aspectos:

- La amplitud de los mismos.
- Los puntos de corte inicial y final de la escala.
- El número de intervalos.

Respecto a la primera cuestión, cuanto menor es la amplitud de los intervalos mejor definimos a la población. Sin embargo, una amplitud demasiado pequeña puede hacer que el encuestado se sienta incómodo o, peor aún, que no responda. Así, por ejemplo, si un ciudadano tiene unos ingresos mensuales de 2.360 euros, probablemente no le resultaría incómodo marcar el intervalo 2.000-2.500 pero sí marcar el 2.300-2.400.

En relación con los extremos de los intervalos, si bien no existe una recomendación clara a la hora de definir sus límites, inferior y superior, la práctica recomienda diseñar este tipo de escala de tal forma que el número de casos que se encuentran en los dos extremos sea bajo, digamos un 10 por ciento sumando ambos[8]. Para evitar un número excesivo de casos en los extremos, con la dificultad que conlleva la estimación de un valor representativo para ese intervalo (¿cero para el límite inferior y un múltiplo de la amplitud para el superior?), es aconsejable conocer a *priori* los límites en los que se mueve la población y ajustar los extremos de los intervalos en consecuencia.

Por último, el número de intervalos vendrá determinado por los dos aspectos anteriores. En cualquier caso es recomendable, por las razones que veremos más adelante, que su número no sea inferior a 5. En general, este número se sitúa entre 5 y 7.

1.6. Codificación de las variables

A la hora del diseño del cuestionario, otro factor a tener en cuenta hace referencia al tratamiento estadístico posterior de los datos. Así, según el formato de la pregunta, por ejemplo pidiendo una cantidad exacta o dando varias opciones para marcar, se condiciona y limita el tipo de técnica estadística a utilizar. Por tanto, es necesario considerar los siguientes aspectos:

(a) Codificación. La codificación traslada la información alfanumérica a numérica (por ejemplo sexo: hombre=1; mujer=2). También es habitual un código para las preguntas en blanco (No sabe / No contesta), por ejemplo el 99. Cuando las variables son ordinales (por ejemplo, nivel de estudios) es conveniente adjudicar el mayor código al encuestado de mayor nivel (por ejemplo, ninguno=0, primarios=1, secundarios=2 y universitarios=3). En el caso de respuestas binarias “sí / no” se suele considerar sí=1 y no=0.

(b) Escala. Por ejemplo, supongamos una pregunta con las siguientes opciones:

a. Absolutamente de acuerdo.
b. Ligeramente de acuerdo.

c. Indiferente.
d. Ligeramente en desacuerdo.
e. Absolutamente en desacuerdo.

Podría utilizarse una escala 6, 4, 3, 2, 0, respectivamente, en lugar de 5, 4, 3, 2, 1, ya que parece que hay una mayor distancia entre “Absolutamente de acuerdo” y “Ligeramente de acuerdo”, que entre “Ligeramente de acuerdo” e “Indiferente”. Si la opción a. se formulara como “Bastante de acuerdo”, podríamos optar por la codificación 5, 4, 3, 2, 1.

(c) Edición de los datos. Es necesaria la comprobación de los datos recogidos con el fin de corregir los posibles errores de trascripción del encuestado (o encuestador en el caso de entrevista personal o por teléfono) al cuestionario y de éste al ordenador. La mayoría de paquetes estadísticos disponen de esta función. Como alternativa, cualquier hoja de cálculo ordena los datos y permite comprobar si el rango es el correcto.

La ciencia más útil es aquella cuyo fruto es el más comunicable
(Leonardo Da Vinci)

CAPÍTULO 2

TIPOS DE VARIABLES Y ALTERNATIVAS DE ANÁLISIS

Capítulo 2.
Tipos de variables y alternativas de análisis

2.1. Introducción

Un primer paso a la hora de analizar datos es conocer la naturaleza de las variables con las que se trabaja, es decir, la escala de medida de las mismas. En el caso de la búsqueda de relaciones o diferencias entre variables (análisis bivariante o multivariante), esta escala determina el tipo de análisis estadístico a realizar. De igual manera, la descripción de las variables (análisis univariante) utiliza diferentes procedimientos según su naturaleza.

2.2. Tipos de variables

Podemos clasificar las escalas de medida de las variables en tres tipos[9]:

- Variables nominales.
- Variables ordinales.
- Variables métricas

Variables nominales

Las categorías no implican ninguna ordenación, simplemente identifican al elemento. Por ejemplo, el encuestado prefiere la marca A, B, C o D (si bien en la codificación utilizaremos *marca*=1, 2, 3 o 4).

Si tenemos sólo dos categorías hablamos de variables dicotómicas (o binarias), como es el caso del SEXO (hombre o mujer). Una práctica habitual en este último caso consiste en utilizar el valor 1 para una de las categorías y el 0 para la otra (hombre=1, mujer=0, o viceversa). Aunque la asignación del 0 y del 1 es arbitraria, en algunos casos, como por ejemplo en el análisis de regresión, facilita la interpretación de los coeficientes asignar el 1 a la categoría que influye positivamente en la variable que queremos explicar y 0 a la otra, así, por ejemplo, en un estudio que busque la relación entre sexo y peso corporal asignamos el 1 a los hombres y el 0 a las mujeres, ya que el peso medio de los hombre es superior al de las mujeres.

Variables ordinales

Indican un orden creciente o decreciente pero no la magnitud de las diferencias entre las categorías. Por ejemplo, al final de la Liga de fútbol los equipos se ordenan según su posición pero no se tiene en cuenta la diferencia de puntos entre ellos. De igual forma, en una carrera de coches sólo hablamos de posiciones finales y no de tiempos de llegada. Existen pruebas estadísticas específicas basadas exclusivamente en la ordenación de los elementos, lo cual resulta muy útil en estudios con información limitada.

Variables métricas

Las variables métricas representan la jerarquía máxima dentro de las escalas de medida ya que permiten realizar cualquier tipo de operación matemática. Incluyen dos subcategorías:

Variables de escala por intervalos

Permiten la ordenación y la cuantificación de las distancias entre dos observaciones. Existen dos diferencias con respecto a las variables de ratio:

- *Cero arbitrario*. Las variables de escala por intervalo fijan el cero de forma arbitraria. Las escalas de temperatura (Kelvin, Fahrenheit o Celsius) son un ejemplo típico de este tipo de variables. Así, es necesario explicar la escala de la medición ya que no es lo mismo 70º Kelvin, Celsius o Fahrenheit. Por el contrario, nadie preguntaría en qué escala se miden 4 kg o 38 km (ambas variables de ratio). Otro ejemplo es la puntuación obtenida en un examen: Si es de 4,3, no es lo mismo en una escala de 0 a 10 que en otra de 1 a 5.
- *Proporcionalidad* entre dos valores. En las variables de escala por intervalo un valor *n* veces mayor no implica *n* veces más intensidad del concepto que representa. Por ejemplo, el doble de 35ºF es 70ºF pero esto no implica que el calor sea el doble, para comprenderlo basta con transformar estas temperaturas a la escala Celsius (1,7ºC y 21,1ºC, respectivamente). De igual forma, si los resultados de dos estudiantes en un examen son 4,3 y 8,6, no se concluye que el segundo tiene el doble de conocimientos que el primero.

Variables de ratio

También conocidas como variables de proporción o razón. En estas variables el cero está definido y es fijo. La magnitud de la distancia entre dos valores no depende de dichos valores. Para comprender estos dos aspectos consideremos la variable métrica distancia (por ejemplo en km). Para esta variable no es necesario definir dónde está el cero (0 km) y la magnitud de la distancia entre dos puntos no depende de estos puntos: Entre 3 y 7 km hay 4 km, los mismos que entre 16 y 20 km.

Desde un punto de vista práctico, el tratamiento estadístico de las variables de escala por intervalos y las de ratio es el mismo. Por este motivo, el análisis estadístico desarrollado en este manual se limita a clasificar a las variables como nominales, ordinales o métricas.

2.3. Consideración métrica de variables ordinales

Utilidad y condiciones

Bajo determinados supuestos es posible tratar una variable ordinal como una métrica y por tanto ampliar el abanico de técnicas de análisis estadístico disponibles. Como veremos en el siguiente apartado, la escala en que se mide la variable (nominal, ordinal o métrica) condiciona el tipo de análisis estadístico. En este sentido, las variables métricas son las que proporcionan más posibilidades al analista[10], por ello resulta

interesante explorar cuándo es posible utilizar estas técnicas aunque sólo tengamos variables ordinales. En concreto nos referimos al caso de las variables definidas por intervalos numéricos, las cuales, según la amplitud de los mismos, podemos dividirlas en dos grupos:

- La amplitud de los intervalos no es la misma. Por ejemplo, definimos los intervalos de ingresos como sigue: <1.000, 1.000-1.200, 1.201-1.600, 1.601-2.000 y >2.000, lo cual implica que la amplitud de cada intervalo es 1.000, 200, 400, 400 y >1.000 (presumiblemente algún encuestado ganará más de 3.000), respectivamente. En este caso, el tratamiento estadístico de esta variable sería el mismo que el de una variable ordinal.
- Igual amplitud en cada intervalo. Si la amplitud de los intervalos es la misma y el número de éstos es superior o igual a 5 (por ejemplo: <500, 500-1.000, 1.001-1.500, 1.501-2.000 o >2.000), el error que cometemos al tratar esta variable como métrica (con los valores 1, 2, 3, 4 y 5) es mínimo desde un punto de vista práctico, como sugieren Lobovitz (1967 y 1970), Kim (1975), Binder (1984), Givon y Shapira (1984), Crask y Fox (1987), Zumbo y Zimmerman (1993) y Jaccard y Wan (1996), entre otros autores.

Ejemplo de variable definida por tramos

En este ejemplo ponemos de manifiesto la importancia de un diseño adecuado de los intervalos así como la ventaja de utilizar una variable ordinal como una métrica. Para ello supongamos que tenemos una población hipotética de profesionales definida por dos variables: Ingresos y años de experiencia profesional. La primera se recogerá como variable ordinal mediante intervalos de ingresos, la segunda como variable métrica. Los datos de ingresos (ING) y experiencia (EXP) de la población se presentan en la Tabla 2.1 junto con dos alternativas posibles de intervalos identificadas como:

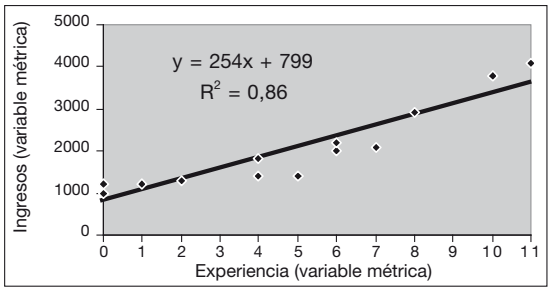
- ING -1 (<1.200, 1.200-1.600, 1.601-2.000 y >2.000; en la tabla como 1, 2, 3 y 4)
- ING -2 (<1.300, 1.300-1.800, 1.801-2.300, 2.301-2.800, 2.801-3.300 y >3.300; en la tabla como 1, 2, 3, 4, 5 y 6).

Tabla 2.1. *Ejemplo de creación de variables ordinales*

EXP	0	0	1	1	2	4	5	4	6	7	6	8	10	11
ING	1000	1200	1200	1300	1300	1400	1400	1800	2000	2100	2200	2900	3800	4100
ING -1	1	2	2	2	2	2	2	3	3	4	4	4	4	4
ING -2	1	1	1	2	2	2	2	2	3	3	3	5	6	6

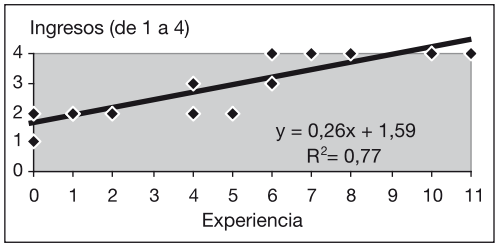
En la tabla anterior las dos filas sombreadas representan los valores reales de experiencia profesional e ingresos mensuales. Esta relación (entre EXP e ING) puede observarse gráficamente en la Figura 2.1.

Figura 2.1. *Ingresos medidos mediante variable métrica*

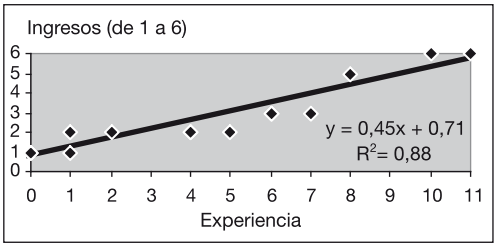


Como refleja el gráfico anterior, por cada año de experiencia profesional el sueldo se incrementa en 254 euros/mes. ¿A qué conclusiones hubiéramos llegado si, en lugar de utilizar el valor exacto (ING), hubiéramos recogido los valores de ingresos mediante las variables ordinales ING-1 o ING-2 (ver las dos últimas filas de la Tabla 2.1)? Veamos gráficamente ambos casos:

Figura 2.2. *Ingresos medidos mediante variables ordinales con diferentes intervalos*



Variable ingresos=*ing*-1



Variable ingresos=*ing*-2

En el primer caso, con cuatro intervalos, un año de experiencia extra supone un incremento de sueldo de 0,26 veces la amplitud del intervalo, es decir, 0,26*400=104 euros/mes. En el segundo caso, con seis intervalos, ese mismo año extra supone un incremento de 0,45*500=225 euros/mes.

Claramente la utilización de la escala ING-2 estima con mucha mayor precisión el incremento real de sueldo por cada año de experiencia. ¿Cuáles son las razones detrás de las diferencias de comportamiento de ambas escalas? Podemos apuntar:

1. La variable ING-1 tiene pocos intervalos. En efecto, para el tratamiento de este tipo de variables de forma similar al tratamiento de una variable métrica es aconsejable contar con 5 o más intervalos.

2. En la variable ING-1, la distancia entre la cota superior del mayor intervalo (2.000) y el valor máximo de los datos (4.100) excede con mucho la amplitud de los intervalos (400). Teniendo en cuenta que era probable encontrar profesionales con un sueldo muy superior a 2.000 hubiera sido aconsejable añadir intervalos superiores.

Este ejemplo pone de manifiesto el error que supone la utilización de la variable ordinal ING-1 como una variable métrica (la variable dependiente de la regresión) cuando no se cumplen los requisitos antes expuestos. Por el contrario, la variable ordinal

ING-2 sí permite esta consideración, por tanto, las estimaciones obtenidas son una aproximación aceptable de la realidad.

2.4. Tratamiento estadístico según el tipo de variable

Análisis univariante

El análisis univariante, como su nombre indica, estudia la distribución individual de cada variable. Este análisis se centra en dos aspectos: La tendencia central de la distribución y su dispersión. En el primer caso hablamos de un valor característico o medio de la distribución, en el segundo de la variabilidad interna de los datos. Según el tipo de variables proceden los siguientes análisis:

1. Variables nominales. Para este tipo de variables el análisis se limita a las frecuencias de cada categoría. Se suele expresar en porcentajes.
2. Variables ordinales. La tendencia central se mide con los estadísticos mediana y moda (pero no la media, ya que ésta implica distancias comparables), mientras que para la dispersión podemos utilizar un diagrama de frecuencias (histograma).
3. Variables métricas. Para el análisis de la tendencia central se utiliza por lo general la media, si bien es aconsejable utilizar la mediana cuando nos encontramos con unos pocos valores extremos cuya magnitud difiere ampliamente del resto (son mucho mayores o mucho menores que la mayoría). Para estudiar el grado de dispersión recurrimos a la desviación típica o la varianza. Es posible estudiar además del momento de primer orden de la distribución (tendencia central) y el momento de segundo orden (dispersión), el momento de tercer orden (simetría) y el de cuarto orden (achatamiento). No es frecuenta estudiar momentos de orden superior.

Análisis bivariante

Para estudiar el efecto de una variable sobre otra podemos clasificar las técnicas estadísticas en dos grandes familias: Las pruebas paramétricas y las no paramétricas. Cuatro son los requisitos para el uso de las pruebas paramétricas (Field, 2000, p. 37):

1. Distribución normal de las variables. Para llevar a cabo pruebas de estadística paramétrica se asume que la variable estudiada de la población sigue una distribución normal.
2. Uniformidad de la varianza. Se requiere que la varianza de una variable no dependa del nivel de otra variable.
3. Escala de medida. Las variables deben medirse en una escala métrica.
4. Independencia. Las respuestas de un sujeto no dependen de las de otro.

Los dos últimos puntos sólo pueden contrastarse siguiendo el sentido común del investigador. Para los dos primeros sí existen pruebas objetivas.

Algunas de las pruebas más comunes utilizadas para determinar la relación entre dos variables son:

- Pruebas paramétricas: Comparación de medias (prueba *t*), análisis de correlación (Pearson) y análisis de la varianza (ANOVA I).
- Pruebas no paramétricas: Comparación de medias (Mann-Whitney), análisis de correlación (Spearman, Kendall tau), análisis de la varianza (Kruskal-Wallis) y tablas de contingencia (Chi-cuadrado, Fisher).

La tabla siguiente resume algunas de las posibilidades de análisis estadístico de dos variables. Si bien el análisis de regresión se ha incluido en las técnicas multivariantes, resulta útil en la comparación bivariante. Las alternativas se agrupan según la naturaleza de la escala de las variables objeto de análisis.

Tabla 2.2. *Alternativas de análisis bivariante según la naturaleza de las variables*

	Nominal	**Ordinal**	**Métrica**
Nominal	Chi-cuadrado Fisher (tablas 2x2)		
Ordinal	Chi-cuadrado Fisher (tablas 2x2)	Chi-cuadrado (indica existencia de relación pero no sentido) Análisis de correlación: Coeficiente de Spearman o Coeficiente de Kendall tau	
Métrica	Si la nominal es dicotómica prueba de comparación de medias: Prueba t o Mann-Whitney Con la variable nominal como factor análisis de la varianza: ANOVA I o Kruskal-Wallis Regresión con variable nominal como ficticia Si se transforma la variable métrica a escala ordinal: Chi-cuadrado (indica existencia de relación pero no sentido)	Análisis de correlación: Coeficiente de Pearson, Coeficiente de Spearman o Coeficiente de Kendall tau Si se transforma la variable métrica a escala ordinal analizar como Ordinal-Ordinal (si el análisis de correlación no revela relaciones claras)	Análisis de correlación: Coeficiente de Pearson o Coeficiente de Spearman Si se transforman ambas variables a escala ordinal analizar como Ordinal-Ordinal (si el análisis de correlación no revela relaciones claras)

Fuente: Elaboración propia y Bryman y Cramer (1997, p. 200).

En la tabla anterior hemos sombreado las pruebas no paramétricas. Como puede observarse, siempre es posible optar por una de estas pruebas cuando no se cumplen los requisitos paramétricos.

Pruebas paramétricas vs. no paramétricas

Si bien el uso de pruebas paramétricas requiere la comprobación previa de los supuestos anteriormente descritos, el poder de estas pruebas es superior al de las no paramétricas, esto es, con las pruebas paramétricas es más probable "descubrir" una relación subyacente entre variables que con las no paramétricas (Siegel, 1985, p. 54).

Desde un punto de vista práctico, aún cuando no sepamos si la variable de la población sigue una distribución normal, para una muestra de tamaño suficiente (habitualmente cuando $n>50$) podemos optar por las pruebas paramétricas. La razón hay que buscarla en el teorema central del límite. Según este teorema, aunque una variable no siga una distribución normal sí lo hacen las estimaciones de las sucesivas muestras.

Por tanto, en la búsqueda de relaciones entre dos variables podemos seguir la siguiente estrategia:

Tabla 2.3. *Tipo de prueba según la naturaleza de las variables y el tamaño muestral*

Análisis bivariante	Grupo de pruebas		
Nominal-Nominal Nominal-Ordinal Ordinal-Ordinal	Pruebas no paramétricas		
Al menos una de las dos variables es métrica	Tamaño muestral $n \geq 50$	Los datos siguen una distribución normal o no	Pruebas paramétricas
	Tamaño muestral $n<50$	Los datos siguen una distribución normal	Pruebas paramétricas
		Los datos no siguen una distribución normal	Pruebas no paramétricas

Análisis multivariante

Bajo la denominación de análisis multivariante, es decir, la consideración simultánea en el análisis de más de dos variables, aparece un amplio abanico de técnicas estadísticas y de algoritmos de cálculo. En el presente manual, sin ánimo de ser exhaustivos, hemos considerado interesante incluir:

Análisis de la varianza (ANOVA II). Se analiza el efecto de dos variables nominales (llamadas factores) sobre una variable métrica.

- *Ejemplo*: ¿Depende el rendimiento de una máquina (variable métrica) del modelo de la máquina (factor 1) y del operario que la maneja (factor 2)?

Análisis multivariante de la varianza (MANOVA). A diferencia de ANOVA, este análisis considera simultáneamente dos o más variables dependientes. En este caso, ¿por qué no utilizar ANOVA para cada una de las variables dependientes? La respuesta es simple: porque es posible que exista una diferencia entre grupos que sólo se ponga de manifiesto considerando varias variables dependientes simultáneamente.

- *Ejemplo*: Estudiando el efecto de un tratamiento sobre el periodo de recuperación de un paciente concluimos que no tiene ningún efecto sobre dicho periodo. Sin embargo, si estudiáramos este efecto de forma simultánea sobre el periodo de recuperación y la edad del paciente (dos variables métricas) mediante MANOVA pudiera ocurrir que sí hubiera un efecto estadísticamente significativo.

Análisis discriminante. La idea central de esta técnica consiste en ponderar las diferentes variables a la hora de asignar un elemento a un grupo u otro. Por tanto, la variable dependiente es nominal mientras que las explicativas pueden ser de cualquier tipo.

- *Ejemplo*: ¿Podríamos predecir el tipo de coche (variable nominal) que un consumidor compraría en función de su nivel de ingresos (variable ordinal), sexo (variable nominal) y edad (variable métrica)?

Análisis de regresión lineal múltiple. Se estudia el efecto simultáneo de un conjunto de variables de cualquier tipo (variables explicativas) sobre una variable métrica (variable dependiente).

- *Ejemplo*: ¿Cómo influye en el rendimiento de una máquina el modelo de la misma, el operario que la maneja, su antigüedad y el turno de trabajo?

Análisis de regresión logística (logit). Similar al análisis de regresión pero en este caso la variable dependiente es dicotómica, es decir, que sólo toma dos valores: 1 si el caso presenta la característica analizada y 0 en caso contrario.

- *Ejemplo*: ¿Qué variables explicativas tienen mayor influencia en la avería (Y=1) o no (Y=0) de las máquinas de una factoría?

Análisis de la covarianza (ANCOVA). Comparamos el efecto de una variable métrica (denominada covariable o cofactor) y otra nominal (denominada factor o tratamiento) sobre una variable métrica (variable dependiente).

- *Ejemplo*: ¿Depende el rendimiento de una máquina (variable métrica dependiente) del modelo de la máquina (factor) y de la antigüedad de la misma (covariable)?

Modelo lineal general. Este enfoque puede considerarse como una extensión del análisis de regresión donde consideramos simultáneamente varias variables dependientes. Permite abordar análisis en circunstancias donde otras técnicas presentan dificultades como es el caso del análisis de la varianza frente a diseños desequilibrados de datos o el análisis de regresión frente a la existencia de correlación entre las variables explicativas (multicolinealidad).

Análisis factorial. Con esta técnica reducimos el número de variables explicativas a un número menor de factores. Cada factor es una combinación lineal de un conjunto de las variables explicativas.

- *Ejemplo*: En lugar de utilizar las variables nivel de ingresos, nivel educativo y nivel profesional, todas correlacionadas entre sí, podemos "fundirlas" en una nueva que podríamos llamar *status*, así, a la hora de explicar el gasto en bienes de lujo bastaría analizar su correlación con el status social en lugar de considerar las tres simultáneamente.

Análisis de conglomerados (o *cluster* en inglés). Puede considerarse como una técnica inversa al análisis discriminante. En efecto, mientras en esta última tenemos grupos de elementos y queremos ponderar las variables que los definen, en el análisis de conglomerados buscamos formar grupos homogéneos a partir de los valores que toma cada elemento respecto a un conjunto de variables.

- *Ejemplo*: Una editorial está interesada en clasificar a los lectores en grupos relativamente homogéneos en función de su nivel de ingresos, edad y tipo de sexo literario más demandado.

La tabla siguiente clasifica cada una de estas técnicas según la naturaleza (métrica o no) de las variables dependientes, esto es, las variables objeto de estudio que intentamos explicar, y las variables independientes, es decir, las que explican el valor alcanzado por las dependientes. No se incluyen en esta tabla el análisis factorial ni el análisis de conglomerados porque estas técnicas consideran todas la variables simultáneamente sin clasificarlas en dependientes e independientes.

Tabla 2.4. *Clasificación de técnicas multivariantes de análisis según el tipo de variable*

		Variables dependientes (Ys)	
		No métricas (nominales u ordinales)	**Métricas** (de razón o de escala)
Variables independientes (Xs)	**No métricas** (nominales u ordinales)	Análisis discriminante Regresión logística	ANOVA n, MANOVA, MLG
	Métricas (de razón o de escala)	Análisis discriminante Regresión logística	Análisis de regresión lineal, ANCOVA, MLG

Es de importancia para quien desee alcanzar una certeza en su investigación el saber dudar a tiempo
(Aristóteles)

CAPÍTULO 3

CONTRASTACIÓN DE HIPÓTESIS ESTADÍSTICAS

Capítulo 3. Contrastación de hipótesis estadísticas

3.1. Diagrama de frecuencias y funciones de distribución

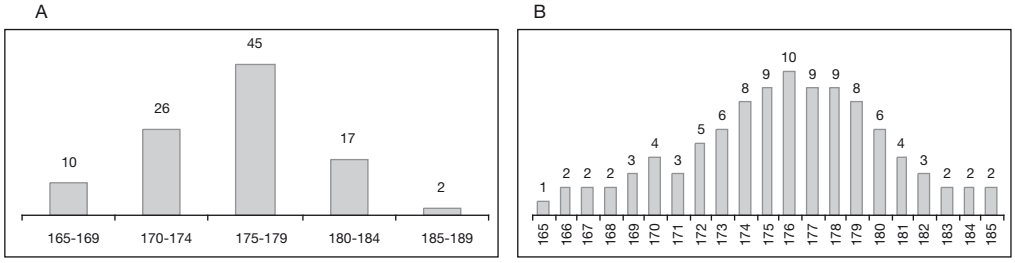
A la hora de realizar cualquier una prueba estadística resulta esencial comprender qué es una función de distribución y la probabilidad de obtener un valor (o rango) determinado al azar. Imaginemos que tenemos una muestra con la altura de 100 individuos, como aparece en la tabla siguiente:

Tabla 3.1. *Muestra con la altura de 100 individuos por orden creciente*

Caso	1	2	3	4	5	6	7	8	9	10	11	...	94	95	96	97	98	99	100
Altura	165	166	166	167	167	168	168	169	169	169	170	...	182	183	183	184	184	185	185

Es posible tener una visualización de la distribución de los datos utilizando un diagrama de frecuencias, esto es, un histograma, agrupando los casos en cada categoría. La altura de cada barra indica el número de casos del intervalo por lo cual nos permite tener una aproximación visual tanto de la tendencia central de los datos de la muestra como de su dispersión.

Figura 3.1. *Diagrama de frecuencias*



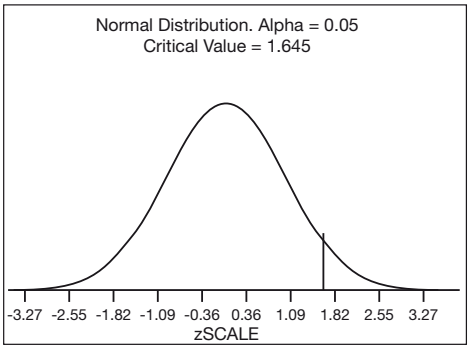
En la Figura 3.1.A, que agrupa los datos de la Tabla 3.1 en cinco categorías, el intervalo con mayor probabilidad es el tercero ya que tiene el mayor número de casos (45), en concreto, esta probabilidad es igual a 45/100=0,45. En la Figura 3.1.B hemos incrementado el número de categorías hasta 21. En este segundo histograma, la moda (el valor que más se repite), la mediana (el valor central de los datos ordenados de forma creciente o decreciente) y la media coinciden (valor igual a 176). La probabilidad de que un individuo mida más de 172 y menos de 180, es igual a 0,59, ya que (6+8+9+ 10+9+9+8)/100=0,59. La probabilidad de la categoría más frecuente en la Figura 3.1.B es 10/100=0,1, frente a 0,45 en la Figura 3.1.A.

Vemos cómo a medida que incrementamos el número de categorías la probabilidad de cada una disminuye, ya que hay menos casos en cada una de ellas. Si elevamos el número de categorías hasta el infinito, como puede hacerse con las variables métricas, el diagrama de frecuencias pasa a denominarse función de distribución. A

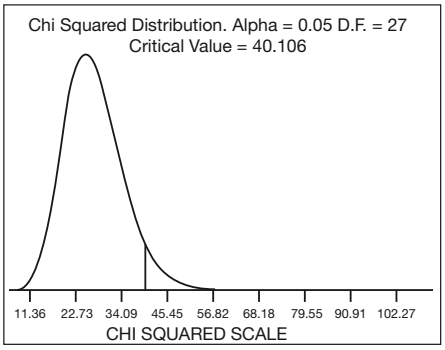
continuación presentamos dos tipos de funciones de distribución frecuentes, la normal y la Chi-cuadrado.

Figura 3.2. *Ejemplos de funciones de distribución*

A. Función de distribución normal



B. Función de distribución Chi^2



En el caso de la función de distribución normal podemos ver que la probabilidad de encontrar un valor superior a 1,645, esto es, el área a la derecha de 1,645, es igual a 0,05. Para la función de distribución Chi-cuadrado el valor que deja a la derecha el 5% de la superficie es 40,106. En ambos casos la probabilidad de encontrar un valor concreto es cero ya que dividimos un caso por infinito, por lo que siempre hablamos de la probabilidad de un intervalo. Los valores que dejan una superficie determinada a la derecha de una distribución están tabulados para algunas funciones (normal, Chi-cuadrado, *F* de Fisher-Snedecor, *t* de Student, etc.).

3.2. Corroboración de hipótesis estadísticas

En cualquier tipo de prueba estadística se compara el valor de un estadístico con su valor teórico en caso de seguir una distribución determinada. Los conceptos relacionados con una prueba de hipótesis son[11]:

- Valor calculado de un estadístico: Es el valor que se obtiene a partir de los datos de la muestra aplicando una fórmula matemática.
- Valor crítico de un estadístico: Es el valor que deja a la derecha de la distribución (o a la izquierda si es negativo) un porcentaje del área total de la misma. Este porcentaje representa la probabilidad de exceder ese valor crítico y se denomina nivel de significación (α).
- Hipótesis nula (H_0): Aquella que rechazamos si el valor calculado del estadístico excede el valor crítico del mismo.
- Hipótesis alternativa (H_1): Aquella que aceptamos cuando rechazamos la hipótesis nula.

Tomemos como ejemplo la prueba Chi-cuadrado de independencia entre dos variables. La prueba Chi-cuadrado calcula el estadístico Chi basándose en las frecuencias observadas y las esperadas. Este estadístico, bajo la hipótesis nula de independencia,

se distribuye como una Chi-cuadrado con (*m*-1)*(*n*-1) grados de libertad, siendo *m* y *n* el número de categorías de cada variable. Nuestras hipótesis serían:

H_0: Las dos variables son independientes.
H_1: No son independientes.

Imaginemos que queremos comprobar si el peso (supongamos que se encuentra codificado en 4 categorías) es independiente del sexo (dos categorías). Los grados de libertad de esta prueba son (4-1)*(2-1)=3. Si miramos el Anejo 2, el valor crítico con tres grados de libertad que deja el 5% del área a la derecha, $Chi_{3,\ 0,05}$, es 7,815. Si el valor calculado del estadístico Chi a partir de los datos de la muestra es superior al valor crítico rechazamos la hipótesis nula de independencia de las dos variables. ¿Por qué? Porque siendo tan improbable obtener un valor tan alto (sólo el 5% de probabilidad) el hecho de obtenerlo nos inclina a pensar que la hipótesis nula es falsa, es decir, el estadístico calculado no sigue una distribución Chi-cuadrado con tres grados de libertad porque las variables no son independientes.

A continuación resumimos algunas de los contrastes de hipótesis más usuales en el análisis estadístico, incluyendo la formulación de la hipótesis nula (la que se rechaza cuando $p<0,05$) y su significado.

Tabla 3.2. *Hipótesis nula en los contrastes de hipótesis más habituales*

Tipo de prueba o análisis	Estadísticos	Hipótesis nula (H_0)
Comparación de medias	t, Mann-Whitney, Wilcoxon	Las medias son iguales, por tanto no hay diferencias entre los dos grupos
Independencia entre dos variables nominales	Chi-cuadrado, Fisher	Las dos variables son independientes
Correlación	Pearson, Spearman, Kendall Tau	El coeficiente de correlación es cero, por tanto no hay relación entre las dos variables
ANOVA	F, Kruskal-Wallis	El factor no tiene efecto sobre la variable dependiente
Homogeneidad de la varianza	Barlett, Cochran, Hartley	Homogeneidad de la varianza entre grupos
Regresión	F	Todos los coeficientes del modelo son iguales a cero, por tanto ninguna variable aporta nada a la explicación de la variable dependiente
Regresión	t	El coeficiente de la variable explicativa es igual a cero, por tanto la variable debe ser eliminada del modelo

3.3. Errores Tipo I y Tipo II

En la mayoría de las pruebas estadísticas se fija una probabilidad del 5%, esto es, el nivel de significación es igual al 5% (α=0,05), como límite para aceptar que el valor que observamos puede haber sido seleccionado aleatoriamente (una probabilidad inferior al 5% es realmente baja por lo que se considera que no se ha obtenido por azar sino que nuestra hipótesis de partida, la hipótesis nula, era falsa). Como hemos explicado en el apartado anterior, para aceptar o rechazar la hipótesis nula basta con obtener el valor tabulado de la función de distribución que deja ese 5% de la superficie a la

derecha (o 2,5% en cada extremo en una prueba de dos colas), denominado valor crítico, y compararlo con el valor obtenido en la muestra. Si el valor calculado en la muestra es mayor que el valor crítico se rechaza la hipótesis nula y aceptamos la hipótesis alternativa. Por el contrario, si este valor calculado es menor que el valor crítico no se puede rechazar nuestra hipótesis nula.

El error del Tipo I consiste en rechazar una hipótesis nula cuando realmente es verdadera. Este error es igual al nivel de significación (α), que por lo general se sitúa en el 5% o para pruebas muy exigentes en el 1%. El error de Tipo II consiste en aceptar una hipótesis nula cuando realmente es falsa. Este error, conocido como β, está relacionado con el valor de α de una forma inversa y no lineal, por lo tanto, cuanto menor es el riesgo de cometer el error de Tipo I, mayor es el riesgo de cometer el error de Tipo II. La tabla siguiente resume la probabilidad de cometer uno u otro error:

Tabla 3.3. *Probabilidad de cometer errores de Tipo I o II*

		Decisión tomada por el investigador	
		H_0 es verdadera	**H_0 es falsa**
Realidad (no la conocemos)	**H_0 es verdadera**	Decisión correcta Probabilidad =1-α = *Nivel de confianza*	Error Tipo I Probabilidad =α = *Nivel de significación*
	H_0 es falsa	Error Tipo II Probabilidad =β	Decisión correcta Probabilidad =1-β = *Potencia de la prueba*

Siendo arbitraria la fijación del nivel de significación, α, muchos investigadores prefieren dar directamente la probabilidad asociada al valor calculado y que sea el propio investigador quien decida si rechaza o no la hipótesis nula (H_0). Todos los paquetes estadísticos proporcionan esta probabilidad (en general, bajo la columna *Prob*.). De igual forma OS4 proporciona *Prob.>t*, que es la probabilidad de obtener dicho valor por azar asumiendo que H_0 es cierta. Cuando el número de casos no es muy elevado (inferior a 50), los valores críticos tabulados se toman de una distribución t de Student en lugar de una distribución normal. Cuanto mayor es el tamaño de la muestra más se acercan los valores críticos de t a los de la distribución normal. Por ejemplo, para una muestra de 120 casos y α=0,05 (valor que deja el 5% de la distribución a su derecha), el valor crítico de t es igual a 1,98 frente a 1,96 en la normal.

3.4. Pruebas de dos colas

Cuando la hipótesis alternativa no indica el sentido de la diferencia debemos utilizar una prueba de dos colas. Por el contrario, si la dirección de la diferencia se recoge en la hipótesis alternativa utilizaremos una prueba de una sola cola. Para aclarar este punto veamos dos ejemplos:

Prueba de una cola. Queremos determinar si el tratamiento A contra una plaga reduce significativamente la incidencia de la misma sobre el cultivo. Para ello medimos el número de gusanos por planta en 5 parcelas con tratamiento y en 5 parcelas sin

tratamiento. En este caso, lo lógico es pensar que el tratamiento no va a incrementar el número de gusanos y que, si tiene algún efecto, será reducir dicho número. Este es un caso de prueba de una cola. La hipótesis se formularía como sigue:

H_0: $\mu_0=\mu_A$ (no existen diferencias entre las dos medias)
H_A: $\mu_0>\mu_A$ (la media de las fincas sin tratar es mayor que la media de las fincas tratadas)

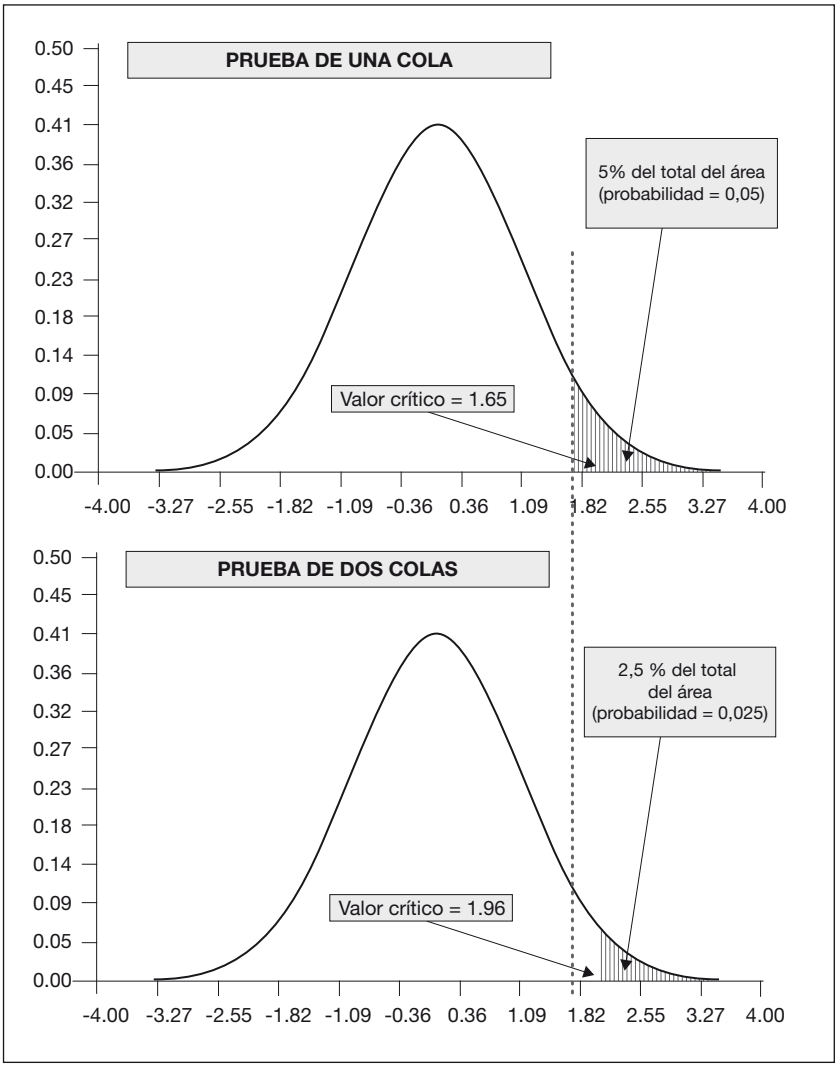
Prueba de dos colas. Siguiendo con el mismo ejemplo, si comparamos 5 parcelas tratadas con el pesticida A con otras 5 parcelas tratadas con el pesticida B, ya no sabemos si el que nos interesa, el A, tiene un comportamiento mejor o peor que el B. En este caso, la hipótesis se formula de la siguiente forma:

H_0: $\mu_A=\mu_B$ (no existen diferencias entre las dos medias)
H_A: $\mu_A\neq\mu_B$ (las medias de las fincas tratadas con A o B son distintas)

¿Qué implica el hecho de considerar una prueba con una o dos colas? En el primer caso, una cola, el valor crítico que determina rechazar o no la hipótesis nula deja a la derecha la totalidad del nivel de significación (α). En el caso de una prueba de dos colas, este valor deja a la derecha (o la izquierda para un valor negativo) la mitad de esta probabilidad ($\alpha/2$). Esto implica que en una prueba de dos colas es más difícil rechazar la hipótesis nula ya que el valor crítico es mayor. Por ejemplo, en una distribución normal el valor crítico que deja el 5% de la probabilidad a la derecha es igual a 1,65, sin embargo este valor aumenta hasta 1,96 con sólo el 2,5% de la superficie a la derecha.

En las ciencias sociales es más frecuente realizar pruebas de dos colas ya que resulta más complicado predecir las relaciones entre variables (¿podríamos asegurar que el nivel de estudios implica un mayor nivel de ingresos?). Teniendo en cuenta que las pruebas de dos colas son más exigentes, el rechazo de la hipótesis nula con una prueba de dos colas implica el rechazo de la misma utilizando una prueba de una cola (pero no viceversa), si bien las hipótesis alternativas no son las mismas. Por tanto, en la mayoría de los casos utilizaremos una prueba de dos colas, es decir, buscaremos en la tabla el valor crítico tabulado que deja a la derecha una superficie igual a 0,025, suponiendo que hemos fijado el nivel de significación en el 5%. La figura siguiente muestra claramente este punto.

Figura 3.3. *Valores críticos para pruebas de una cola y de dos colas*



Largo es el camino de la enseñanza por medio de teorías;
breve y eficaz por medio de ejemplos

Para saber algo, no basta con haberlo aprendido

(Séneca)

CAPÍTULO 4

INTRODUCCIÓN A OS4

Capítulo 4. Introducción a OS4

4.1. Obtención del paquete estadístico

OS4 es un programa de análisis estadístico bajo entorno Windows con un manejo e interpretación de resultados similar a otros paquetes de amplia difusión. Para descargar de forma gratuita la última versión junto con el manual del programa el lector puede dirigirse al sitio de internet: http://www.statpages.org/miller/openstat

4.2. Manejo de los archivos de datos

Al abrir el programa en primer lugar decidimos si el separador decimal es la coma o el punto. Para una mayor estabilidad del programa se recomienda utilizar el punto como separador decimal. Si el entorno Windows que utilizamos tiene la coma como separador decimal y el punto como separador de miles podemos intercambiarlos en el menú de "Configuración regional" dentro del "Panel de control". Otra opción es transformar los datos originales para evitar los decimales, así por ejemplo en lugar de utilizar el dato 1,73 m podemos transformarlo en 173 cm.

Introducción de datos directamente en OS4

Podemos seleccionar "Variables → Define." e incluir la primera variable indicando su tipo y etiquetas (opcional) para cada valor de la variable. Si presionamos ENTER en una celda de la última variable existente aparece de nuevo el menú de creación de una nueva variable. Para introducir datos basta con situarse en la celda correspondiente y utilizar las flechas del teclado para ir a la celda adyacente.

Importación de datos desde una hoja de cálculo

En el caso frecuente de tener los datos en una hoja de cálculo (por ejemplo Excel), prepararemos la hoja dejando sólo la tabla que vamos a exportar, con los nombres de las variables (máximo de 8 caracteres) comenzando en A1, B1, etc. A continuación exportamos el fichero en formato texto siguiendo los comandos:

Archivo → Guardar como → Nombre del archivo: *ejemplo*; Guardar como tipo: *Texto (MS-DOS)*.

Para importar el fichero "ejemplo.txt" en OS4 (en el caso de tener alguna coma podemos utilizar el comando Reemplazar (Edición, CTRL+H) del Bloc de notas, sustituir todas las comas por puntos y guardar de nuevo el fichero como texto) seleccionamos

Files → Open Character Separated Values File → Tab File; *En el cuadro de diálogo seleccionar* Tipo de archivo:: All *.* y buscamos el fichero "ejemplo.txt".

Para guardarlo en formato OS4: Files → Save as: OpenStat4 File → *ejemplo*.

En la sección siguiente comenzaremos a trabajar con una base de datos hipotética donde se incluyen los datos de 50 pacientes ingresados por sufrir algún episodio cardiaco (ver Anejo 3). Las variables consideradas a *priori* relacionadas con la variable dependiente (sufrir o no el ataque) aparecen en la Tabla 4.1.

Tabla 4.1. *Descripción de las variables de la base de datos médica*

Nombre	Descripción de la variable
ALTURA	Estatura en cm
PESO	Peso en kg
SEXO	1=Hombre; 2=Mujer
CARDIO	0=No tuvo crisis cardiaca; 1=Sí la tuvo
EDAD	Edad del paciente
ALIMENTA	1=dieta equilibrada; 2=dieta preferentemente vegetariana; 3=dieta con gran contenido en grasas
EJERCICIO	1=2 horas o menos; 2=Más de 2 y menos de 4 horas; 3=Entre 4 y 6 horas; 4=Más de 6 horas

4.3. Análisis preliminar de los datos

Para comprobar que no hemos cometido ningún error en la introducción de los datos podemos obtener el mínimo y el máximo de cada variable. Esta práctica debería llevarse a cabo de forma rutinaria antes de comenzar con el análisis estadístico para evitar llegar a conclusiones erróneas. Sirve además para tener una descripción inicial de las variables.

En el caso de las variables métricas, en nuestro ejemplo las variables ALTURA, PESO y EDAD, seleccionamos en el menú:

Analyses → Descriptive → Distribution Statistics: *Altura*.

El resultado del análisis incluye la media (*Mean*), la varianza (*Variance*), la desviación típica (*Std. Dev.*), el rango (*Range*), el momento de tercer orden (*Skewness*), el cual indica el grado de simetría, y el momento de cuarto orden (*Kurtosis*), indicador del grado de achatamiento de la distribución. Para la variable ALTURA tenemos:

```
DISTRIBUTION PARAMETER ESTIMATES

Altura (N =50) Sum = 8665.000
Mean = 173.300 Variance = 87.684 Std.Dev. = 9.364
Std.Error of Mean = 1.324
Range = 43.000 Minimum = 151.000 Maximum = 194.000
Skewness = -0.006 Std. Error of Skew = 0.337
Kurtosis = 0.478 Std. Error Kurtosis = 0.662
```

De los errores estándar que aparecen (media, simetría y achatamiento) nos interesa el error estándar de la media, que se calcula como sigue: $\sqrt{(s^2/n)} = \sqrt{(87{,}68/50)} = 1{,}75$, siendo un indicador de la variabilidad de la media muestral (sacando diferentes muestras de la población obtenemos diferentes medias muestrales, cuanto menor sea su dispersión menor será su error estándar).

Es frecuente en estadística descriptiva comparar la distribución de una variable con la distribución normal. Dicha distribución es simétrica y tiene el 95% de los casos en un intervalo de amplitud $2 \cdot \sigma$ (el doble de la desviación estándar). El valor de la simetría y del achatamiento en este tipo de distribución es igual a 0. La forma de esta distribución puede verse en el Anejo 1.

En la práctica, lo habitual es encontrarnos con variables que no siguen una distribución normal. En estos casos los coeficientes de simetría (*skewness*) y de achatamiento (*kurtosis*) no son cero, por tanto podemos indicar:

- Coef. de simetría (Skewness)>0 → distribución alargada hacia la derecha.
- Coef. de simetría (Skewness)<0 → la cola alargada se encuentra a la izquierda.
- Kurtosis>0 → la distribución es alta y estrecha en la base.
- Kurtosis<0 → la distribución es baja y ancha en la base.

Así pues, cuanto mayores sean los valores de simetría y kurtosis de la distribución mayor es el riesgo de asumir que siguen una distribución normal.

Con respecto a las variables nominales u ordinales, en nuestro ejemplo tenemos tres variables nominales (SEXO, CARDIO y ALIMENTA) y una ordinal (EJERCICIO), por lo que podemos utilizar al análisis de frecuencias:

Analyses → Descriptive → Distribution Frecuencies: *Sexo.*

En el cuadro que aparece tenemos:

```
FREQUENCY ANALYSIS BY BILL MILLER

Frequency Analysis for Sexo
FROM      TO       FREQ.    PCNT    CUM.FREQ.   CUM.PCNT.   %ILE RANK
1.00      2.00     23       0.46    23.00       0.00        0.23
2.00      3.00     27       0.54    50.00       1.00        0.73
```

Donde vemos que hay 23 datos con el valor 1 (hombres) y otros 27 con el valor 2 (mujeres). La columna FROM (desde) indica el valor de la categoría. PCNT muestra la frecuencia en porcentaje de cada categoría.

4.4. Creación de variables

Podemos obtener una variable nueva a partir de otra existente mediante una transformación matemática. Así, por ejemplo, podríamos expresar la altura en metros, en lugar de en centímetros. Bastaría con crear una nueva variable, que llamaremos ALTURA_M, de forma que ALTURA_M =ALTURA/100.

Variables → Transform → First Var. Argument (V1): *Altura*; Constant: *100*; Save new variable as: *Altura_m*; Select Transformation: *New=V1/C* → OK.

Seleccionando el tipo de transformación de las que aparecen en el menú superior. Como podemos comprobar en la hoja de datos, aparece una nueva columna con la altura en metros. Para eliminar esta columna nos situamos en una celda de la columna:

Edit → Delete Column. (no confundir con *Delete Row* que elimina una fila)

IMPORTANTE: Al no poder deshacer esta operación perderemos los datos de esa columna (o de la fila). Si esto ocurre de forma accidental bastaría con salir del programa sin guardar los últimos cambios y volver a recuperar el fichero con los datos originales.

4.5. Creación de tablas resumen

Tablas de frecuencias

Podemos obtener el número de casos en una tabla con dos variables categóricas (nominales u ordinales) como paso previo a la búsqueda de relaciones entre dos variables. Supongamos que deseamos saber cuántos hombres y mujeres han tenido una crisis cardiaca:

Analyses → Descriptive → CrossTabs → *Sexo, Cardio.*

```
…
Cell Frequencies by Levels

          Cardio
              0          1
Block 1   14.000     9.000
Block 2   20.000     7.000

Grand sum for all categories =50
```

En la parte superior del resultado nos indica el número de casos en cada combinación de sexo (Block 1 y 2) y sufrimiento o no de la crisis. Por ejemplo, en la combinación Sexo=1 (hombre) y Cardio=0 (no sufrió ninguna crisis) encontramos 14 casos.

Tablas con valores medios

Imaginemos que queremos saber la altura media por sexo. En este caso manejamos una variable nominal (SEXO) y otra métrica (ALTURA). Con OS4 se obtiene como sigue:

Analyses → Descriptive → Breakdown → Categorical Variables Selected: *Sexo*; Continuous Variables to Break Down: *Altura*. (un nuevo click baja la variable continua)

```
Variable levels:
Sexo       level = 1
Freq.      Mean          Std. Dev.
23         177.870       834.762
Variable levels:
Sexo       level = 2
Freq.      Mean          Std. Dev.
27         169.407       863.849

Number of observations accross levels = 50
Mean accross levels = 173.300
Std. Dev. accross levels =1214.617
```

Así, la altura media del grupo de hombres es de 177,9 cm frente a 169,4 en el caso de las mujeres. La media general de la muestra, sin distinción por sexo, es de 173,3 cm.

Lo que es afirmado sin prueba
puede ser negado sin prueba
(Euclides)

CAPÍTULO 5

ANÁLISIS BIVARIANTE

Capítulo 5. Análisis bivariante

5.1. Pruebas de normalidad

Teniendo en cuenta el tamaño reducido de la muestra con la que estamos trabajando, y con el objetivo de buscar relaciones entre las variables mediante pruebas paramétricas, es necesario comprobar si las variables en cuestión cumplen con los requisitos necesarios para este tipo de pruebas. Como apuntamos anteriormente, estos requisitos son:

1. Distribución normal de las variables.
2. Uniformidad de la varianza.
3. Escala de medida métrica.
4. Independencia de los datos.

De los cuatro requisitos, el tercero y el cuarto no presentan problemas para las variables ALTURA y PESO. Para la comprobación de la homogeneidad de la varianza podemos recurrir a la prueba de Hartley (ver apartado ANOVA), si bien sería necesario categorizar una de ellas para comprobar la homogeneidad de la varianza de la otra. Esta dificultad se traduce en que, desde un punto de vista práctico, los investigadores se limiten a la comprobación de la normalidad de la distribución de cada variable y asuman la homogeneidad de la varianza respecto a los diferentes niveles de la otra variable. Otra razón para omitir esta comprobación es la robustez de las pruebas paramétricas incluso ante el incumplimiento de este requisito.

Existen numerosas pruebas de normalidad, entre las más populares se encuentra la de *Kolmogorov-Smirnov*. Esta prueba tiene un poder inferior (mayor probabilidad de rechazar una distribución como normal cuando realmente lo es) comparada con otras pruebas especialmente diseñadas para este propósito. Por este motivo utilizaremos las pruebas de *Shapiro-Wilk* y de *Lilliefors*. En el caso de la variable ALTURA tenemos:

Analyses → Descriptive → Normality Tests → Test normality of: *Altura* → Apply.

El estadístico de *Shapiro-Wilk* (W) es igual a 0,9728. La probabilidad de obtener este valor al azar en el supuesto de que la distribución sea normal es de 0,2993. Como esta probabilidad excede de 0,05 no rechazamos la hipótesis nula de normalidad de la variable.

De igual manera, el estadístico de *Lilliefors* alcanza un valor de 0,087. El valor crítico para muestras superiores a 35 casos y un error del 5% se calcula mediante la siguiente fórmula:

$D_{0,05}=1,36 / \sqrt{n} = 1,36 / \sqrt{50} = 1,36 \cdot 7,07 = 9,62$

Siendo el valor calculado del estadístico menor que el valor crítico (0,0087 < 9,62), no podemos rechazar la hipótesis nula de normalidad de la variable. El programa nos ahorra esta comprobación dándonos la conclusión: No hay evidencias contra la normalidad (*No evidence against normality*).

En el caso de repetir el análisis para las variables PESO y EDAD llegaríamos a la conclusión de que la primera sí cumple con los criterios de normalidad y la segunda no (*Strong evidence against normality*).

5.2. Análisis de variables nominales

En esta sección estudiaremos algunas de las posibilidades de análisis de dos variables nominales mediante las tablas de contingencia. Si bien es posible aplicar estas técnicas para el estudio de dos variables ordinales, las tablas de contingencia no hacen uso de la ordenación intrínseca de los datos sino simplemente de su frecuencia en cada categoría por lo que no se explotan al máximo sus posibilidades: En estos casos podemos sacar "mayor rendimiento" utilizando el análisis de correlación.

Para analizar tablas de contingencia podemos utilizar diferentes estadísticos según el tamaño de la muestra y el número de categorías de cada variable, como podemos ver en la tabla siguiente:

Tabla 5.1. *Tipo de estadístico en tablas de contingencia*

Número de categorías de ambas variables	Tamaño de la muestra	Estadístico
2 x 2	n<100	Fisher
	n≥100	Chi-cuadrado con la corrección de Yates
Más de dos categorías en una de las variables (o en las dos)	Cualquiera	Chi-cuadrado

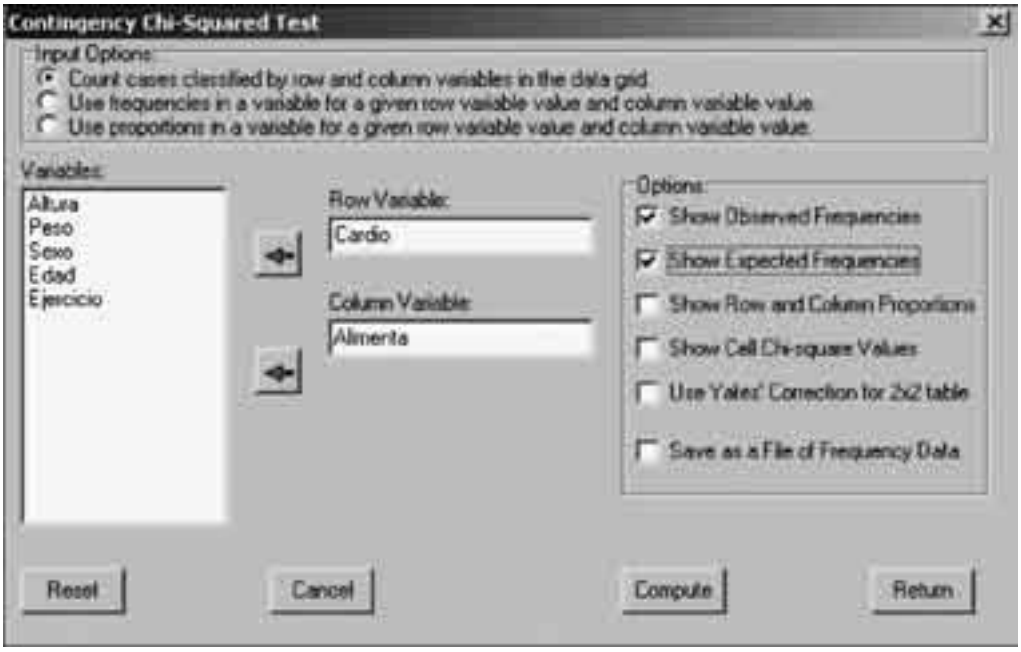
Prueba Chi-cuadrado

La búsqueda de relación entre dos variables categóricas se realiza mediante la prueba no paramétrica Chi-cuadrado (χ^2). Por ejemplo, ¿existe relación entre la ocurrencia o no de alguna crisis cardiaca en el pasado año y el tipo de alimentación?

Teniendo en cuenta que una de las variables (ALIMENTA) tiene más de dos categorías, procedemos a utilizar el estadístico Chi-cuadrado, cuya hipótesis nula de la prueba es la independencia entre ambas variables.

Analyses → Nonparametric → Contingency Chi-Squared Analysis → Variable for Row Categories: *Cardio*; Variable for Column Categories: *Alimenta*. Marcando las casillas de "*Show Observed Frequencies*" y "*Show Expected Frequencies*" para ver las frecuencias observadas y las frecuencias teóricas, respectivamente, que habría en el caso de que ambas variables fueran independientes.

Figura 5.1. *Cuadro de diálogo de la prueba Chi-Cuadrado*




```
Chi-square Analysis Results for Cardio and Alimenta
No. of Cases =50

OBSERVED FREQUENCIES

           Frequencies
              COL. 1     COL. 2     COL. 3     Total
Row 1             18         10          6        34
Row 2              3          3         10        16
Total             21         13         16        50

EXPECTED FREQUENCIES

           Expected Values
              COL. 1     COL. 2     COL. 3
Row 1         14.280      8.840     10.880
Row 2          6.720      4.160      5.120
```

En la tabla *Observed Frequencies*, Row 1 indica el primer valor de la primera variable (CARDIO=0) y Row 2, el segundo (CARDIO=1). Las columnas COL representan los valores de la variable ALIMENTA. Así, el cruce de Row 1 y COL 3 significa que hay 6 casos que no sufrieron ningún episodio cardiaco (CARDIO=0) y siguen una dieta rica en grasas (ALIMENTA=3).

La siguiente tabla, *Expected Frequencies*, nos informa de que una celda tiene una frecuencia teórica inferior a 5 (CARDIO=1 y ALIMENTA=2), en concreto esta combinación tiene una frecuencia teórica igual a 4,16. Si bien la prueba Chi-cuadrado exige que todas las frecuencias teóricas sean superiores a 5, puede admitirse esta prueba si:

1. Todas las frecuencias son superiores a 1, y
2. Las celdas que tienen una frecuencia entre 1 y 5 representan menos del 20% del total de celdas.

En nuestro caso sólo hay una celda con una frecuencia teórica inferior a 5, lo cual representa el 17% del total (1 de 6), por tanto la prueba Chi-cuadrado es válida.

Cuanto mayores sean las diferencias entre las frecuencias observadas y las teóricas mayor será el valor del estadístico Chi-cuadrado y mayor la probabilidad de rechazar la hipótesis nula de independencia entre ambas variables. Continuando con el resultado de la prueba tenemos:

```
Chi-square = 10.344 with D.F. =2. Prob. > value = 0.006

Liklihood Ratio = 10.247 with prob. > value =0.0060

phi correlation =0.4548

Pearson Correlation r =0.4316

Mantel-Haenszel Test of Linear Association = 9.126 with probability
> value =0.0025

The coefficient of contingency = 0.414

Cramer's V = 0.455
```

En este caso, el valor del estadístico es igual a 10,247, con una probabilidad asociada de ocurrencia de 0,0060 (es decir, 0,6%). Como esta probabilidad es inferior a 0,05 (es decir, 5%) rechazamos la hipótesis nula de independencia entre ambas variables, por tanto concluimos que la dieta sí influye en la ocurrencia o no de la crisis cardiaca.

También podemos comparar el valor calculado de la Chi y el valor crítico. En nuestro caso el estadístico calculado, si H0 es cierta, se distribuye como una Chi-cuadrado con 2 grados de libertad ([2-1]*[3-1], ya que una variable tiene dos categorías y la otra tres). En esta distribución la probabilidad de encontrarnos al azar un valor superior a 5,991 es de 0,05 (ver Anejo 2, fila=2 y columna=0,05). En nuestro ejemplo el valor del estadístico es 10,247, un valor muy superior al valor crítico (incluso superior a 9,21 que es el valor crítico correspondiente al 1% de probabilidad), por tanto rechazamos la hipótesis nula de independencia entre las dos variables.

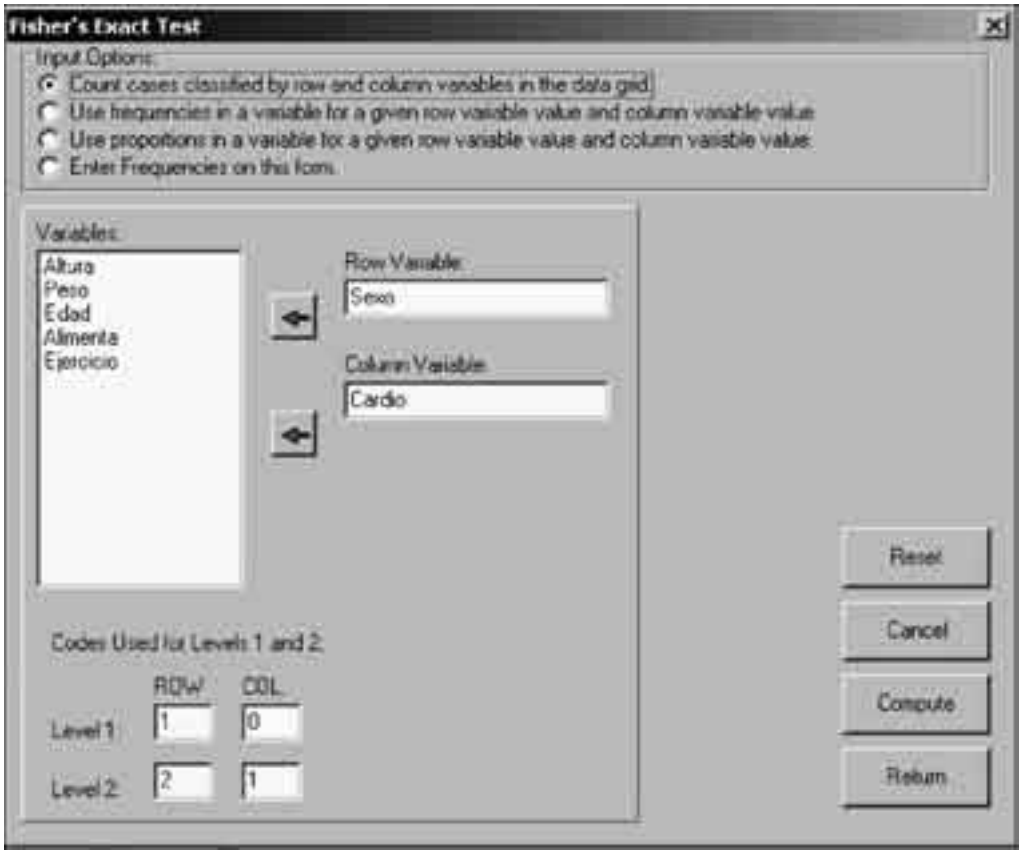
Prueba de Fisher

Esta prueba está diseñada para tablas de contingencia 2x2 con un número de casos inferior a 100. Supongamos que queremos determinar la influencia de sexo en la ocurrencia o no de una crisis cardiaca:

Analyses → Nonparametric → Fisher's Exact Test for 2x2 Table: Variable for Row Categories: *Sexo*; Variable for Column Categories: *Cardio*.

Por defecto, el programa asume que los valores de las variables son 1 y 2. Esto es correcto para la variable SEXO pero no para la variable CARDIO, que toma los valores 0 y 1. Como esta variable la hemos situado en *Column Variable*, cambiamos los valores en la columna COL por 0 y 1 (la columna ROW es para la variable SEXO, que permanece con los valores 1 y 2), tal y como aparece en el cuadro de diálogo siguiente:

Figura 5.2. *Cuadro de diálogo de la prueba Fisher de tablas de contingencia*



Nota: Tanto en esta prueba como en el resto de análisis de tablas de contingencia es posible introducir los datos directamente en formato tabla indicando el número de casos en cada una de las celdas de la tabla de contingencia seleccionando la opción "Enter Frequencies on this form".

El resultado de la prueba es el siguiente:

```
Fisher Exact Probability Test
Accumulating Values of the Hypergeometric Distribution
Contingency Table for Fisher Exact Test
       Column
Row    1      2
1      14     9
2      20     7
Probability =0.1474
...
Cumulative Probability =0.2439

Null hypothesis accepted.

Chi-squared =0.995 with 1 d.f. and prob. > chi-square =0.3185

Log Odds = -0.608

Relative Risk = 0.822

Liklihood Ratio = 0.995 with prob. > value =0.3186

phi correlation =0.1411
Mantel-Haenszel Test of Linear Association=0.975 with probability >
value =0.3234
The coefficient of contingency = 0.140
Cramer's V = 0.141
```

Teniendo en cuenta que la probabilidad acumulada (*Cumulative Probability*) es superior a 0,05 (alcanza un valor de 0,2439), se acepta la hipótesis nula (Null hypothesis accepted), esto es, ambas variables son independientes. La probabilidad del estadístico Chi-cuadrado también es superior a 0,05 (en concreto 0,3185) por lo que también se llega a la misma conclusión.

5.3. Diferencia entre dos grupos: Comparación de las medias

Es frecuente en estadística comparar el valor medio de una variable en dos grupos y preguntarse si las diferencias observadas se deben a que ambos grupos provienen de dos poblaciones diferentes o provienen de la misma pero las diferencias son producto del azar.

A modo de ejemplo, exploremos la relación entre las variables SEXO y ALTURA y entre CARDIO y EDAD. En el primer caso, y teniendo en cuenta que la variable métrica ALTURA sigue una distribución es normal, es apropiada una prueba de comparación de medias paramétrica como la prueba t. En el segundo caso, la variable métrica EDAD no cumple este requisito, por lo que recurriremos a una prueba de comparación de medias no paramétrica, la prueba de Mann-Whitney.

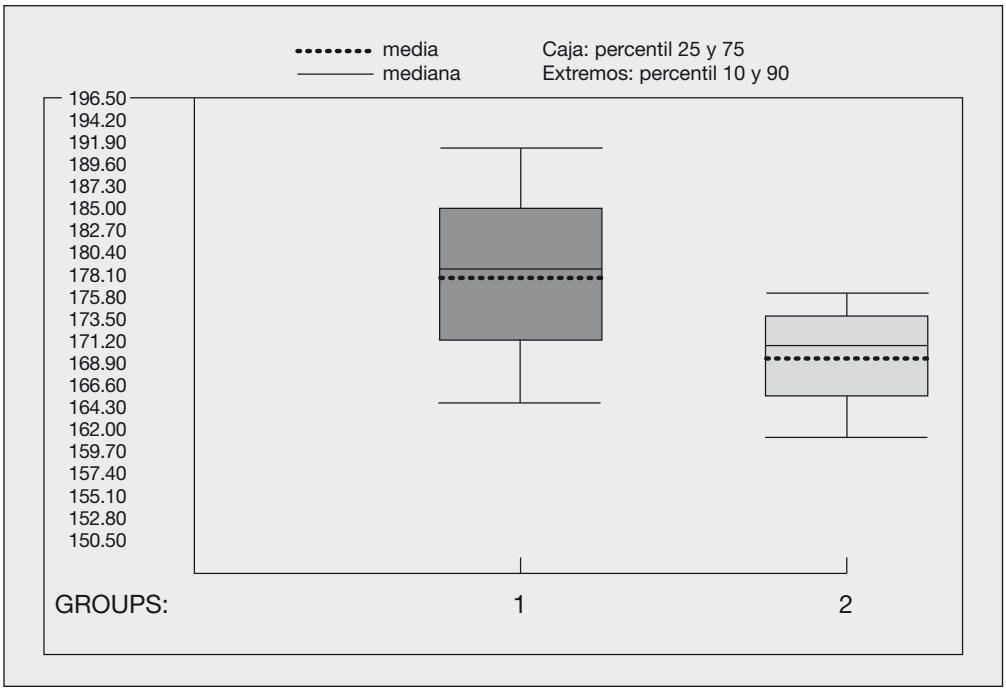
Diagrama de dispersión de una variable por grupos

Podemos ver gráficamente si existe o no una relación entre el SEXO y la ALTURA mediante el diagrama de cajas. Para ello ejecutamos:

Analyses → Descriptive → Box and Whisker Plot → Group Variable: *Sexo*, Measurement Variable: *Altura*.

La figura siguiente nos muestra la relación entre ambas variables. De los dos grupos, el primero se corresponde con los hombres y el segundo con las mujeres. La línea discontinua se corresponde con la media de cada grupo, la línea continua interna con la mediana, la caja contiene los percentiles 25% y 75% mientras que los extremos horizontales representan los percentiles 10% y 90%.

Figura 5.3. *Diagrama de cajas de la relación Sexo-Altura*



Parece claro que sí existe una relación entre ambas variables ya que los hombres tienen valores superiores respecto a la media, la mediana y todos los percentiles. Para corroborar esta hipótesis inicial realizaremos una prueba de comparación de medias. Como apuntamos al principio, en el caso de que la variable métrica siga una distribución normal utilizaremos la prueba paramétrica t. Por el contrario, si la dicha variable no se distribuye como una normal recurriremos a la prueba de comparación de medias no paramétrica de Mann-Whitney.

Prueba paramétrica t

En el caso de la relación entre la ALTURA y el SEXO, para corroborar nuestra hipótesis inicial de que los hombres son más altos que las mujeres procedemos de la siguiente forma:

Analyses → Two Sample Tests → Two Means → Values in the Data Grid: First Variable: *Altura*; Group Variable: *Sexo*; Group 1 Code: *1*; Group 2 Code: *2*.

Figura 5.4. *Cuadro de diálogo de la prueba paramétrica de comparación de medias t*

Con los siguientes resultados:

```
COMPARISON OF TWO MEANS

Variable    Mean        Variance    Std.Dev.    S.E.Mean    N
Group 1     177.87      101.75      10.09       2.10        23
Group 2     169.41      44.94       6.70        1.29        27

Assuming =variances, t = 3.540 with probability =0.0009 and 48
degrees of freedom
Difference = 8.46 and Standard Error of difference = 2.39
Confidence interval =( 3.66, 13.27)

Assuming unequal variances, t = 3.429 with probability =0.0011 and
37.21 degrees of freedom
Difference = 8.46 and Standard Error of difference = 2.47
Confidence interval =( 3.46, 13.46)
F test for equal variances = 2.264, Probability =0.0238
```

La prueba de homogeneidad de la varianza (F-test) arroja un estadístico con probabilidad inferior a 0,05 (Probability =0,0238) por lo que rechazamos la hipótesis nula de igualdad de varianzas y por tanto utilizamos el estadístico *t* de no igualdad de varianzas (*Assuming unequal variances*), es decir el valor *t* =3,429 con una probabilidad de 0,0011, por tanto, siendo inferior a 0,05 esta probabilidad, rechazamos la hipótesis nula de independencia entre ambas variables. Recordemos que:

- p<0,05 → rechazamos la hipótesis nula de que todas las medias son iguales.
- p>0,05 → no se rechaza la hipótesis nula de que todos los grupos tienen medias iguales (las diferencias observadas no son estadísticamente significativas).

Prueba no paramétrica de Mann-Whitney

En el caso de que la variable no cumpla los requisitos paramétricos, una de las pruebas de comparación de medias disponible es la de Mann-Whitney. Aplicándola a la variable EDAD y a la incidencia o no de una crisis cardiaca tenemos:

Analyses → Nonparametric → Mann-Whitney U Test → Treatment Group Codes: *Cardio*; Dependent Variable: *Edad*.

```
Sum of Ranks in each Group
Group      Sum No.    in Group
0          802.00     36
1          473.00     14

No. of tied rank groups = 12
Statistic U =368.0000
z Statistic (corrected for ties) = 2.5064, Prob. > z =0.0061
z test is approximate. Use tables of exact probabilities in Siegel.
(Table J or K, pages 271-277)
```

El estadístico U se distribuye como una normal para muestras superiores a 20 (nuestra muestra excede esta cantidad por lo que podemos asumir que esta condición se cumple) bajo la hipótesis nula de no diferencia en las medias.

La probabilidad del estadístico z, `Prob > z =0,0061` es inferior a 0,05 por lo que rechazamos la hipótesis nula de igualdad de medias entre los dos grupos, esto es, la media de edad de los que sufrieron un ataque es estadísticamente diferente de la media de edad de los que no lo sufrieron.

5.4. Diferencia entre dos o más grupos: Análisis de la varianza

Prueba paramétrica ANOVA I

En el caso de que necesitemos analizar la influencia de una variable categórica nominal, conocida como factor, sobre una variable métrica (por ejemplo determinar si cuatro tratamientos contra una plaga producen resultados estadísticamente diferentes) podemos recurrir al análisis de la varianza, también conocido como ANOVA. Si existe un solo factor se denomina ANOVA I, con dos factores ANOVA II, y en general, con n factores, ANOVA n. En este capítulo de análisis bivariante corresponde estudiar ANOVA I.

Si el factor (por ejemplo tipo de abonado) toma sólo dos valores (abono A y abono B) podemos realizar una prueba de comparación de medias, como hemos visto en la sección anterior. En general, el factor tomará más de dos valores (abono A, B, C y D) por lo que descartamos la prueba de comparación de medias y utilizamos el análisis de la varianza.

Para llevar a cabo ANOVA es necesario que la variable métrica cumpla con los requisitos paramétricos ya descritos. Sin embargo, mientras el incumplimiento de la normalidad de la variable tiene un efecto reducido en la validez de la prueba, la diferente varianza de los grupos (heterocedasticidad) sí puede afectar a dicha validez[12]. Por tanto, en el caso de rechazar esta igualdad de la varianza (prueba de Hartley y Barlett Chi-cuadrado) procederemos como sigue:

- Grupos de tamaño similar → Pruebas paramétricas de análisis.
- Grupos de tamaño muy diferente → Pruebas no paramétricas de análisis.

En la práctica, las conclusiones del análisis de la varianza siguiendo una prueba paramétrica o una prueba no paramétrica son, en la mayoría de los casos, las mismas.

Para el análisis de la varianza utilizaremos un los datos reales recogidos en cinco encuestas realizadas a visitantes de parques naturales situados en las provincias Córdoba y Jaén en 2001[13]. En esta base de datos (a efectos prácticos, el lector puede trabajar con la submuestra de 50 casos que aparece en el Anejo 3) podemos encontrar las siguientes variables:

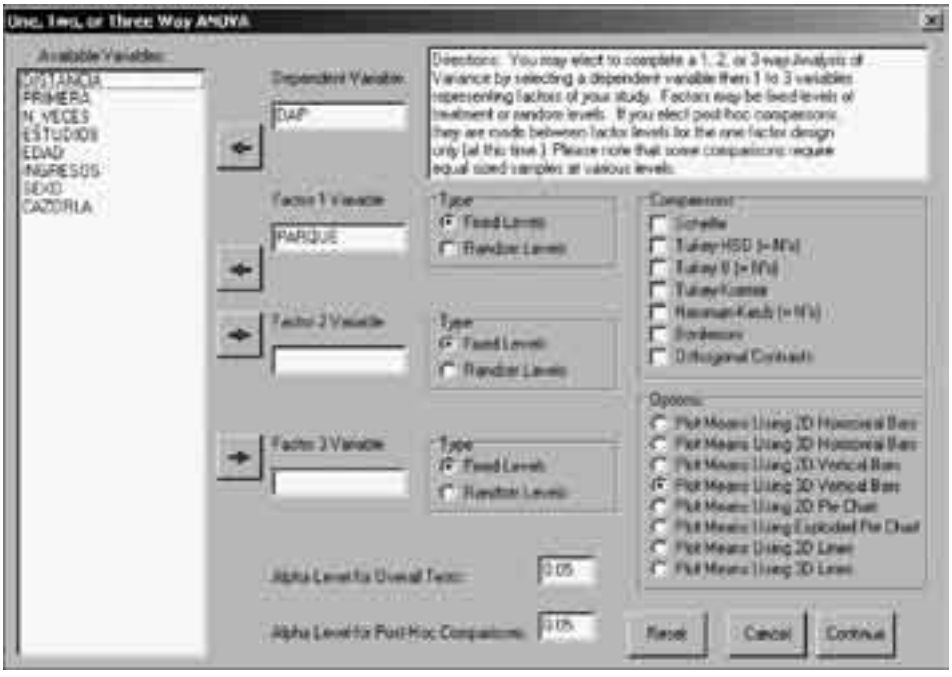
Tabla 5.2. *Variables de la encuesta sobre la disposición a pagar por la visita a un parque*

Nombre	Descripción de la variable
DAP	Disposición a pagar por la visita al parque (euros)
PARQUE	Código identificativo del parque natural (de 1 a 5)
DISTANCIA	Distancia recorrida hasta llegar al parque (km)
PRIMERA	Primera vez que visita el parque (Sí=1; No=0)
N_VECES	Número de veces que visita el parque al año
ESTUDIOS	Nivel de estudios (Min=1; Max=4)
EDAD	Edad del encuestado (Min=1; Max=6)
INGRESOS	Nivel de ingresos familiares (Min=1; Max=4)
SEXO	Sexo del encuestado (Hombre=1; Mujer=2)

Imaginemos que queremos estudiar si existe o no relación entre la cantidad que el encuestado está dispuesto a pagar por entrar al parque (DAP) y el parque visitado (PARQUE). En este caso, DAP es la variable dependiente, la que queremos explicar, mientras PARQUE es la variable de grupo o factor. No confundir el número de factores con el número de categorías del factor. En este ejemplo tenemos un único factor (variable PARQUE) con cinco categorías (hay cinco parques).

Analyses → Analyses of Variance → One, Two or Three Way ANOVA → Dependent Variable: *Dap*; Factor 1 Variable: *Parque*.

Figura 5.5. *Cuadro de diálogo del Análisis de la Varianza (ANOVA)*



```
ONE WAY ANALYSIS OF VARIANCE RESULTS
Dependent variable is: DAP, Independent variable is: PARQUE
-------------------------------------------------------------------------
SOURCE   D.F.     SS         MS       F        PROB.>F  OMEGA    SQR.
-------------------------------------------------------------------------
BETWEEN  4        117.13     29.28    21.71    0.00     0.29
WITHIN   195      262.99     1.35
TOTAL    199      380.12
-------------------------------------------------------------------------
MEANS AND VARIABILITY OF THE DEPENDENT VARIABLE FOR LEVELS OF THE
INDEPENDENT VARIABLE
-------------------------------------------------------------------------
GROUP    MEAN     VARIANCE   STD.      N
                             DEV.
-------------------------------------------------------------------------
1        1.09     2.15       1.47      38
2        3.11     2.75       1.66      24
3        1.40     1.44       1.20      40
4        0.86     1.02       1.01      48
5        0.53     0.33       0.57      50
-------------------------------------------------------------------------
TOTAL    1.20     1.91       1.38      200
-------------------------------------------------------------------------
TESTS FOR HOMOGENEITY OF VARIANCE
-------------------------------------------------------------------------
Hartley Fmax test statistic = 8.40 with deg.s freem: 5 and 49.
Cochran C statistic =  0.36 with deg.s freem: 5 and 49.
Bartlett Chi-square =  45.83 with 4 D.F. Prob. > Chi-Square = 0.000
-------------------------------------------------------------------------
```

La columna SS indica la suma de las desviaciones al cuadrado de la media de cada grupo respecto a la media total ($SS_{BETWEEN}$) y las desviaciones al cuadrado de cada dato respecto a la media de su grupo (SS_{WITHIN}). El estadístico F se calcula como sigue:

$F=(SS_{BETWEEN} / df) / (SS_{WITHIN} / df) =(117,13 / 4) / (262,99 / 195) =21,71$

o lo que es lo mismo, $F=MS_{BETWEEN} / MS_{WITHIN} =29,28 / 1,35 =21,71$

La probabilidad de obtener un valor mayor que 21,71 en una distribución F con estos grados de libertad es de 0,00, por lo que rechazamos la hipótesis nula de no efecto de la variable de grupo sobre la variable dependiente, o lo que es equivalente, aceptamos que el *PARQUE* sí influye en la *DAP*.

El estadístico `OMEGA SQR.` Indica el grado de relación entre ambas variables (de interpretación similar a un coeficiente de correlación). Varía entre 0 (totalmente independientes) y 1 (totalmente dependientes). En este caso el valor 0,29, sugiere una clara relación entre *DAP y PARQUE*.

A continuación tenemos la media de la variable *DAP* para cada parque. Como puede observarse, con medias tan dispares (desde 3,11 euros hasta 0,53 euros) era probable que el análisis de la varianza revelara diferencias significativas.

Finalmente se presenta una prueba de homogeneidad de la varianza en cada grupo. Puede ocurrir que las medias sean diferentes pero con el mismo grado de dispersión de los datos. En este caso, el estadístico de `Hartley F` tiene un valor de 8,40. Este

estadístico, bajo la hipótesis nula de homogeneidad de la varianza, se distribuye como una $F_{5,49}$, para saber el valor crítico de esta distribución computamos:

Simulation → F-Distribution Plot → Enter the Type I Error Rate: *0.05*[14]; Enter the numerator degrees of Freedom: *5*; Enter the denominator degrees of Freedom: *49*.

Si comparamos el resultado tabulado de $F_{5,\ 49\ 0,05}$ =2,40 con el obtenido `Hartley F =8,40`, rechazamos la hipótesis nula de homogeneidad de la varianza, por lo que estamos incumpliendo uno de los requisitos de la prueba.

De igual manera, sin necesidad de obtener el valor crítico de la distribución F, podemos utilizar el estadístico de `Bartlett Chi-square` que nos proporciona OS4 junto con la probabilidad asociada al valor. En este caso el estadístico alcanza un valor igual a 45,83 lo cual implica una probabilidad asociada igual a 0,000, inferior a 0,05 por lo que rechazamos la hipótesis nula de homogeneidad de la varianza.

¿Es grave esta violación? Fijándonos en el número de casos de cada categoría vemos que la diferencia entre el grupo menor (parque número 2 con 24 casos) y el mayor (parque número 5 con 50 casos) sí es lo suficientemente amplia para no pasar por alto este incumplimiento. Ante este problema debemos abandonar el análisis de la varianza mediante ANOVA y utilizar la prueba no paramétrica de Kruskal-Wallis.

Prueba no paramétrica Kruskal-Wallis

Esta prueba no paramétrica de análisis de la varianza se puede utilizar cuando incumplimos el requisito de igualdad de la varianza entre grupos (homocedasticidad) y el número de casos de cada grupo está muy desequilibrado. Siguiendo con el ejemplo anterior tenemos:

Analyses → Nonparametric → Kruskal Wallace One-Way ANOVA → Treatment group codes: *Parque*; The dependent variable: *Dap*.

```
Kruskal - Wallis One-Way Analysis of Variance
      Score         Rank      Group
       0.00        40.00          3
        ...
       6.01       200.00          1

Sum of Ranks in each Group
      Group          Sum   No. in Group
          1         3526             38
          2         3923             24
          3         4601             40
          4         4238             48
          5         3811             50

No. of tied rank groups = 9
Statistic H uncorrected for ties =42.5480
Correction for Ties =0.9299
Statistic H corrected for ties =45.7558
Corrected H is approx. chi-square with 4 D.F. and probability =0.0000
```

La hipótesis nula de igualdad de las cinco distribuciones (una por cada parque), y por tanto no influencia del factor sobre la variable dependiente, es rechazada ya que tiene una probabilidad igual a 0,0000 (inferior a 0,05). Así, habiendo determinado un efecto estadísticamente significativo de la variable PARQUE sobre la variable dependiente DAP procedemos a identificar qué parques son los responsables del mismo. Para ello obtenemos el valor medio de la variable DAP en cada parque:

Analyses → Descriptive → Breakdown → Categorical Variables Selected: *Parque*; Continuous Variable to Break Down: *Dap*.

```
BREAKDOWN ANALYSIS PROGRAM

PARQUE     level =    1
Freq.      Mean       Std. Dev.
38         1.087      1.467

PARQUE     level =    2
Freq.      Mean       Std. Dev.
24         3.108      1.660

PARQUE     level =    3
Freq.      Mean       Std. Dev.
40         1.405      1.200

PARQUE     level =    4
Freq.      Mean       Std. Dev.
48         0.864      1.008

PARQUE     level =    5
Freq.      Mean       Std. Dev.
50         0.534      0.573

Number of observations accross levels =200
Mean accross levels = 1.201
Std. Dev. accross levels = 1.382
```

De acuerdo con las medias de los cinco parques vemos que existe uno, el parque 2, con una media muy superior al resto (3,1 euros frente a una media global de 1,2). Sería interesante aislar el efecto de este parque, el Parque Nacional de Segura, Cazorla y las Villas para ser más precisos, del resto. Para ello crearemos una nueva variable, CAZORLA, que tomará el valor 1 para el visitante de este parque y el valor 0 para el visitante del resto de parques. Situándonos en la columna de la variable PARQUE:

Variables → Recode; ⊙Into a new column, (para crear otra variable en otra columna);

New Name: *CAZORLA*,; New value: ⊙Value: *1*,; Old value: ⊙Value: *2*, (el valor 2 en la variable PARQUE se convierte en 1 en la nueva variable CAZORLA) → Apply.

Figura 5.6. *Creación de la variable Cazorla. Paso 1*

La nueva columna, CAZORLA, sólo tiene el valor 1 en el parque de Cazorla. Para asignar el 0 al resto de parques nos situamos en cualquier celda de la nueva columna:

Variables → Recode; ⊙Into the same column; New value:⊙Value:0, Old value:⊙Blanks → Apply.

Figura 5.7. *Creación de la variable Cazorla. Paso 2*

Y obtenemos el fichero con la nueva variable CAZORLA. Para comprobar el efecto neto de este nueva variable podemos repetir el análisis de la varianza:

Analyses → Nonparametric → Kruskal Wallace One-Way ANOVA → Treatment group codes: *Cazorla*; The dependent variable: *Dap*.

```
Kruskal - Wallis One-Way Analysis of Variance

                  Score           Rank       Group
                   0.00          40.00           1
                   0.00          40.00           1
                    ...
                   6.01         197.50           2
                   6.01         200.00           1

Sum of Ranks in each Group
      Group         Sum   No. in Group
          0    16176.50            176
          1     3923.50             24

Statistic H corrected for ties =34.7252
Corrected H is approx. chi-square with 1 D.F. and probability =0.0000
```

En este caso, como era de esperar, la probabilidad de que la variable CAZORLA no tenga ninguna influencia en la variable DAP es cero (0,0000).

5.5. Análisis de correlación

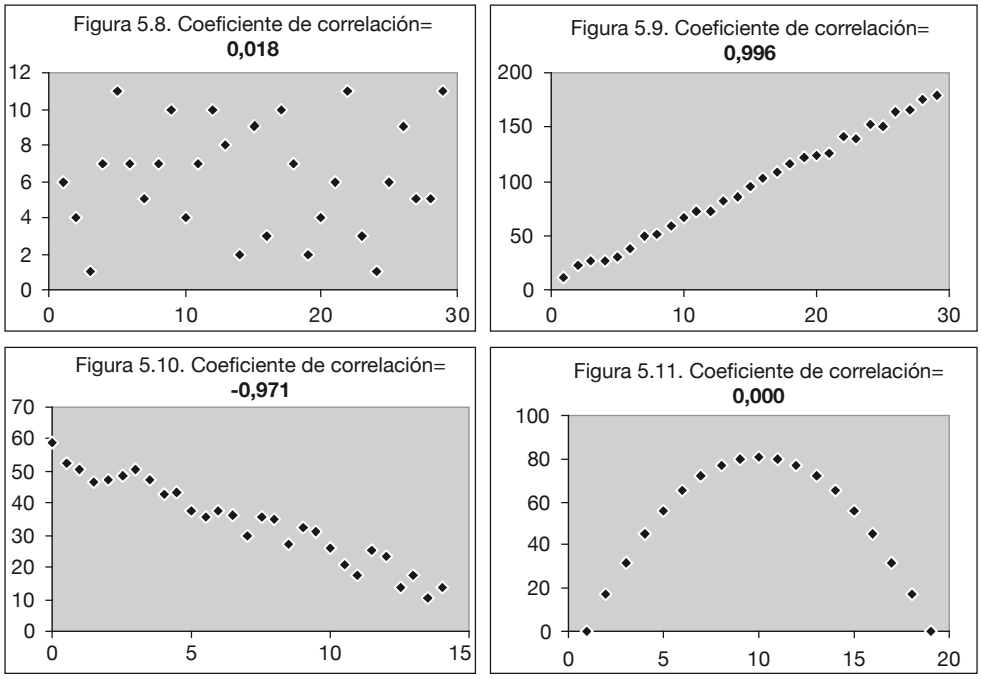
El objetivo de este análisis es cuantificar la relación entre dos variables métricas u ordinales. Si bien un alto grado de correlación, medido por el coeficiente de correlación, puede sugerir una relación causa-efecto (por ejemplo dosis de fertilizantes y rendimiento de una parcela) lo contrario no es siempre cierto. Esto es debido a que el análisis de correlación mide el grado de relación lineal entre dos variables. Conviene además aclarar que un coeficiente de correlación alto no implica necesariamente una relación de causalidad, ya que dos variables independientes pueden moverse en la misma dirección por efecto de otra tercera sin ser una el resultado de la otra.

Para el análisis de correlación calculamos el coeficiente de Pearson (prueba paramétrica) y los coeficientes de Spearman y Kendall tau[15] (pruebas no paramétricas). Estos coeficientes de correlación varían entre –1 (relación lineal negativa perfecta) y +1 (relación lineal positiva perfecta).

En la Figura 5.8. las dos variables son independientes por lo que el coeficiente de correlación es prácticamente 0. En la Figura 5.9 la relación es positiva y casi lineal por lo que el coeficiente de correlación alcanza un valor cercano al 1. Igualmente en la Figura 5.10 la relación es casi lineal pero negativa por lo que el valor se aproxima a –1. Finalmente en la Figura 5.11 el coeficiente de correlación es 0 pero, a diferencia de la Figura 5.8, las dos variables tienen una relación de dependencia perfecta ($y=- x^2 + 20x – 19$). A la vista de este último caso, conviene ser cauteloso a la hora de afirmar

la independencia de dos variables por tener un coeficiente de correlación cercano a cero.

Figura 5.8 a 5.11. *Ejemplos de correlación entre dos variables*



Para determinar el tipo de coeficiente de correlación que se debe utilizar en el análisis es necesario considerar el tipo de variable y el tamaño muestral. Esquemáticamente podemos indicar:

Ordinal-ordinal

- Número de categorías de ambas ordinales ≥ 5 → Coeficiente de Spearman.
- Número de categorías de una o ambas < 5 → Coeficiente de Kendall tau[16].

Métrica-métrica

- n (tamaño de la muestra) ≥ 100 → Coeficiente de Pearson.
- n < 100 y distribución normal de ambas → Coeficiente de Pearson.
- n < 100 y distribución no normal de al menos una → Coeficiente de Spearman.

Métrica-ordinal

- Número de categorías < 5 → Coeficiente de Kendall tau.
- Número de categorías ≥ 5 y n ≥ 100 → Coeficiente de Pearson.
- Número de categorías ≥ 5 y n < 100 → Coeficiente de Spearman.

Coeficiente de correlación de Pearson

Consideremos las variables DAP (cantidad de dinero que el visitante está dispuesto a pagar por la entrada al parque) y DISTANCIA (distancia recorrida para llegar al parque). El análisis de correlación se lleva a cabo mediante las instrucciones:

Analyses → Correlation → Product-Moment Correlations: *Dap; Distancia.*

```
Correlations Matrix

Variables
                      DAP     DISTANCIA
        DAP         1.000         0.765
  DISTANCIA         0.765         1.000
```

Como indica el coeficiente de correlación de Pearson, existe una fuerte correlación positiva entre las variables. Parece lógico pensar que a medida que el visitante recorre más kilómetros para visitar el parque mayor será su disposición a pagar una entrada por su menor importancia relativa dentro de los costes totales en que incurre.

Coeficiente de correlación de Spearman

Utilizando de nuevo los datos sobre incidencia de infartos, procedemos a analizar la relación entre el peso del individuo y su edad. Teniendo en cuenta el tamaño de la muestra, 50 casos, el primer paso consiste en determinar si las variables siguen aproximadamente una distribución normal. Para ello basta con utilizar el comando de análisis de normalidad:

Analyses → Descriptive → Normality Tests → Test normality of: *Peso →Apply.*

Repitiendo el mismo análisis con la variable EDAD, resumimos los resultados de las pruebas en la tabla siguiente:

Tabla 5.3. *Pruebas de normalidad de las variables PESO y EDAD*

Variable	Shapiro-Wilkes		Lilliefors	
	W	Probabilidad	L	Conclusión
PESO	0,97	0,244	0,084	*No evidence against normality*
EDAD	0,92	0,032	0,143	*Strong evidence against normalit.*

En el caso de la variable PESO tanto la prueba de Shapiro-Wilkes (probabilidad inferior a 0,05) como la prueba de Lilliefors (*No evidence against normality*) sugieren que dicha variable sigue aproximadamente una distribución normal. Sin embargo, con respecto a la variable EDAD llegamos a la conclusión contraria. Por tanto, siendo una de las variables métricas no normal procedemos a calcular el coeficiente de correlación de Spearman:

Analyses → Nonparametric → Spearman Rank Correlation → *Peso, Edad.*

```
…
Spearman Rank Correlation =0.034
t-test value for hypothesis r =0 is 0.237
Probability > t =0.8138
```

Como indica el coeficiente de correlación de Spearman (`r=0,034`), prácticamente cero, no existe relación entre ambas variables. Esta afirmación se ve corroborada por la probabilidad de la hipótesis nula de la prueba, esto es, el coeficiente de correlación es cero, cuyo valor se sitúa en 0,8138. Como es habitual, una probabilidad del estadístico t superior a 0,05 implica la aceptación de la hipótesis nula, es decir, que ambas variables son independientes. Resumiendo:

- `Spearman Rank Correlation =`indica el grado de correlación entre ambas variables, siendo mayor cuanto mayor es su valor absoluto (entre –1 y +1).
- `Probability > t =`si es menor que 0,05 rechazamos H_0, esto es, rechazamos que el coeficiente de correlación sea cero y por tanto aceptamos que existe una relación estadísticamente significativa entre las variables.

Coeficiente de correlación de Kendall Tau

Siguiendo con los mismos datos ¿existe alguna relación entre el grado de sedentarismo del individuo y su edad? En este caso, la variable ordinal EJERCICIO tiene cuatro categorías por lo que es apropiado utilizar el coeficiente de correlación Kendall Tau

Analyses → Nonparametric → Kendall's Rank Correlation Tau and Partial Tau → X Variable: *Ejercicio*; Y Variable: *Edad*.

```
Kendall Tau for File: Cardio.OS4

Kendall Tau for variables Ejercicio and Edad
Tau =-0.3334 z =3.416 probability > |z| =0.000

NOTE: Probabilities are for large N (>10)
```

El valor del estadístico de Tau indica una relación negativa entre la práctica de ejercicio y la edad, es decir, los individuos de mayor edad dedican menos horas a la semana al ejercicio físico. La hipótesis nula de esta prueba es que el coeficiente de correlación es igual a cero (es decir, que ambas variables son independientes). En nuestro ejemplo la probabilidad de esta hipótesis es 0,000, por tanto, al ser inferior a 0,05, nos hace rechazar la hipótesis nula.

La ignorancia afirma o niega rotundamente;
la ciencia duda
(Voltaire)

CAPÍTULO 6

ANÁLISIS DE LA VARIANZA

Capítulo 6. Análisis de la varianza

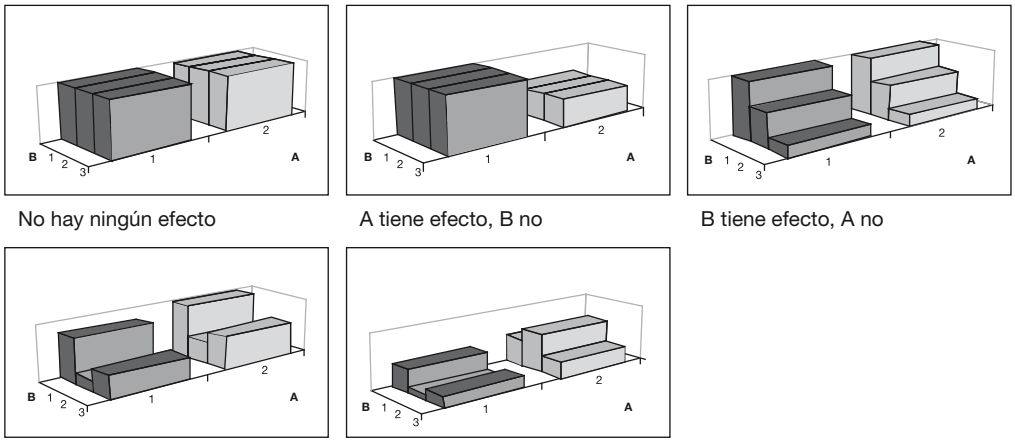
6.1. Análisis de la varianza con una variable dependiente

Prueba paramétrica ANOVA II

Con ANOVA II se analiza el efecto de dos variables nominales (o factores) sobre una variable métrica. Al igual que en ANOVA I, la variable métrica debe cumplir con todos los requisitos paramétricos, aunque en la práctica se suele continuar con ANOVA si no se incumple el requisito de homocedasticidad. En el caso de incumplir este requisito podemos intentar estabilizar la varianza transformando la variable dependiente, substituyendo ésta por su raíz cuadrada o por su logaritmo neperiano. Estas transformaciones no siempre consiguen su objetivo, por lo que, de persistir el problema, es necesario indicar en el análisis de la varianza el grado de incumplimiento de este supuesto.

Con el objeto de visualizar el efecto de dos factores sobre una variable métrica, la siguiente figura presenta las posibles alternativas.

Figura 6.1. *Posibles efectos de dos factores sobre una variable dependiente* (el factor A con dos categorías y el B con tres, la variable dependiente se mide por la altura del sólido)



No hay ningún efecto | A tiene efecto, B no | B tiene efecto, A no

A y B tienen efecto | A y B tienen efecto y además hay interacción

Aplicación de ANOVA II

Continuando con las encuestas realizadas en los parques naturales, estudiaremos el efecto simultáneo de las variables PARQUE y SEXO (factor 1 y 2) sobre la variable dependiente métrica DAP (disposición a pagar una entrada por la visita al parque):

Analyses → Analyses of Variance → One, Two or Three Way ANOVA → Dependent Variable: *Dap*; Factor 1 Variable: *Parque*; Factor 2 Variable: *Sexo*.

Figura 6.2. *Cuadro de diálogo de ANOVA II*

Tras presionar “Continue” obtenemos:

```
Two Way Analysis of Variance
```

```
Variable analyzed: DAP

Factor A (rows) variable: PARQUE (Fixed Levels)
Factor B (columns) variable: SEXO (Fixed Levels)

SOURCE          D.F.       SS      MS       F  PROB.> F  Omega Squared
Among Rows         4  117.129  29.282  22.512     0.000          0.293
Among Columns      1    0.190   0.190   0.146     0.703          0.000
Interaction        4   15.664   3.916   3.011     0.019          0.027
Within Groups    190  247.137   1.301
Total            199  380.120   1.910

Omega squared for combined effects =0.318
...
TESTS FOR HOMOGENEITY OF VARIANCE
------------------------------------------------------------------------
Hartley Fmax test statistic =14.92 with deg.s freedom: 10 and 24.
Cochran C statistic =0.24 with deg.s freedom: 10 and 24.
Bartlett Chi-square statistic =140.27 with 9 D.F. Prob. larger value =0.000
------------------------------------------------------------------------
...
```

Centrándonos en la primera parte del resultado podemos ver cómo la variable PARQUE (*rows*) sí tiene un efecto estadísticamente discernible sobre DAP (con una F=22,512, PROB.>F=0,000, lo que implica un 0,0% de probabilidad de que ambas variables sean independientes). Por el contrario, la variable SEXO (*columns*) no está relacionada con la variable DAP (F=0,146, PROB.>F=0,703, indicando un 70,3% de probabilidad de que ambas variables sean independientes). El efecto combinado de ambas es alto (0,318), si bien hay que atribuirlo en su mayor parte a la variable PARQUE (0,293).

En la segunda parte podemos comprobar cómo el estadístico de Hartley, cuya hipótesis nula es la homocedasticidad (no diferencia de varianzas entre grupos), tiene un valor de 14,92, muy superior al valor crítico del estadístico $F_{10, 24\ 0,05}$ =2,26. Esta valor crítico, en caso de no tener una tabla tabulada de F podemos calcularlo de la siguiente forma:

Simulation → F-Distribution Plot → Enter the Type I Error Rate: *0.05*; Enter the numerator degrees of Freedom: *10*; Enter the denominator degrees of Freedom: *24*.

Al exceder el valor calculado del estadístico (14,92) el valor crítico (2,26) rechazamos la hipótesis nula de homocedasticidad. De igual manera, el estadístico de Barlett con un valor de 140,27 tiene una probabilidad asociada de 0,000, lo que también nos hace rechazar dicha hipótesis nula. A la vista de este resultado sería recomendable estabilizar la varianza de la variable DAP, como explicamos en el siguiente apartado.

Estabilización de la varianza

En la práctica, para estabilizar la varianza procedemos a transformar la variable dependiente obteniendo su raíz cuadrada o su logaritmo neperiano. Una u otra transformación suelen ser suficientes para eliminar, o reducir, el problema de la heterocedasticidad. En la transformación logarítmica deberá tenerse en cuenta los siguientes aspectos[17]:

- Si hay valores iguales a cero se suma una unidad a todos los valores, o
- si existen valores decimales conviene multiplicar todos los valores por 10 (o 100 si son dos decimales, etc.).

Por tanto, el primer paso consiste en crear una nueva variable RAIZDAP que es la raíz cuadrada de la variable DAP. Para ello:

Variables → Transform → First Var. Argument (V1): *Dap*; Constant: *0.5*; Save new variable as: *Raizdap*; Select Transformation: *New=V1**C* → OK.

Repetimos el análisis de la varianza con la variable RAIZDAP en lugar de DAP:

Analyses → Analyses of Variance → One, Two or Three Way ANOVA → Dependent Variable: *Raizdap*; Factor 1 Variable: *Parque*; Factor 2 Variable: *Sexo*.

```
Two Way Analysis of Variance

Variable analyzed: Raizdap

Factor A (rows) variable: PARQUE (Fixed Levels)
Factor B (columns) variable: SEXO (Fixed Levels)

SOURCE          D.F.        SS      MS       F  PROB.> F    Omega Squared
Among Rows         4    23.299   5.825  13.621     0.000        0.200
Among Columns      1     0.009   0.009   0.021     0.884        0.000
Interaction        4     2.904   0.726   1.698     0.152        0.011
Within Groups    190    81.250   0.428
Total            199   107.463   0.540

Omega squared for combined effects =0.207

...

TESTS FOR HOMOGENEITY OF VARIANCE
------------------------------------------------------------------------
Hartley Fmax test statistic =3.19 with deg.s freedom: 10 and 24.
Cochran C statistic =0.17 with deg.s freedom: 10 and 24.
Bartlett Chi-square statistic =29.40 with 9 D.F. Prob. =0.001
------------------------------------------------------------------------
```

Si bien no hemos corregido completamente el problema de heterocedasticidad, éste se ha reducido considerablemente, pasando de un valor del estadístico de Hartley de 14,92 a 3,19 (en el caso del estadístico de Barlett hemos pasado de 140,27 a 29,40).

A partir de estos resultados, debemos tomar con cautela las conclusiones que se derivan del análisis de la varianza. En el caso de aceptar la validez del análisis actual, las conclusiones a las que llegamos sobre el efecto de los factores PARQUE y SEXO sobre la variable RAIZDAP son los mismos que los apuntados anteriormente: se descarta el efecto SEXO y se acepta el efecto del PARQUE sobre la variable dependiente.

Pruebas post-hoc de diferencias entre grupos

El análisis de la varianza permite no sólo determinar si los factores tienen o no influencia en la variable dependiente sino también descubrir qué grupos dentro de cada factor tienen una media estadísticamente diferente del resto de grupos. Para este propósito utilizamos las pruebas *post-hoc* de Scheffe y Tukey. La primera está indicada ante un número desigual de casos por grupo (situación más frecuente) mientras la segunda se utiliza cuando este número es similar. Repetimos el análisis de la varianza marcando la opción Scheffe:

Analyses → Analyses of Variance → One, Two or Three Way ANOVA → Dependent Variable: *Raizdap*; Factor 1 Variable: *Parque*; Factor 2 Variable: *Sexo -marcar la opción Scheffe-.*

Two Way Analysis of Variance

Variable analyzed: raizdap

Factor A (rows) variable: PARQUE (Fixed Levels)
Factor B (columns) variable: SEXO (Fixed Levels)

...
Descriptive Statistics

GROUP	Row	Col.	N	MEAN	VARIANCE	STD.DEV.
Cell	**1**	**1**	**25**	**0.806**	0.715	0.845
Cell	**1**	**2**	**13**	**0.478**	0.400	0.633
Cell	2	1	15	1.507	0.435	0.659
Cell	2	2	9	1.882	0.287	0.536
Cell	3	1	23	0.981	0.628	0.793
Cell	3	2	17	0.936	0.321	0.567
Cell	4	1	26	0.767	0.445	0.667
Cell	4	2	22	0.542	0.409	0.639
Cell	5	1	34	0.453	0.224	0.473
Cell	5	2	16	0.684	0.322	0.568
Row	1	**38**	**0.694**	0.618	0.786	
Row	2	24	1.647	0.399	0.632	
Row	3	40	0.962	0.487	0.698	
Row	4	48	0.664	0.432	0.657	
Row	5	50	0.527	0.261	0.511	
Col	1	**123**	**0.818**	0.557	0.746	
Col	2	77	0.805	0.520	0.721	
TOTAL	200	0.813	0.540	0.735		

Según la tabla anterior, en la muestra hay 25 hombres en el parque 1, cuya DAP media es igual a 0,806 euros. El número de mujeres en este mismo parque es igual a 13, con una DAP media de 0,478 euros.
También se indican los datos agregados en cada una de las categorías de los factores: 38 personas en el parque número 1 con una DAP media de 0,694 euros.
Del total de personas en todos los parques 123 son hombres y tienen una DAP media igual a 0,818 euros.

...
COMPARISONS AMONG ROWS *(comparación de DAP entre parques)*

Scheffe contrasts among pairs of means. Alpha selected =0.05

Group vs	Group	Difference	Scheffe Statistic	Critical Value	Significant?
1	**2**	**-0.95**	**5.59**	**3.110**	**YES**
1	3	-0.27	1.81	3.110	NO
1	4	0.03	0.21	3.110	NO
1	5	0.17	1.19	3.110	NO
2	3	0.69	4.06	3.110	YES
2	4	0.98	6.02	3.110	YES
2	5	1.12	6.90	3.110	YES
3	4	0.30	2.13	3.110	NO

```
    3        5        0.44        3.14        3.110        YES
    4        5        0.14        1.04        3.110        NO
-------------------------------------------------------------
```

En el primer caso, la comparación entre el parque 1 y el 2, el valor calculado del estadístico de Scheffe es igual a 5,59, por encima de su valor crítico que es 3,11, por tanto rechazamos la hipótesis nula (no diferencia entre grupos) y aceptamos la hipótesis alternativa (las medias son estadísticamente diferentes).

```
COMPARISONS AMONG COLUMNS WITHIN EACH ROW
ROW 1 COMPARISONS (comparamos la DAP media entre hombres y mujeres en el parque 1)
-------------------------------------------------------------
                Scheffe contrasts among pairs of means.
Group vs  Group      Diff        Scheffe      Crit        Significant?
                                  stat.      value
   ----------------------------------------------------------------
    1        2        0.33        1.47        2.064        NO
-------------------------------------------------------------
ROW 2 COMPARISONS (comparamos la DAP media entre hombres y mujeres en el parque 2)
-------------------------------------------------------------
Group vs  Group      Diff        Scheffe      Crit        Significant?
                                  stat.      value
-------------------------------------------------------------
    1        2       -0.38        1.36        2.064        NO
-------------------------------------------------------------
ROW 3 COMPARISONS (comparamos la DAP media entre hombres y mujeres en el parque 3)
-------------------------------------------------------------
Group vs  Group      Diff        Scheffe      Crit        Significant?
                                  stat.      value
-------------------------------------------------------------
    1        2        0.04        0.21        2.064        NO
-------------------------------------------------------------
ROW 4 COMPARISONS (comparamos la DAP media entre hombres y mujeres en el parque 4)
-------------------------------------------------------------
Group vs  Group      Diff        Scheffe      Crit        Significant?
                                  stat.      value
-------------------------------------------------------------
    1        2        0.22        1.19        2.064        NO
-------------------------------------------------------------
ROW 5 COMPARISONS (comparamos la DAP media entre hombres y mujeres en el parque 5)
-------------------------------------------------------------
Group vs  Group      Diff        Scheffe      Crit        Significant?
                                  stat.      value
-------------------------------------------------------------
    1        2       -0.23        1.16        2.064        NO
-------------------------------------------------------------
```

Según indica la última columna, en ningún parque es estadísticamente significativa la diferencia de DAP media entre hombres (Group 1) y mujeres (Group 2).

```
COMPARISONS AMONG ROWS WITHIN EACH COLUMN
COLUMN 1 COMPARISONS (comparamos la media de dos parques para los hombres)
-------------------------------------------------------------
                Scheffe contrasts among pairs of means.
Group vs  Group  Difference    Scheffe    Critical    Significant?
                              Statistic    Value
-------------------------------------------------------------
    1        2       -0.70        3.28        3.110        YES
    1        3       -0.17        0.92        3.110        NO
    1        4        0.04        0.22        3.110        NO
    1        5        0.35        2.05        3.110        NO
```

2	3	0.53	2.42	3.110	NO
2	4	0.74	3.49	3.110	**YES**
2	5	1.05	5.20	3.110	**YES**
3	4	0.21	1.14	3.110	NO
3	5	0.53	2.99	3.110	NO
4	5	0.31	1.84	3.110	NO

Considerando exclusivamente a los hombres, existe una DAP media estadísticamente diferente entre los parques 1 y 2, los parques 2 y 4 y los parques 2 y 5.

COLUMN 2 COMPARISONS (*comparamos la media de dos parques para las mujeres*)

Scheffe contrasts among pairs of means.
alpha selected =0.05

Group vs	Group	Difference	Scheffe Statistic	Critical Value	Significant?
1	2	-1.40	4.95	3.110	**YES**
1	3	-0.46	1.90	3.110	NO
1	4	-0.06	0.28	3.110	NO
1	5	-0.21	0.84	3.110	NO
2	3	0.95	3.51	3.110	**YES**
2	4	1.34	5.18	3.110	**YES**
2	5	1.20	4.40	3.110	**YES**
3	4	0.39	1.87	3.110	NO
3	5	0.25	1.11	3.110	NO
4	5	-0.14	0.66	3.110	NO

Considerando exclusivamente a las mujeres, existe una DAP media estadísticamente diferente entre el parque 2 y el resto de parques.

Prueba paramétrica ANOVA n

Al igual que hemos analizado el efecto de dos factores (variables nominales) sobre una variable dependiente (variable métrica), es posible extender el análisis incluyendo tres (ANOVA III) o más factores (ANOVA n). El paquete estadístico OS4 permite analizar hasta tres factores simultáneamente. Veamos, por ejemplo, si existe relación entre la distancia recorrida por el visitante para visitar el parque y tres variables explicativas: Factor 1=el parque es Cazorla; Factor 2=el SEXO del visitante; y Factor 3=si es la primera visita.

Analyses → Analyses of Variance → One, Two or Three Way ANOVA → Dependent Variable: *Distancia*; Factor 1 Variable: *Cazorla*; Factor 2 Variable: *Sexo*; Factor 3 Variable: *Primera*.

OS4 muestra el siguiente mensaje:

> *"ERROR! A negative SS found. Unbalanced Design? Ending analysis"*

Es decir, que no podemos continuar con ANOVA ya que el diseño no es equilibrado. Este problema surge por alguna de las siguientes causas:

- El número de casos en cada celda no es igual para cada factor o las diferencias no son proporcionales a lo largo de cada fila.
- No existen casos en alguna celda.

Esta circunstancia produce un error en la partición tradicional de la varianza utilizada por ANOVA ya que los efectos no son ortogonales (Searle, 1971, p. 138; Shaw and Mitchell-Olds, 1993). En estos casos podemos analizar el efecto de varios factores sobre una variable métrica mediante el Modelo Lineal General (MLG, o su acrónimo anglosajón GLM). Esta técnica puede considerarse como una extensión del modelo de regresión múltiple y permite analizar el efecto de cualquier tipo de variable sobre varias variables dependientes de forma simultánea. Debido a su flexibilidad hemos dedicado un capítulo exclusivo a su uso una vez abordados los capítulos de análisis de la varianza y análisis de regresión.

6.2. Análisis de la varianza con dos o más variables dependientes

Justificación del análisis multivariante de la varianza (MANOVA)

El análisis multivariante de la varianza (MANOVA) puede considerarse una extensión del análisis de la varianza (ANOVA). En efecto, mientras ANOVA analiza el efecto de los factores sobre una única variable dependiente, MANOVA considera simultáneamente su efecto en más de una variable dependiente[18].

¿Cuáles son las ventajas de usar MANOVA en lugar de una prueba ANOVA para cada una de las variables dependientes? Básicamente, dos son las ventajas:

1. A diferencia de múltiples ANOVA, con MANOVA no incrementamos el error de tipo I, esto es, la probabilidad de rechazar una hipótesis verdadera. Por ejemplo, con dos pruebas ANOVA la probabilidad de no incurrir en ningún error del Tipo I, asumiendo que son independientes, es igual a $0,95 \cdot 0,95 = 0,86$, o lo que es lo mismo, la probabilidad de incurrir al menos en un error del Tipo I es igual a 1- 0,86=0,14, muy por encima del máximo fijado en 0,05 en MANOVA.
2. Con MANOVA se consideran las posible interacciones entre variables dependientes. En este sentido, es posible que no podamos diferenciar grupos en función de una sola variable dependiente pero sí hacerlo incluyendo varias simultáneamente.

Requisitos paramétricos de MANOVA

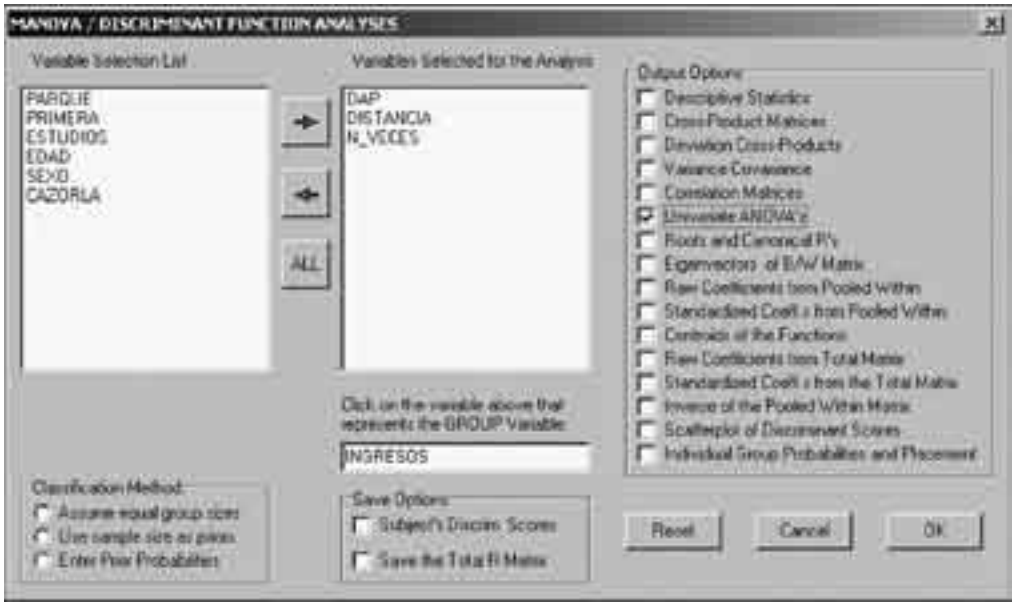
Al igual que en ANOVA se requería que la variable métrica no incumpliera el requisito de normalidad y que las varianzas internas de los grupos fueran iguales, MANOVA extiende estos requisitos hacia una dimensión multivariante. Aunque raramente éstos se satisfacen, desde un punto de vista práctico, y gracias a la robustez de los estadísticos que se usan, podemos sacar conclusiones válidas incluso en situaciones de no normalidad multivariante y/o de heterocedasticidad (desigualdad de las varianzas entre grupos).

Ejemplo MANOVA

Vamos a estudiar el efecto de la variable INGRESOS sobre la disposición a pagar una entrada (DAP), el número de veces que visita el parque (N_VECES) y la distancia recorrida hasta el mismo (DISTANCIA). Nótese que la variable independiente (INGRESOS) en este caso es una variable ordinal, sin embargo en este análisis se trata como una variable nominal. En efecto, MANOVA no tendrá en cuenta que los individuos del grupo 1 cobran menos que los individuos del grupo 4 sino simplemente que pertenecen a grupos distintos[19]. El análisis se lleva a cabo con las instrucciones:

Analyses → Multivariate → Discriminant Function / MANOVA → Variables selected for the analysis: *Dap, Distancia, N_veces*; GROUP Variable: *Ingresos*; ⊙ Univariate ANOVA's -

Figura 6.3. *Cuadro de diálogo de MANOVA*



Nota: Para introducir la variable grupo en el cuadro inferior (GROUP variable) primero se introduce en el cuadro superior al igual que el resto de variables y después se vuelve a seleccionar.

La primera pantalla de resultados muestra el análisis de la varianza univariante (es decir, ANOVA I):

```
MULTIVARIATE ANOVA / DISCRIMINANT FUNCTION

Total Cases:=200, Number of Groups:=4

UNIVARIATE ANOVA FOR VARIABLE DAP
SOURCE              DF            SS           MS        F    PROB > F
BETWEEN              3       116.368       38.789   28.825       0.000
ERROR              196       263.752        1.346
TOTAL              199       380.120

UNIVARIATE ANOVA FOR VARIABLE DISTANCIA
SOURCE              DF            SS           MS        F    PROB > F
BETWEEN              3    210096.627    70032.209    7.713       0.000
ERROR              196   1779708.553     9080.146
TOTAL              199   1989805.180

UNIVARIATE ANOVA FOR VARIABLE N_VECES
SOURCE              DF            SS           MS        F    PROB > F
BETWEEN              3      1561.633      520.544    1.149       0.331
ERROR              196     88805.562      453.090
TOTAL              199     90367.195
```

El análisis univariante de la varianza sugiere una clara influencia de la variable INGRESOS sobre las variables DAP y DISTANCIA (en ambos casos la probabilidad de la hipótesis nula de independencia es de 0,000). Por el contrario, es probable que los ingresos no influyan en el número de veces que el usuario visita el parque (0,331>0,05→ no se rechaza la hipótesis nula de independencia entre ambas).

Procedemos a eliminar la variable N_VECES y volvemos a calcular MANOVA con las variables DAP y DISTANCIA y la variable de grupo INGRESOS:

```
MULTIVARIATE ANOVA / DISCRIMINANT FUNCTION
Reference: Multiple Regression in Behavioral Research
Elazar J. Pedhazur, 1997, Chapters 20-21
Harcourt Brace College Publishers

Total Cases:=200, Number of Groups:=4

UNIVARIATE ANOVA FOR VARIABLE DAP
SOURCE      DF          SS           MS          F          PROB > F
BETWEEN     3           116.368      38.789      28.825     0.000
ERROR       196         263.752      1.346
TOTAL       199         380.120

UNIVARIATE ANOVA FOR VARIABLE DISTANCIA
SOURCE      DF          SS           MS          F          PROB > F
BETWEEN     3           210096.627   70032.209   7.713      0.000
ERROR       196         1779708.553  9080.146
TOTAL       199         1989805.180
```

Mientras MANOVA puede indicarnos la posible existencia de una relación entre la variable nominal (u ordinal) y las variables dependientes, la explicación de esa relación y su cuantificación se consigue a través del análisis discriminante, técnica que

abordaremos en el siguiente capítulo. En este sentido, el análisis discriminante puede considerarse como la continuación lógica de MANOVA. Esta es la causa por la que el paquete estadístico OS4 engloba ambas técnicas en el mismo comando (Analyses→ Multivariate→Discriminant Function / MANOVA). Limitándonos por ahora a MANOVA, utilizamos la primera pantalla de resultados (ANOVAs) y la última:

```
Corr.s Between Variables and Functions with 200 valid cases.

Variables
                              1              2
           DAP            0.954          0.299
     DISTANCIA            0.537          0.844

Wilkes Lambda =0.6566.
F =15.2148 with D.F. 6 and 390. Prob > F =0.0000
Bartlett Chi-Square =82.4456 with 4 D.F. and prob. =0.0000
Pillia Trace =0.3476
```

El valor del estadístico de Wilkes Lambda alcanza un valor de 0,6566, lo que se traduce en un valor de F igual a 15,21 con una probabilidad de ocurrencia de 0,0000. Esta probabilidad supone el rechazo de la hipótesis nula de no diferencia entre los grupos de INGRESOS respecto a las variables DAP y DISTANCIA, por tanto la variable INGRESOS sí influye en la cantidad que el visitante está dispuesto a pagar por la entrada y en la distancia que recorre para visitar dicho parque.

De nuevo hay que tomar las conclusiones con cautela debido al incumplimiento del requisito de homogeneidad de la varianza entre grupos ya que el estadístico de Barlett arroja una probabilidad igual a 0,000, rechazando la hipótesis nula de homogeneidad de la varianza.

Como hemos explicado anteriormente, MANOVA permite analizar el efecto de un factor (F), sobre dos o más variables dependientes (Y_1, Y_2, ...Y_n) simultáneamente, de forma que puede darse el caso de no encontrar diferencias significativas analizando la varianza de forma univariante (ANOVA F-Y_1, ANOVA F-Y_2,..., ANOVA F-Y_n) y sí descubrirlas mediante el análisis multivariante de la varianza.

El hombre que ha cometido un error y no lo corrige,
comete otro error mayor
(Confucio)

CAPÍTULO 7

ANÁLISIS DISCRIMINANTE

Capítulo 7. Análisis discriminante

7.1. Introducción teórica

Con esta técnica multivariante se pretende explicar o predecir la pertenencia de un caso a un determinado grupo en función de un conjunto de variables explicativas. Por tanto, en este análisis la variable dependiente es de tipo categórico (nominal u ordinal) mientras las variables explicativas pueden ser de cualquier tipo. Por tanto, la idea central de esta técnica estadística consiste en determinar el peso de las distintas variables explicativas (X_1,·X_2,...·X_n) en la clasificación de los individuos en los diferentes grupos a partir de la función discriminante (D). Matemáticamente:

$D=b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + ... + b_n \cdot X_n$

Una vez estimados los pesos b_1, b_2,... b_n basta con sustituir los valores de las variables X_1, X_2,... X_n para obtener un valor de la función discriminante, el cual nos indicará el grupo al que pertenecerá con mayor probabilidad el individuo.

Como el lector puede apreciar, el análisis discriminante sólo se diferencia del análisis de regresión en la naturaleza categórica de la variable dependiente, es decir, la variable dependiente es de tipo nominal (por ejemplo, Consumidor A, Consumidor B y Consumidor C) u ordinal (por ejemplo, nivel de ingresos Alto, Medio, Bajo). En el caso de que la variable dependiente tenga sólo dos categorías (por ejemplo, sexo) se podría abordar el problema mediante un modelo de regresión logístico como alternativa al análisis discriminante.

La clasificación de los casos se realiza a través de las llamadas funciones canónicas (o factores). Estas funciones canónicas representan los ejes del hiperplano donde se sitúan los casos. Así, el procedimiento de clasificación de los casos sigue los siguiente pasos:

1. Obtención de las expresiones matemáticas de los factores, esto es, las funciones canónicas, en función de las variables explicativas.
2. Cálculo del valor de cada función canónica para cada caso.
3. Aplicación de la regla de Bayes de probabilidad condicional utilizando los valores canónicos y las probabilidades a *priori*, esto es, las frecuencias observadas de cada grupo.
4. Asignación de cada caso al grupo con mayor probabilidad de pertenencia.

7.2. Ejemplo de análisis discriminante con cuatro grupos

Comenzaremos con el análisis preliminar de las posibles relaciones entre el nivel de ingresos del visitante y un conjunto de variables explicativas:

Analyses → Multivariate → Discriminant Function/MANOVA → Variables selected for the analysis: *Dap, Estudios, Edad, Sexo*; GROUP Variable: *Ingresos*; ⊙ Univariate ANOVA's -

```
MULTIVARIATE ANOVA / DISCRIMINANT FUNCTION
Reference: Multiple Regression in Behavioral Research
Elazar J. Pedhazur, 1997, Chapters 20-21
Harcourt Brace College Publishers

Total Cases:=200, Number of Groups:=4

UNIVARIATE ANOVA FOR VARIABLE DAP
SOURCE                 DF         SS         MS          F    PROB > F
BETWEEN                 3    116.368     38.789     28.825       0.000
ERROR                 196    263.752      1.346
TOTAL                 199    380.120

UNIVARIATE ANOVA FOR VARIABLE ESTUDIOS
SOURCE                 DF         SS         MS          F    PROB > F
BETWEEN                 3      2.065      0.688      0.795       0.498
ERROR                 196    169.730      0.866
TOTAL                 199    171.795

UNIVARIATE ANOVA FOR VARIABLE EDAD
SOURCE                 DF         SS         MS          F    PROB > F
BETWEEN                 3     19.569      6.523      3.717       0.012
ERROR                 196    343.951      1.755
TOTAL                 199    363.520

UNIVARIATE ANOVA FOR VARIABLE SEXO
SOURCE                 DF         SS         MS          F    PROB > F
BETWEEN                 3      1.418      0.473      2.017       0.113
ERROR                 196     45.937      0.234
TOTAL                 199     47.355
```

En la primera pantalla de resultados vemos cómo las variables ESTUDIOS y SEXO tienen un valor de F pequeño (0,795 y 2,017, respectivamente), con probabilidades de ocurrencia superior a 0,05 (0,498 y 0,113, respectivamente) por lo que no podemos rechazar la hipótesis nula de independencia entre estas variables y el nivel de ingresos. A continuación procedemos a repetir el análisis sin estas variables sin la opción ANOVA:

Analyses → Multivariate → Discriminant Function / MANOVA → Variables selected for the analysis: *Dap*, *Edad*; GROUP Variable: *Ingresos*.

```
MULTIVARIATE ANOVA / DISCRIMINANT FUNCTION
Reference: Multiple Regression in Behavioral Research
Elazar J. Pedhazur, 1997, Chapters 20-21
Harcourt Brace College Publishers

Total Cases:=200, Number of Groups:=4

Number of roots extracted:=2
Percent of trace extracted:=100.0000
Roots of the W inverse time B Matrix

 No.     Root  Proportion   Canonical R  Chi-Square   D.F.      Prob.
  1    0.4416      0.8974        0.5535     81.1344      6      0.000
  2    0.0505      0.1026        0.2192      9.6288      2      0.008
```

En la columna **Root** de esta primera pantalla de resultados se computan el valor característico de la función discriminante (*eigenvalue*) cuya interpretación es similar al estadístico F de ANOVA.

El número de factores (**No.** o *variates*: 2 en nuestro caso) es igual al mínimo del par (*v,g*), siendo *v* el número de variables explicativas y *g* el número de grupos menos uno. Por tanto, en nuestro caso, *v*=2 y *g*=4-1=3, y así, el mínimo (2,3)=2, que será el número de factores que explicarán la pertenencia o no al grupo. Podemos considerar estos factores como dimensiones o ejes del hiperplano que se utilizan para situar a los casos.

La columna **Proportion** nos informa del peso de cada eje en la clasificación de los casos. Vemos cómo el primer eje tiene el mayor peso en la clasificación (89,74%). En ocasiones en las que un eje contribuye con un porcentaje muy bajo a la clasificación de los casos, por ejemplo inferior al 5%, podemos evaluar la conveniencia de incluirlo o no, decidiendo si el incremento del poder predictivo del modelo compensa su mayor complejidad. Así, si tuviéramos tres ejes y uno de ellos fuera poco significativo podríamos no considerarlo y situar los casos en un plano.

La correlación entre el nivel de ingresos y los factores se mide en la columna **Canonical R**. Este coeficiente alcanza un valor de R=0,55 para el primer factor, lo que implica que este único factor explica el 31% ($0{,}55^2$ =0,31) de la variabilidad de la variable dependiente INGRESOS. De igual forma podemos calcular la contribución del otro eje (4,8%) hasta totalizar aproximadamente un 36%.

Finalmente la columna **Chi-Square** y su probabilidad, **Prob.,** revela que los dos factores son estadísticamente significativos, rechazando la hipótesis nula de no influencia individual sobre el nivel de ingresos.

La siguiente pantalla de resultados muestra los coeficientes no estandarizados de la función canónica. A partir de ellos se obtienen los coeficientes estandarizados, equivalentes a los coeficientes beta del análisis de regresión. OS4 utiliza el método directo de cálculo de los coeficientes. A diferencia del método por pasos (*stepwise*), el método directo incluye todas las variables independientes en la función canónica[20].

```
Raw Function Coeff.s from Pooled Cov. with 200 valid cases.
Variables
                                  1               2
             DAP              0.859          -0.111
            EDAD              0.024           0.758

Raw Discriminant Function Constants with 200 valid cases.
Variables                         1               2
                             -1.113          -2.383
```

De acuerdo con estos resultados las funciones canónicas no estandarizadas serían:

- V_1=-1,113 + 0,859*DAP + 0,024*EDAD
- V_2=-2,383 - 0,111*DAP + 0,758*EDAD

Alternativamente, también es posible clasificar los casos mediante las funciones discriminantes de Fisher. Tanto las funciones canónicas como las de Fisher dan lugar a la misma clasificación de los casos.

```
Fisher Discriminant Functions

Group 1    Constant:=-3.732
Variable   Coefficient
       1    0.262
       2    2.028

Group 2    Constant:=-2.525
Variable   Coefficient
       1    0.387
       2    1.630

Group 3    Constant:=-4.704
Variable   Coefficient
       1    1.429
       2    1.831

Group 4    Constant:=-5.874
Variable   Coefficient
       1    1.765
       2    1.931
```

Así, las funciones de Fisher tendrían la siguiente expresión matemática:

- Grupo 1 =-3,732 + 0,262*DAP + 2,028*EDAD
- Grupo 2 =-2,525 + 0,387*DAP + 1,630*EDAD
- Grupo 3 =-4,704 + 1,429*DAP + 1,831*EDAD
- Grupo 4 =-5.874 + 1,765*DAP + 1,931*EDAD

Estas funciones son útiles para clasificar nuevos casos no incluidos en la muestra. Por ejemplo, imaginemos que conocemos la DAP (1,20) y la EDAD (3) de un individuo pero

no su nivel de ingresos. Utilizando las funciones de Fisher podemos predecir el grupo de ingresos con mayor probabilidad de pertenencia:

- Grupo 1 =-3,732 + 0,262*(1,20) + 2,028*(3) =2,67
- Grupo 2 =-2,525 + 0,387*(1,20) + 1,630*(3) =2,83
- Grupo 3 =-4,704 + 1,429*(1,20) + 1,831*(3) =2,50
- Grupo 4 =-5.874 + 1,765*(1,20) + 1,931*(3) =2,04

Según los valores anteriores, este individuo se situaría en el segundo grupo ya que es éste donde la función de Fisher alcanza su valor máximo (2,83). El siguiente grupo que le sigue en probabilidad es el primero (2,67).

Para clasificar automáticamente los casos de la muestra OS4 utiliza las funciones canónicas V_1 y V_2 en lugar de las funciones de Fisher, si bien ambos procedimiento darían la misma ordenación de probabilidad de pertenencia.

```
CLASSIFICATION OF CASES

                    HIGH PROBABILITY       SEC.D HIGH          DISCRIM
SUBJECT    ACTUAL   IN GROUP    P(G/D)     GROUP    P(G/D)       SCORE
ID NO.      GROUP
       1         2          2    0.4387         1    0.4325     -1.0396
                                                                 -0.1094
       2         1          3    0.3237         4    0.2484      0.5065
                                                                 -0.3089
       3         2          2    0.4387         1    0.4325     -1.0396
                                                                 -0.1094
       4         1          2    0.5275         1    0.3494     -1.0639
                                                                 -0.8674
```

La columna ACTUAL GROUP (**i**) de la salida de OS4 indica el grupo al que pertenece realmente el caso. A continuación aparece el grupo estimado con la mayor probabilidad de pertenencia. En el Caso 1, la función discriminante estima correctamente que es el Grupo 2 el que tiene la mayor probabilidad de contener a este caso (0,4387). La segunda mayor probabilidad (0,4325) lo sitúa en el Grupo 1. En la Tabla 7.1 resumimos los valores de los primeros cuatro casos y la clasificación de la función discriminante. En ella vemos cómo sólo clasifica correctamente los casos 1 y 3.

Tabla 7.1. *Asignaciones de grupo a partir del análisis discriminante*

	Variables explicativas		Grupo (nivel de ingresos)	
	DAP	**EDAD**	**Real**	**Estimado**
Caso 1	0,00	3	2	2
Caso 2	1,80	3	1	3
Caso 3	0,00	3	2	2
Caso 4	0,00	2	1	2

La columna **DISCRIM SCORE** muestra los valores de las funciones canónicas. Por ejemplo, para el primer caso (DAP=0,00 y EDAD=3) tenemos:

- V_1=-1,113 + 0,859*(0,00) + 0,024*(3) =-1,0396
- V_2=-2,383 - 0,111*(0,00) + 0,758*(3) =-0,1094

OS4 resume el número de casos clasificados correctamente en la tabla siguiente:

CLASSIFICATION TABLE

Variables	PREDICTED GROUP 1	2	3	4	TOTAL
1	**26**	31	5	5	67
2	21	**40**	5	7	73
3	8	11	**5**	16	40
4	0	3	5	**12**	20
TOTAL	55	85	20	40	200

Según la tabla anterior, sólo 26 de los 67 individuos del Grupo 1 son clasificados correctamente. En total, la función discriminante clasifica correctamente 83 casos (26+40+5+12), lo que implica una capacidad clasificatoria del 42% (83/200).

Como hemos dicho anteriormente, para poder comparar la importancia relativa de cada variable a la hora de definir los factores es necesario estandarizar los coeficientes anteriores. OS4 proporciona estos coeficientes estandarizados:

Standardized Coeff. from Pooled Cov. with 200 valid cases.

Variables	1	2
DAP	0.996	-0.129
EDAD	0.032	1.004

Según estos resultados, la variable DAP es la que define casi totalmente el primer eje (V_1), mientras en el segundo eje (V_2) es la variable EDAD la que tiene un mayor peso.

Por último tenemos la correlación entre las variables independientes y las dos funciones canónicas. Según estos coeficientes, la primera función tiene una correlación total con la variable DAP (r=1,000) y en menor medida con la variable EDAD (r=0,149). La segunda función está correlacionada casi exclusivamente con EDAD (r=0,989).

Corr.s Between Variables and Functions with 200 valid cases.

Variables	1	2
DAP	1.000	-0.027
EDAD	0.149	0.989

Wilk's Lambda =**0.6603.**
F =14.9892 with D.F. 6 and 390. **Prob > F =0.0000**
Bartlett Chi-Square =81.3419 with 4 D.F. and prob. =0.0000
Pillai Trace =0.3544

El estadístico Wilk's Lambda tiene un valor 0 cuando la función discriminante clasifica correctamente todos los casos. Por el contrario, si el porcentaje de acierto es cero el estadístico alcanza un valor igual a 1. En este ejemplo su valor (0,6603) indica un ajuste no muy preciso. También aparece el valor equivalente F del estadístico de Wilk junto con la probabilidad de la hipótesis nula de que todos los coeficientes son cero. En este caso el valor de F (14,99) y su probabilidad (`Prob > F=0,0000`) implican el rechazo de la hipótesis nula, por lo que aceptamos que el modelo propuesto sí es correcto, esto es, las variables seleccionadas sí tienen capacidad de clasificación, si bien sólo lo hacen correctamente en el 42% de los casos.

Respecto a los supuestos del análisis, el estadístico de Bartlett, con una probabilidad igual a 0,000, implica el rechazo de la hipótesis nula de igualdad de las varianzas entre grupos (supuesto del análisis multivariante de las varianzas). Ante grupos de un tamaño similar, situación que no se da en nuestro ejemplo, el incumplimiento de este requisito no presenta tantos problemas. En cualquier caso, la utilidad del análisis discriminante se medirá por su capacidad de clasificación correcta de los individuos.

Predecir es extremadamente difícil,
especialmente si es acerca del futuro
(Niels Bohr)

CAPÍTULO 8

REGRESIÓN LINEAL MÚLTIPLE

Capítulo 8. Regresión lineal múltiple

8.1. Formulación lineal del problema

En la regresión lineal múltiple intentamos explicar los valores observados de una variable métrica (variable dependiente) en función de un conjunto de variables explicativas de cualquier tipo (nominales, ordinales y/o métricas). La popularidad de esta técnica se explica por su versatilidad, facilidad de uso y gran respaldo teórico. En este sentido, podemos sustituir un análisis de la varianza o de correlación fácilmente por una regresión lineal simple (una sola variable explicativa) o múltiple (varias).

El modelo que se propone, como su propio nombre indica, es lineal. Sin embargo, podemos utilizar esta técnica en situaciones no lineales recurriendo a la transformación de los datos iniciales. Por ejemplo, si el análisis visual de los datos nos inclina a pensar que la relación entre la variable dependiente y la independiente se ajusta a una función del tipo: $Y = b_0 X^{b1}$; podemos proceder a modelizar el logaritmo de ambas eliminando la relación curva. Así, en lugar de estudiar la relación entre Y y X, aplicamos el análisis de regresión a su logaritmo: $\ln(Y) = \ln(b_0) + b_1 \cdot \ln(X)$. Una vez obtenidos los parámetros b_0 y b_1 por esta técnica estadística podemos deshacer la transformación y volver así a la formulación inicial.

Por tanto, la forma general del modelo lineal múltiple estimado es:

$$\hat{Y} = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + \ldots + b_n \cdot X_n$$

Donde $\hat{Y}$ es el valor estimado de la variable dependiente para los valores observados de las variables explicativas $X_1 \ldots X_n$. La diferencia entre el valor observado del caso *i* de la variable dependiente (Y_i) y el estimado ($\hat{Y}_i$) se denomina residuo (u_i). Por tanto:

$$Y_i = \hat{Y}_i + u_i$$

Cuanto menores son los valores absolutos de los residuos mejor es el ajuste de regresión. La técnica más habitual de cálculo de los coeficientes del modelo ($b_0 \ldots b_n$) es el método de mínimos cuadrados ordinarios, el cual minimiza la suma de los cuadrados de los residuos.

8.2. Supuestos del modelo de regresión lineal

Es importante no pasar por alto los supuestos en los que se basa el modelo de regresión lineal, ya que, de hacerlo, podemos llegar a conclusiones completamente erróneas, tanto en la especificación del modelo, es decir, el número de variables del modelo, como en la dimensión de los coeficientes. Los aspectos a tener en cuenta a la hora de validar un modelo de regresión son[21]:

1. Tipo de variables. La variable dependiente deber ser de naturaleza métrica y no limitada (existen casos para todos los posibles valores de la variable dependiente).

2. Variabilidad de las variables independientes. La varianza de las variables independientes no es cero. Por tanto, no tiene ningún sentido incluir como variable explicatoria una cuyo valor es el mismo para todos los casos.
3. Tamaño de la muestra. El número de observaciones es superior al número de variables explicativas. En general, se recomienda al menos un número de casos no inferior a 15 por cada variable explicativa que se incluya en el modelo.
4. No existencia de multicolinealidad. Las variables explicativas no tienen un grado de correlación significativo.
5. Supuestos relativos a la distribución de los residuos (u_i):
   - La media de los residuos es cero. Este supuesto implica que el valor esperado de y, E(y), para un conjunto de valores del vector X es $E(y)=b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 +... + b_n \cdot X_n$.
   - La varianza de los residuos es constante (residuos homocedásticos). Para cada nivel de las variables explicativas la varianza de los residuos es constante, esto es, los residuos son homocedásticos. Si se incumple este requisito nos encontramos ante un problema de heterocedasticidad.
   - Independencia de los errores. Este supuesto implica la no existencia de autocorrelación entre los residuos.
   - Distribución de los residuos. Se asume que dicha distribución es normal.

8.3. Efectos de la violación de los supuestos

Según el objetivo de la regresión (explicatorio o predictivo), el incumplimiento de alguno de estos supuestos influye de forma diferente en la validez de la misma. En general, para el análisis de encuestas los puntos más conflictivos y sus efectos son:

Especificación del modelo de regresión

El supuesto de linealidad del modelo de regresión implica que ninguna otra forma funcional es relevante para explicar el valor de la variable dependiente en función de las variables independientes. Aunque este supuesto es bastante restrictivo, se justifica por su conveniencia a la hora de estimar los parámetros del modelo. Por este motivo es necesario comprobar si la adición de términos no lineales en el modelo puede o no justificarse estadísticamente.

Obviamente, si tratamos de explicar una relación intrínsecamente no lineal mediante un modelo de regresión lineal estaremos cometiendo un error de especificación del modelo. Como consecuencia de esto la capacidad explicativa de nuestro modelo se verá seriamente comprometida así como su validez desde el punto de vista del resto de supuestos del modelo de regresión lineal clásico.

Multicolinealidad

A medida que se incrementa la correlación entre las variables explicativas también lo hacen los errores estándar de los coeficientes, lo cual a su vez incrementa la probabilidad de rechazar erróneamente una variable explicativa estadísticamente válida.

En la práctica, un valor alto de *F* y del coeficiente de determinación (R^2) frente a valores pequeños de *t* sugieren que estamos ante un problema de multicolinealidad y cuya gravedad dependerá del objetivo de la regresión. En este sentido, podemos encontrarnos ante dos situaciones:

- El objetivo de la regresión es explicar la variabilidad de la variable dependiente. Por ejemplo, queremos saber cuáles son las variables determinantes del gasto en ropa deportiva. Si la respuesta es la edad y la educación (los jóvenes estudiantes gastan más) nuestra campaña publicitaria debería orientarse a este segmento de la población. En este caso el problema de multicolinealidad sí debe preocuparnos.
- El objetivo de la regresión es predecir el valor de la variable dependiente. Siguiendo con el ejemplo anterior, sólo nos interesa saber el volumen estimado de ventas de ropa deportiva en una nueva localidad. En este caso, la posible multicolinealidad de las variables explicativas no tiene tanta importancia ya que el objetivo principal es buscar un modelo con la mayor R^2 posible.

Heterocedasticidad

Este problema implica que la varianza de los residuos no es constante a lo largo del valor que toman las variables explicativas del modelo. Más frecuente en modelos longitudinales que en el análisis de series temporales, la presencia de heterocedasticidad[22] tiene como consecuencia la sobreestimación del error estándar de los parámetros del modelo.

En la práctica, si los residuos presentan heterocedasticidad las pruebas *t* y *F* de validez individual de los parámetros y del modelo global, respectivamente, no son aplicables ya que los errores estándar de los estimadores son mayores de lo que deberían, dando lugar a valores menores de *t*, y por tanto, incrementando el riesgo de rechazar un parámetro estadísticamente significativo.

Siendo la presencia de residuos heterocedásticos más bien la norma que la excepción, y teniendo en cuenta la cierta complejidad de su corrección, conviene evaluar la gravedad del incumplimiento de este requisito. Si bien no existe un criterio definido, podemos afirmar que la eficiencia del método de mínimos cuadrados ordinarios se ve seriamente comprometida la ratio s^2_{max}/s^2_{min} (máxima y mínima varianza de los residuos, respectivamente) es mayor que 10 (Fox, 1997, p. 305).

No normalidad de los residuos

Como hemos explicado anteriormente, en el modelo de regresión lineal clásico tres son los requisitos para que los estimadores sean insesgados y eficientes (mínima varianza):

1. $E(u_i)=0 \rightarrow$ la esperanza matemática es igual a cero, esto es, las desviaciones (valor observado – valor estimado) positivas tienden a compensar las negativas.
2. $Cov\ (u_i, u_j)=0 \rightarrow$ no existe correlación entre los residuos.
3. $Var\ (u^2_i)=\sigma^2 \rightarrow$ la varianza de los residuos es constante.

En efecto, si nuestro objetivo es simplemente la estimación de los parámetros del modelo y no queremos inferir los resultados al total de la población basta con que

los residuos cumplan los tres requisitos anteriores. Sin embargo, si el objetivo de la regresión es la inferencia de los resultados al total de la población, caso habitual en el análisis de encuestas, se asume que los residuos siguen una distribución normal. Bajo este supuesto, los estimadores son insesgados y eficientes, y los parámetros tienen mínima varianza.

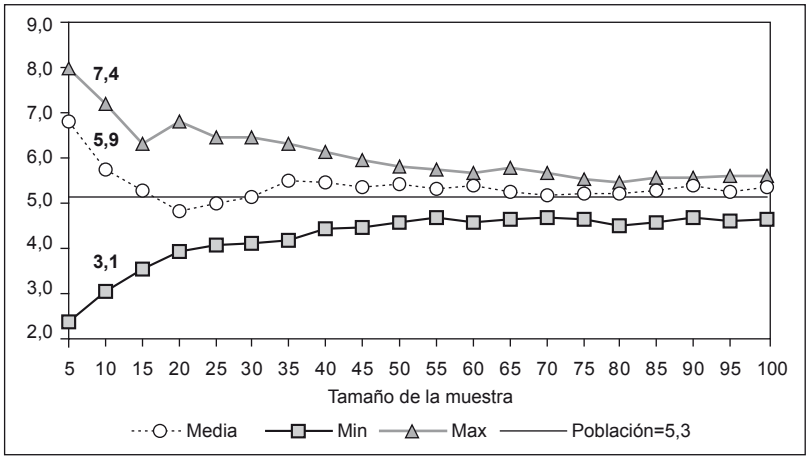
La ventaja de tener un modelo de regresión con los residuos distribuidos normalmente radica en el hecho de poder utilizar las pruebas *t* y *F* independientemente del tamaño de la muestra. Sin embargo, en el caso de no cumplir este requisito, las pruebas *t* y *F* sólo son válidas asintóticamente[23], es decir, para muestras de gran tamaño. A continuación profundizamos en este aspecto.

8.4. Efectos del tamaño de la muestra

El concepto de muestra de tamaño "suficiente" es bastante subjetivo, sin embargo, para el análisis de regresión, la mayoría de autores consideran que una muestra es de gran tamaño cuando supera los 50 casos. Para tratar de visualizar este punto llevaremos a cabo un ejercicio de estimación a partir de varias muestras de diferentes tamaños. Esta simulación tiene las siguientes fases:

1. Generación aleatoria de 2.000 números entre 0 y 10.
2. Extracción aleatoria de 20 de muestras de diferentes tamaños (n=5, 10,... 100). Así, tenemos 20 muestras con 5 casos cada una (grupo 1), 20 muestras con 10 casos cada una (grupo 2), 20 con 20 cada una (grupo 3), 20 con 30 cada una (grupo 4),... 20 muestras con 100 casos cada una (grupo 20).
3. Para cada grupo de muestras selección aleatoria de una y cálculo de su media (círculos en la línea discontinua de la Figura 8.1).
4. En cada grupo de muestras determinación de la media mínima y máxima (líneas con cuadrados y triángulos, respectivamente) de todas las medias del grupo.

Figura 8.1. *Estimación de la media de la población según muestra de diferentes tamaños*



Por ejemplo, fijémonos en las 20 muestras con 10 elementos cada una. Si seleccionamos al azar una de las 20 muestras y calculamos la media de estos 10 elementos obtenemos el valor 5,9. Por tanto concluiríamos que la estimación de la media de la población es 5,9. En realidad, si hubiéramos podido acceder a los 2.000 casos, en lugar de sólo 10, determinaríamos con exactitud que la media de la población es 5,3 (línea horizontal continua). En este caso, sólo con una muestra de 10 elementos hemos obtenido una estimación bastante buena, sin embargo, también podríamos haber seleccionado por azar cualquier otra de las 19 muestras restantes de 10 elementos cada una. Una de estas muestras da una media de 7,4 (media máxima del grupo) y otra una media de 3,1 (media mínima del grupo). Es decir, con sólo 10 casos corremos un gran riesgo de obtener una estimación muy alejada de la realidad.

Por el contrario, con muestras a partir de 50 casos la diferencia entre la media más baja y la media más alta de las 20 muestras prácticamente se estabiliza. Esta es la justificación de fijar el límite de una muestra de gran tamaño a partir de este valor.

8.5. Detección y corrección de violación de supuestos

Considerando los efectos que la violación de un supuesto tiene en la validez del modelo de regresión lineal clásico es necesario disponer de pruebas para su detección así como de medidas de corrección. Es importante no olvidar que todas estas técnicas no sirven de nada en el caso de no disponer de las variables adecuadas para el análisis. En este sentido, todo modelo de regresión debe buscar relaciones donde a *priori* existan y no simplemente obtener el modelo con el mayor coeficiente de determinación (R^2). El mensaje es claro: Es preferible un modelo con una menor R^2 pero con un mayor sentido socio-económico (en el caso del análisis de encuestas). A continuación resumimos las técnicas de detección y las medidas de corrección para los casos de multicolinealidad, heterocedasticidad y no normalidad de los residuos.

Especificación del modelo de regresión

En primer lugar resulta útil visualizar la relación entre la variable dependiente y cada una de las variables explicativas, sin embargo, en la mayoría de las ocasiones, no resulta fácil descubrir relaciones no lineales a partir de la nube de puntos. Para solucionar este problema podemos recurrir a la prueba de Ramsey que tiene por objetivo descartar una relación no lineal entre las variables explicativas y la variable dependiente. Para realizar esta prueba utilizamos la siguiente regresión auxiliar:

$$Y = B \cdot X + c_1 \cdot \hat{Y}^2 + c_2 \cdot \hat{Y}^3 + \ldots + c_n \cdot \hat{Y}^n$$

Donde Y y X son las variables de la regresión inicial $Y = B \cdot X$ y las variables $\hat{Y}^n$ representan la potencia de grado n de los valores estimados de la regresión inicial. Si todos los coeficientes c_n no son estadísticamente significativos (prob>0,05) podemos descartar una relación no lineal en la regresión inicial, aceptando por tanto el modelo lineal propuesto. En la práctica se suele estudiar la significación de los términos de segundo y tercer grado.

En el caso de rechazar la hipótesis de un modelo lineal sería preciso corregir este problema detectando la relación (o relaciones) no lineales entre la variable dependiente

y la explicativa (o explicativas) mediante transformaciones logarítmicas o potenciales de los datos iniciales.

Multicolinealidad

Para la detección de un posible problema de multicolinealidad en nuestro modelo podemos llevar a cabo diversas pruebas que resumimos a continuación:

1. Correlaciones de las variables explicativas. Si el coeficiente de correlación entre dos variables (r_{ij}) es alto, digamos superior a 0,30, es probable que tengamos este problema. Si este coeficiente excede de 0,80 estamos ante un problema grave de multicolinealidad. Sin embargo, no podemos asegurar que no existe un problema de multicolinealidad incluso cuando todos los coeficientes de correlación son bajos (r_{ij}<0,30), ya que la baja correlación entre variables es condición necesaria pero no suficiente[24].

2. Valores de los estadísticos *t* y *F*. A simple vista también es posible detectar, en algunas ocasiones, un problema de multicolinealidad en nuestro modelo. Los indicios son: algunos de los coeficientes no son significativos (Prob *t* >0,05) pero no se rechaza el modelo global (Prob *F* < 0,05) y a su vez tenemos un coeficiente de determinación muy alto ($R^2 > 0,90$)[25].

3. Regla de Klien. Si el coeficiente de determinación de alguna de las regresiones de la variable dependiente sobre cada una de las variables independientes ($Y=b_0+b_iX_i$), es decir, regresiones con una sola variable explicativa, es superior al correspondiente coeficiente de determinación (R^2) del modelo global ($Y=b_0+b_1X_1+...+b_nX_n$) tenemos un problema de multicolinealidad (Klien, 1962, p. 101).

4. Factor de inflación de la varianza (FIV) y Tolerancia. Si el factor de inflación de la varianza -*variance inflation factor* (VIF) en inglés- excede en alguna variable de 10 concluimos que esa variable está altamente correlacionada (Myers, 1990; Bowerman y O'Connell, 1990). Respecto a la tolerancia (inversa de FIV), según Menard (2002), un valor inferior a 0,20 indica un problema potencial de multicolinealidad. OS4 proporciona automáticamente ambos indicadores.

Una vez detectado un problema de multicolinealidad por alguna, preferiblemente varias, de las técnicas anteriores es necesario introducir una de las medidas de corrección que se enumeran a continuación:

1. Eliminación de variables. Simplemente eliminamos aquella variable que está altamente correlacionada con otra u otras, ya que también puede ser una combinación lineal de varias variables explicativas. La elección estadísticamente más eficiente, que puede no coincidir con nuestro interés investigador, consistiría en eliminar la variable con mayor FIV (o menor tolerancia).

2. Transformación de las variables. Podemos transformar una variable explicativa que dé lugar a otra no correlacionada con el resto. Por ejemplo, podríamos categorizar una variable métrica (transformando una variable métrica en ordinal), fusionar categorías (reduciendo el número de categorías) o utilizar el logaritmo neperiano de la variable original.

3. Fusión de variables. Mediante técnicas estadísticas como el análisis factorial y el análisis de componentes principales podemos fusionar dos o más variables en una sola.

Heterocedasticidad

Existen numerosas pruebas para la detección de residuos heterocedásticos, esto es, de desigual varianza según el nivel de las variables independiente. Las tres que a continuación describimos, Park, Breusch-Pagan-Godfrey y White, realizan una regresión auxiliar de los residuos sobre las variables independientes.

1. Prueba de Park. Esta prueba (Park, 1966; Harvey, 1976) de tipo exploratorio, consta de dos etapas:

   a. Obtenemos los residuos de nuestro modelo: $Y_i=b_0 + b_1X_1 +... + b_nX_n + u_i$.

   b. Estimamos la siguiente regresión: $u_i^2=b_0+b_iX_i+v_i$, siendo u_i los residuos de la primera regresión y X_i cada una de las variables independientes. Si alguno de los parámetros b_i resultase significativo en alguna de las regresiones anteriores concluiríamos que la varianza de los residuos depende del nivel de la variable, por lo que se incumpliría el supuesto de homocedasticidad.

Sin embargo, como limitación de esta prueba, es necesario apuntar que los residuos de las regresiones auxiliares, v_i, pueden no ser ellos mismos homocedásticos, por lo que queda comprometida la prueba *t* de significatividad de los parámetros.

2. Prueba de Breusch-Pagan-Godfrey. De las tres, esta prueba (Godfrey, 1978; Breusch y Pagan, 1979) es la que requiere un mayor tamaño muestral, por lo que no debería utilizarse para muestras inferiores a 50 casos. Los pasos a seguir son:

   a. Se obtienen los residuos de la regresión inicial, u_i.
   b. Calculamos $\sigma^2=\sum u_i^2/n$, es decir, un estimador de la varianza de los residuos, siendo n el tamaño de la muestra.
   c. Obtener una nueva variable $p_i=u_i^2/\sigma^2$.
   d. Realizar la siguiente regresión: $p_i=b_0 + b_1X_1 +... + b_nX_n + v_i$.
   e. Utilizando la suma de los cuadrados de la regresión (SC_{reg})[26] calcular el siguiente estadístico: $\theta=½\ SC_{reg}$.
   f. Bajo el supuesto de residuos homocedásticos en la regresión inicial el coeficiente θ se distribuye como una Chi-cuadrado con *k* grados de libertad (*k*=número de variables explicativas de la regresión excluyendo el término constante). Por tanto, si el valor de θ excede el valor crítico de la Chi-cuadrado rechazaríamos la hipótesis nula de homocedasticidad.

3. Prueba de White. A diferencia de las dos pruebas anteriores, la de White no requiere la normalidad de los residuos (White, 1980).

   a. Al igual que en las dos pruebas anteriores, obtenemos los residuos del modelo inicial.
   b. Calculamos el coeficiente de determinación de la regresión de los residuos al cuadrado sobre el modelo original más todas las combinaciones multiplicativas de las mismas y sus cuadrados[27].

c. El producto $n \cdot R^2_{auxiliar}$ se distribuye como una Chi-cuadrado con *k* grados de libertad (*n*=tamaño muestral y *k*=número de variables en la regresión auxiliar excluyendo el término constante) bajo la hipótesis nula de homocedasticidad, por tanto si el valor de $n \cdot R^2_{auxiliar}$ excede el valor crítico rechazamos que los residuos sean homocedásticos.

Una vez determinada la gravedad del problema de heterocedasticidad, resulta conveniente representar gráficamente la relación entre el cuadrado de los residuos y cada una de las variables independientes de la regresión. Esta representación es importante ya que, de no encontrar una relación clara entre el cuadrado de los residuos y las variables independientes, podríamos aceptar la utilización del método MCO, incluso bajo el supuesto de heterocedasticidad, en el caso de disponer de una muestra de gran tamaño (por ejemplo, superior a 50 casos). Por tanto, si bien no es el método más eficiente, al menos no nos llevará a conclusiones erróneas (Greene, 1997, p. 547).

En cualquier caso, si detectamos que los residuos no son homocedásticos, la forma correcta de actuar pasaría por estabilizar la varianza de los mismos. Para ello podemos recurrir a la estimación por mínimos cuadrados ponderados (en inglés, *weighted least squares* - WLS-[28], también conocido como *generalized least squared* –GLS-). Mientras en el método de estimación que hemos utilizado hasta ahora (mínimos cuadrados ordinarios *–ordinary least squared* –OLS-) se minimiza la suma de los cuadrados de los residuos, en el método de mínimos cuadrados ponderados se minimiza la suma ponderada de esos mismos cuadrados. Es decir, no todos los residuos tienen el mismo peso en el sumatorio que vamos a minimizar. Matemáticamente:

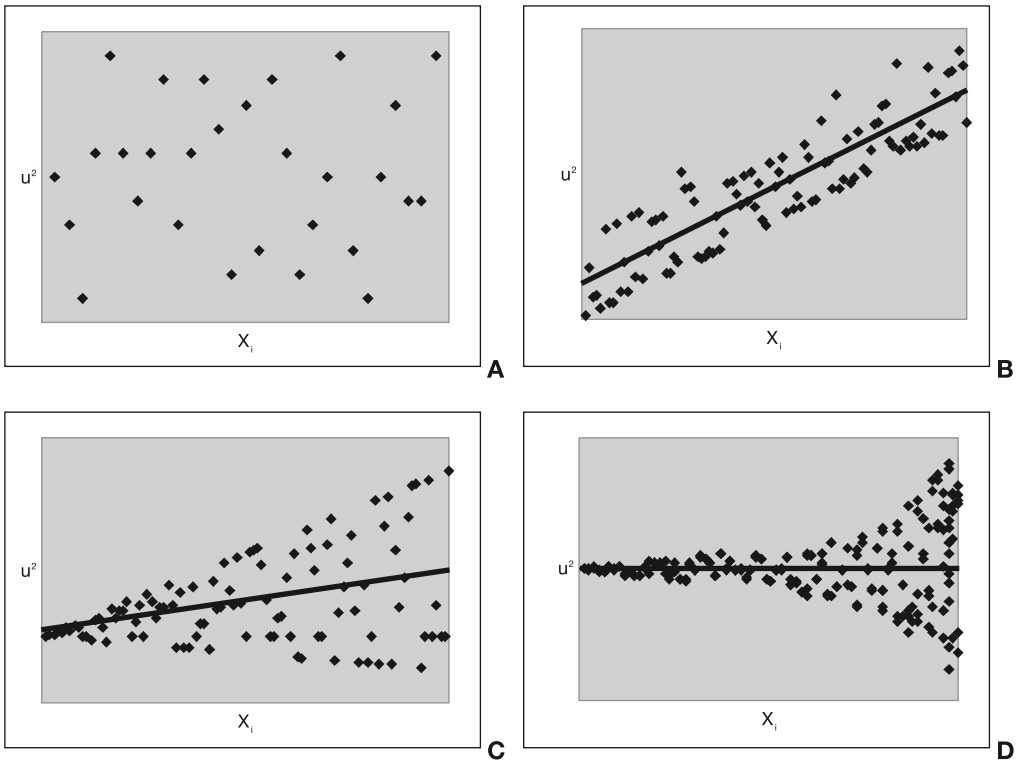
Mínimos Cuadrados Ordinarios: $\text{Min SSE}=\sum_{i=1}^{n} u_i^2$

Mínimos Cuadrados Ponderados: $\text{Min WSSE}=\sum_{i=1}^{n} w_i u_i^2$

El método de mínimos cuadrados ponderados método permite tener en cuenta la varianza heterocedástica y obtener valores fiables del estadístico t, si bien es necesario conocer la varianza poblacional de la variable dependiente para los diferentes niveles de las variables explicativas o al menos la relación funcional entre los residuos heterocedásticos y alguna de las variables explicativas.

Como hemos apuntado anteriormente, en el caso de encontrar alguna relación entre los residuos (su cuadrado) y alguna de las variables independientes, utilizaremos esta información para corregir el problema de heterocedasticidad de los residuos. A continuación mostramos algunos ejemplos de posibles relaciones entre los residuos y las variables independientes.

Figura 8.2. *Posible relación entre los residuos y las variables explicativas*



En la Figura 8.2.A no existe ninguna relación entre la varianza de los residuos y la variable independiente X_i. Bajo este supuesto se puede admitir la utilización de los resultados del modelo de regresión inicial, incluso habiendo confirmado el problema mediante las pruebas de Park, Breusch-Pagan-Godfrey o White, si disponemos de una muestra de un tamaño suficiente[29]. Por el contrario, en el caso de encontrar una relación entre u^2 y X_i, como aparece en las Figuras 8.2.B, 8.2.C y 8.2.D, es necesaria la transformación de las variables para corregir este problema. Así, el tipo de transformación dependerá de la relación encontrada. De entre varias alternativas disponibles, dos posibles opciones son:

- Relación proporcional entre u^2 y X_i. Para solucionar un problema del tipo representado en las figuras 8.2.B y 8.2.C podemos dividir todas las variables por X_i. Así, en lugar de estimar el modelo inicial con residuos heterocedásticos $Y=b_0+b_1X_1+...+b_iX_i+...+b_nX_n$, estimaremos $Y/X_i=c_0+c_1X_1/X_i+...+c_iX_i/X_i+...+c_nX_n/X_i$, que no presenta este problema. Una vez obtenidos los parámetros del segundo modelo (c_i) podemos deshacer la transformación y realizar predicciones de Y según los valores de las variables independientes.
- Relación cuadrática entre u^2 y X_i. En el caso de encontrarnos con una relación del tipo 8.2.D, podemos dividir cada variable por la raíz cuadrada de X_i.

No normalidad de los residuos

Si bien se ha sugerido anteriormente que podríamos obviar el hecho de que los residuos no se distribuyan como una normal para muestras superiores a 50 casos[30], esto no es óbice para que admitamos que, incluso amparados por el teorema central del límite y el comportamiento asintótico de los estimadores, estamos cometiendo un error difícilmente cuantificable. Por tanto, según algunos autores (Shanken, 1996; Campbell, 1997; Dufour *et al.,* 1998; Dufour y Khalaf, 2002), no es justificable este proceder incluso para muestras de tamaño muy superior al mencionado anteriormente.

En consecuencia, es aconsejable abordar el problema de la no normalidad de los residuos incluso ante muestras de gran tamaño. El primer paso consiste en su detección. Con OS4 podemos computar:

Analyses → Descriptive → Normality Tests: *ui* →Apply.,

siendo ui los residuos de la regresión inicial. No es necesario calcular estos residuos (diferencia entre los valores de Y observados y los estimados) ya que son proporcionados por OS4 (Analyses → Regression; ⊙ Predictions,residuals,C.I. to grid) en el panel de datos principales como una nueva variable (*Raw Resid*.).

Si los residuos siguen una distribución normal su media es cero y su varianza es constante. En OS4 podemos utilizar la prueba de normalidad Analyses → Descriptive → Normality Tests: Test Normality of: *Raw Resid* → Apply para corroborar o rechazar esta hipótesis.

En el caso de rechazar la normalidad de los residuos podemos transformar los datos iniciales para corregir este problema. El diagrama de dispersión de los residuos y los valores estimados por un lado, y el de los residuos y cada una de las variables explicativas, por otro, puede ayudarnos a definir el tipo de transformación a realizar. Así, una relación no aleatoria en el primer caso podría sugerir un error de especificación del modelo, esto es, un modelo alternativo no lineal podría ser una mejor representación de los datos. Si se descubre alguna relación entre los residuos y alguna de las variables explicativas podemos intentar corregirla mediante la aplicación del logaritmo o la raíz cuadrada a la variable explicativa correspondiente.

8.6. Interpretación de los resultados de la regresión

Pruebas *t* y *F*

Si el modelo cumple con los supuestos de no multicolinealidad, homocedasticidad y residuos con distribución normal podemos utilizar las pruebas *F* y *t* de validez del modelo global y de los coeficientes individuales de la regresión, respectivamente.

- La prueba *F* está asociada al análisis de la varianza. A partir de la variabilidad explicada por la regresión y la variabilidad residual calculamos el estadístico *F*, cuya hipótesis nula es el rechazo global del modelo. Si la probabilidad de *F* es inferior a 0,05 rechazamos la hipótesis nula y aceptamos que nuestro modelo es válido.

- La prueba *t* tiene como hipótesis nula que el coeficiente de la variable es igual a cero (H_0: b_j=0), y por tanto que la variable explicativa no tiene relación significativa con la variable dependiente. Por consiguiente, aquellos coeficientes con una probabilidad del estadístico *t* inferior a 0,05 son aceptados en el modelo, ya que en el caso de los coeficientes con probabilidad mayor que 0,05 no rechazamos la hipótesis nula de que su verdadero valor es cero y el valor obtenido se debe al azar. En OS4 y otros paquetes estadísticos los coeficientes del modelo aparecen en la columna B.

Podemos determinar la contribución de cada una de las variables independientes en la explicación de la variabilidad de la variable dependiente mediante los coeficientes tipificados (denominados coeficientes tipificados o estandarizados). Cuanto mayor sea este coeficiente mayor será la importancia relativa de la variable independiente correspondiente. En OS4 y otros paquetes estadísticos dichos coeficientes aparecen en la columna Beta.

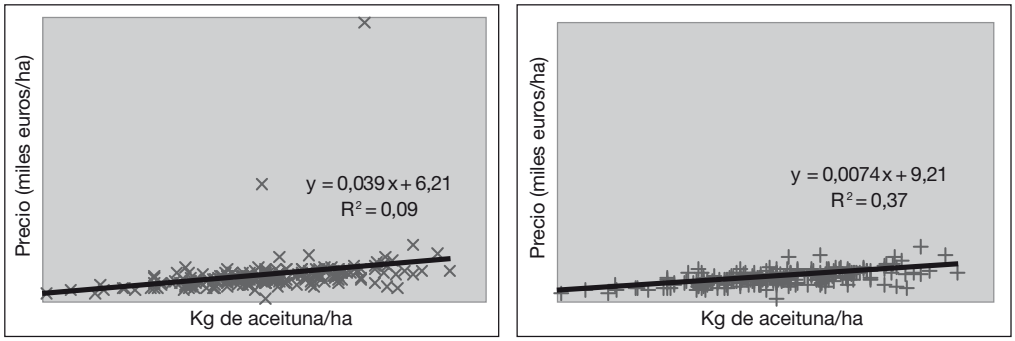
Criterio del coeficiente de determinación

Para determinar la capacidad explicativa de un modelo de regresión utilizamos el coeficiente de determinación (R^2) que representa el porcentaje de variabilidad de la variable dependiente que el modelo explica, por tanto siempre es positivo y con un valor entre 0 y 1, excepto en modelos sin constante, donde R^2 puede tomar valores negativos que se interpretan como cero. El coeficiente de determinación presenta el inconveniente de incrementarse progresivamente con la inclusión de variables explicativas adicionales, incluso cuando éstas no tengan capacidad explicativa. En este sentido conviene advertir que el objetivo del modelo de regresión no es maximizar el coeficiente de determinación sino conseguir un modelo consistente con la teoría que tenga un coeficiente de determinación aceptable[31].

Una vez descartado el juego de maximizar dicho coeficiente, en el caso de no conseguir un valor aceptable podemos analizar qué datos están distorsionando nuestro análisis. Un simple diagrama de dispersión con la variable independiente más importante (ver valores Beta de modelo) en el eje X y la independiente en el eje Y puede ayudarnos a descubrir valores extremos. También los valores estandarizados de los residuos (`z Resid`) mayores que 2,5 o menores que –2,5 pueden servir para este propósito. Una vez detectados la pregunta es ¿qué hacer con ellos? no pueden eliminarse porque no “encajen” en nuestro modelo ya que entonces nuestro modelo no sería una representación válida de la población objetivo, sin embargo, podemos descubrir qué características los hacen diferentes del resto e incluso separarlos del resto y analizarlos como provenientes de una población diferente. Un ejemplo real aclara este punto:

En un estudio sobre el valor de mercado fincas rústicas obtenemos el siguiente diagrama de dispersión entre el valor de mercado por hectárea y el rendimiento por hectárea:

Figura 8.3. *Regresión con y sin valores extremos*



En el primer caso con todas las fincas (n=224 casos) podemos ver cómo el modelo planteado Precio=b_0 + b_1·rendimiento tiene un coeficiente de determinación muy bajo (9%). La razón de este pobre ajuste hay que atribuírselo a dos puntos que claramente se distinguen del resto. En efecto, estas fincas tienen un precio por hectárea muy superior al resto. Revisando los datos referentes a las mismas descubrimos que son las únicas dos fincas que incluyen edificaciones donde vive el agricultor cuyo valor es muy superior a las edificaciones de uso industrial del resto. En este caso no estamos comparando lo mismo por lo que habría que eliminar estas dos fincas o deducirles el valor de dichas edificaciones. En el segundo gráfico vemos el mismo ajuste pero sin estos valores extremos, en este caso el coeficiente de determinación se eleva hasta un 37%.

Teniendo en cuenta que el coeficiente de determinación se incrementa con la mera inclusión de más variables explicativas, incluso en contra del marco teórico establecido, muchos autores recomiendan el uso en su lugar del coeficiente de determinación ajustado ($\overline{R^2}$), que sí tiene en cuenta el número de variables explicativas y de casos (Verbeek, 2000, p. 53; Gujarati, 1995, p. 208) y no incrementa su valor por la mera inclusión de nuevas variables.

Varianza de los residuos

El valor MS (siglas anglosajonas que se corresponden con *Mean Square*) nos informa de la varianza de la regresión y la de los residuos, cuyo cociente nos proporciona el valor del estadístico F. El valor de la varianza de los residuos (en algunos paquetes estadísticos se denomina MSE) es útil ya que el doble de su raíz cuadrada nos sirve como un indicador de la precisión de la estimación del valor de la variable dependiente. Por tanto, la predicción que hacemos, $\hat{y}$, a partir de los valores que toman las variables explicativas se encuentra con un 95% de probabilidad en el intervalo ($\hat{y}$ -2s, $\hat{y}$ +2s), siendo s la desviación típica de los residuos.

Es posible tener un modelo con un coeficiente de determinación alto (por ejemplo superior a 0,90) pero con una varianza de los residuos (s^2) relativamente alta por lo que, si el objetivo de nuestro modelo es predecir un valor de la variable dependiente a partir de los valores que toman las variables explicativas, es modelo obtenido será poco útil por presentar un intervalo de confianza excesivamente amplio.

Criterios alternativos de comparación

Existen alternativas al coeficiente de determinación (ajustado o no) que pueden utilizarse para la comparación de dos modelos alternativos. En el caso de comparar dos modelos con diferente número de variables explicativas podemos utilizar el Criterio de Akaike –AIC- (Akaike, 1973) o el Criterio bayesiano de Schwarz –BIC- (Schwarz, 1978). El modelo que presente menores valores AIC y BIC es generalmente preferido. Su cálculo es como sigue:

$$\text{AIC}= \ln\frac{1}{N}\sum_{i=1}^{N}u_i^2+\frac{2K}{N}$$

$$\text{BIC}= \ln\frac{1}{N}\sum_{i=1}^{N}u_i^2+\frac{K}{N}\ln N$$

donde *N* es el tamaño de la muestra, *K* es el número de parámetros del modelo (incluyendo la constante) y u_i los residuos de la regresión.

Otro criterio alternativo es el de Davidson-MacKinnon (Davidson y MacKinnon, 1981), conocido como la prueba J. Imaginemos que tenemos dos modelos alternativos:

Modelo A: $Y=a_0 + a_1\cdot X + u$

Modelo B: $Y=b_0 + b_1\cdot Z + v$

Es decir, explicamos la variable Y en función de la variable X en el primer caso, y en función de Z en el segundo, más una parte no explicada representada por los residuos *u* y *v*. Según esta prueba, obtenemos los valores estimados del primer modelo, $\hat{Y}^A$ y procedemos a incluirlos como variable explicativa en el segundo, es decir:

Modelo B': $Y=b_0 + b_1\cdot Z + b_2\cdot\hat{Y}^A + v$

Si el coeficiente b_2, utilizando la prueba *t*, no resulta significativo (la probabilidad de *t* es superior a 0,05) aceptamos el modelo B. Por el contrario, si b_2 resulta significativo aceptaríamos el modelo A.

Repitiendo la operación pero cambiando los modelos tenemos:

Modelo A': $Y=a_0 + a_1\cdot X + a_2\cdot\hat{Y}^B + u$

donde $\hat{Y}^B$ son los valores estimados del modelo B. Según la significación o no de a^2 llegamos a las mismas conclusiones que anteriormente. Sin embargo, ya que las pruebas se realizan de forma independiente, podemos llegar a conclusiones que no sean concluyentes, como muestra la siguiente tabla:

Tabla 8.1. *Posibles resultados de la prueba Davidson-MacKinnon de comparación de modelos*

	a_2 no es significativo ($a_2=0$)	a_2 es significativo ($a_2\neq0$)
b_2 no es significativo ($b_2=0$)	Aceptamos A y B	Aceptamos B
b_2 es significativo ($b_2\neq0$)	Aceptamos A	Rechazamos A y B

Como muestra la tabla anterior, existe la posibilidad de no llegar a ninguna conclusión utilizando esta prueba. Otro inconveniente es que sólo puede utilizarse con muestras superiores a 50 casos para que el comportamiento asintótico de los estimadores asegure la validez de la prueba *t*.

8.7. Ejemplo de regresión múltiple

Especificación general del modelo

Como ejemplo de regresión múltiple, esto es, un modelo con más de una variable independiente, investigaremos la posible influencia de todas las variables explicativas disponibles en nuestra encuesta sobre la disposición a pagar una entrada por visitar el parque (DAP). Planteamos por tanto el siguiente modelo tentativo:

$DAP = b_0 + b_1 \cdot CAZORLA + b_2 \cdot DISTANCIA + b_3 \cdot PRIMERA + b_4 \cdot N\_VECES + b_5 \cdot ESTUDIOS + b_6 \cdot EDAD + b_7 \cdot INGRESOS + b_8 \cdot SEXO$

Resulta útil comenzar por los coeficientes de correlación entre la variable dependiente (DAP) y cada una de las variables independientes métricas (DISTANCIA, N_VECES) u ordinales (ESTUDIOS, EDAD, INGRESOS).

Analyses → Correlation → Product-Moment Correlations *Dap, Distancia, N_Veces, Estudios, Edad, Ingresos* ⊙ Show Intercorrelations Matrix.

Product-Moment Correlations Matrix with 200 cases.

Variables

	DAP	DISTANCIA	N_VECES	ESTUDIOS	EDAD	INGRESOS
DAP	1.000	0.765	-0.107	-0.020	0.122	0.507
DISTANCIA	0.765	1.000	-0.164	-0.003	0.178	0.265
N_VECES	-0.107	-0.164	1.000	0.053	-0.075	-0.128
ESTUDIOS	-0.020	-0.003	0.053	1.000	-0.005	0.072
EDAD	0.122	0.178	-0.075	-0.005	1.000	-0.004
INGRESOS	0.507	0.265	-0.128	0.072	-0.004	1.000

Según los resultados anteriores el modelo de regresión debería incluir las variables explicativas DISTANCIA e INGRESOS, ya que el coeficiente de correlación (r) es suficientemente alto ($|r| > 0{,}30$). También podría estudiarse la inclusión de las variables N_VECES, EDAD con un grado de correlación moderado ($0{,}10 < |r| < 0{,}30$). Hemos subrayado una correlación alta entre dos de las variables explicativas (DISTANCIA e INGRESOS) que podrían suponer un problema de multicolinealidad en el modelo de regresión múltiple que a continuación vamos a analizar:

Analyses → Regression → Block Entry: Dependent Variable: *Dap*; Independent Var's to enter in block: *Distancia, Primera,·N_Veces, Edad, Ingresos, Cazorla, Sexo* ⊙ Print Means and Std. Dev. s.

```
Dependent variable: DAP

Variable     Beta    B         Std.Err.   t        Prob.>t   VIF     TOL
DISTANCIA    0.656   0.009     0.001      10.019   0.000     2.654   0.377
PRIMERA      0.047   0.153     0.164      0.934    0.352     1.575   0.635
N_VECES      0.049   0.003     0.003      1.176    0.241     1.054   0.948
EDAD         0.009   0.009     0.042      0.206    0.837     1.054   0.948
INGRESOS     0.321   0.459     0.063      7.319    0.000     1.193   0.838
CAZORLA      0.007   0.032     0.248      0.128    0.898     2.108   0.474
SEXO         -0.038  -0.108    0.115      -0.935   0.351     1.025   0.975
Intercept    0.000   -0.316    0.263      -1.200   0.231

SOURCE       DF      SS        MS         F        Prob.>F
Regression   7       262.126   37.447     60.933   0.0000
Residual     192     117.994   0.615
Total        199     380.120

R2 =0.6896, F = 60.93, D.F. =7 192, Prob>F =0.0000
Adjusted R2 =0.6783

Standard Error of Estimate = 0.78
F =60.933 with probability = 0.000
Block 1 met entry requirements
```

De acuerdo con los resultados anteriores, sólo dos variables (DISTANCIA e INGRESOS) tienen coeficientes con probabilidad inferior a 0,05 de ser cero (y por tanto, por ser tan improbable, rechazamos que sean cero y aceptamos el valor de los coeficientes estimados en la columna `B`: 0,009 y 0,459, respectivamente). Si nos fijamos en el valor de los coeficientes tipificados (columna `Beta`) comprobamos que la variables más importante en el modelo es DISTANCIA (0,656) seguida por INGRESOS (0,321). Por tanto, teniendo en cuenta la significación de los coeficientes el modelo quedaría:

$DAP = b_0 + b_1 \cdot DISTANCIA + b_2 \cdot INGRESOS$

Para descartar un modelo no linear recurrimos a la prueba Ramsey. Como explicamos anteriormente, esta prueba utiliza una regresión auxiliar que tiene por variable dependiente los valores observados de DAP y por variables explicativas las dos variables inciales (DISTANCIA e INGRESOS) y el cuadrado y el cubo de los valores estimados de DAP en el modelo inicial. Para obtener los valores estimados de DAP incluimos la siguiente opción en el menú de regresión:

Regression → Block Entry: Dependent Variable: *Dap*; Independent Var's to enter in block: *Distancia, Ingresos* ⊙Predictions,residuals,C.I.'s to grid.

De entre las nuevas columnas que aparecen en el panel de datos, "`Pred.Raw`" presenta los valores estimados de DAP para cada par de valores de las variables explicativas

DISTANCIA e INGRESOS. La columna "`Raw Resid`." calcula la diferencia entre el valor estimado por el modelo (`Pred.Raw`) y el valor real de la variable dependiente (`DAP`). A continuación calculamos el cuadrado y el cubo de los valores estimados (`Pred.Raw`) mediante las transformaciones siguientes:

Variables → Transform; First Var. Argument (V1): *Pred.Raw*; Constant: *2*; Save New Variable As: *pred2*; Select Transformation: *New =V1^C* → Compute → Return.

Y de igual forma para obtener el cubo: Variables → Transform; First Var. Argument (V1): *Pred.Raw*; Constant: *3*; Save New Variable As: *pred3*; Select Transformation: *New =V1**C* → Compute → Return.

Una vez tenemos todos los términos necesarios para la regresión auxiliar estimamos el siguiente modelo: $DAP = b_0 + b_1 \cdot DISTANCIA + b_2 \cdot INGRESOS + b_3 \cdot pred2 + b_4 \cdot pred3$.

```
Block Entry Multiple Regression by Bill Miller

-------------- Trial Block 1 Variables Added ---------------
R             R2      F          Prob.>F     DF1     DF2
0.830         0.689   108.105    0.000       4       195
Adjusted R Squared =0.683
Std. Error of Estimate = 0.778

Variable         Beta          B   Std.Error       t   Prob.>t
   DISTANCIA    0.309      0.004       0.004   1.145     0.254
    INGRESOS    0.222      0.318       0.139   2.278     0.024
       pred2    0.684      0.180       0.171   1.054     0.293
       pred3   -0.281     -0.016       0.021  -0.728     0.468

Constant =-0.103
Increase in R Squared = 0.689
F =108.105 with probability = 0.000
Block 1 met entry requirements
```

La probabilidad de los valores de *t* asociados a los coeficientes de *pred2* y *pred3* (0,293 y 0,468, respectivamente) es superior a 0,05, por lo que no rechazamos la hipótesis nula de que dichos coeficientes son cero. Al no ser significativos estos parámetros descartamos una forma funcional no lineal en nuestro modelo, por lo que continuamos con la regresión inicial:

Analyses → Regression → Block Entry: Dependent Variable: *Dap*; Independent Var's to enter in block: *Distancia, Ingresos* -sin ninguna opción-.

```
Dependent variable: DAP

Variable      Beta   B          Std.Err.  t         Prob.>t  VIF    TOL
DISTANCIA     0.678  0.009      0.001     16.341    0.000    1.075  0.930
INGRESOS      0.328  0.468      0.059     7.895     0.000    1.075  0.930
Intercept     0.000  -0.410     0.131     -3.142    0.002

SOURCE           DF        SS        MS        F  Prob.>F
Regression        2   260.236   130.118  213.816   0.0000
Residual        197   119.884     0.609
Total           199   380.120

R2 =0.6846, F = 213.82, D.F. =2 197, Prob>F =0.0000
Adjusted R2 =0.6814

Standard Error of Estimate = 0.78
F =213.816 with probability = 0.000
Block 1 met entry requirements
```

Los resultados indican que el modelo propuesto explica el 68,1% (valor de R^2 ajustado) de la variabilidad de la variable dependiente (DAP). Este valor es relativamente alto en las ciencias sociales ya que es frecuente encontrar estudios socioeconómicos donde se acepta un valor de R^2 incluso inferior a 0,30.

El valor MS nos informa de la varianza de la regresión (130,118) y la varianza de los residuos (0,609). Su cociente proporciona el valor del estadístico F, esto es, F=130,118 / 0,609=213,816. Como explicamos anteriormente, el valor de la varianza de los residuos es útil ya que el doble de su raíz cuadrada nos sirve como un indicador de la precisión de la estimación del valor de la variable dependiente. En nuestro caso, $2^*\sqrt{0,609} = 2^*0,78 = 1,56$, por tanto las estimaciones de la variable dependiente tendrán un margen de error de aproximadamente 1,56 euros, lo cual es un valor relativamente alto que resta utilidad al modelo.

El valor del estadístico *F* es alto (213,8), con una probabilidad de ocurrencia igual a 0,000 (menor que 0,05) por lo que rechazamos la hipótesis nula de no validez global del modelo y aceptamos, por tanto, que el modelo es válido. En general, si `Prob.>F` es mayor que 0,05 rechazamos el modelo en su totalidad ya que no podemos rechazar la hipótesis nula de un valor igual a cero para todos los coeficientes (b_1 y b_2, en este caso).

Según los valores tipificados de los coeficientes (Beta) la variable DISTANCIA tiene una capacidad de explicación de DAP doble que la variable INGRESOS. Teniendo en cuenta que ambas variables tienen cierta correlación (coeficiente de correlación=0,265), el posible problema de multicolinealidad limita la posibilidad de interpretar los valores Beta. Posteriormente llevaremos a cabo las diferentes pruebas de validez

A partir del signo de los coeficientes (B) de las variables podemos determinar su influencia sobre la variable dependiente. Al ser ambos positivos indican una mayor DAP cuanto mayores son las distancias recorridas y el nivel de ingresos. Matemáticamente el modelo quedaría como sigue a continuación:

DAP=-0,410 + 0,009·*DISTANCIA* + 0,468·*INGRESOS*

Por ejemplo, un visitante que haya recorrido 200 km y tenga un nivel de ingresos igual a 2 tendría una disposición a pagar una entrada por visitar el parque estimada en:

DAP=-0,410 + 0,009·200 + 0,468·2 =2,33 euros

Validez del modelo

Para aceptar el modelo anterior es necesario comprobar que se cumplen los supuestos del análisis de regresión. Comenzamos con el estudio de casos con valores extremos. A continuación analizaremos los supuestos de no multicolinealidad de las variables explicativas, la homogeneidad de la varianza y, finalmente, la normalidad de los residuos.

Estudio de casos extremos

La existencia de casos con valores extremos (*outliers* en la terminología anglosajona) puede tener un efecto notable sobre los coeficientes de las variables explicativas de la regresión. Para detectar estos casos con una influencia muy superior a su peso relativo en la muestra podemos utilizar los residuos estandarizados obtenidos en la sección anterior. Así, tras haber ejecutado las instrucciones:

Analyses → Regression → Block Entry: Dependent Variable: *Dap*; Independent Var's to enter in block: *Distancia, Ingresos* ⊙Predictions,residuals,C.I.'s to grid.

Nos fijamos en la columna "`z Resid.`" que proporciona los residuos normalizados, cuyo valor absoluto se encuentra entre –2 y +2 en el 95% de los casos, y entre –2,5 y +2,5 en el 99%. Desde un punto de vista práctico, valores absolutos superiores a 3 requieren nuestra atención, por ser altamente improbables. Una vez descartada la posibilidad de error en la introducción de los datos que justifique tales valores extremos, queda a juicio del investigador determinar la conveniencia o no de incluir dichos casos en el análisis de regresión. En general, cuanto menor es el número de casos y mayor el valor normalizado del residuo de los mismos mayor es la probabilidad de que el analista descarte dichos valores a la hora de definir el modelo de regresión. Como regla general podemos decir que nuestro modelo es inaceptable si se da alguna de las circunstancias siguientes (Field, 2000, p. 123):

- Más del 1% de los casos tienen residuos normalizados cuyo valor absoluto es mayor que 2,5.
- Más del 5% de los casos tienen residuos normalizados cuyo valor absoluto es mayor que 2,0.

Podemos descartar la existencia de valores extremos simplemente obteniendo el valor máximo y el mínimo de los residuos. Para ello utilizamos la siguiente línea de comandos:

Analyses → Descriptive → Distribution Statistics: *z Resid.*

```
DISTRIBUTION PARAMETER ESTIMATES

z Resid. (N =200) Sum =-0.060
Mean =0.000 Variance =0.316 Std.Dev. =0.562
Std.Error of Mean =0.040
Range =2.100 Minimum =-0.830 Maximum =1.270
Skewness = 0.511 Std. Error of Skew = 0.172
Kurtosis =-0.684 Std. Error Kurtosis =0.342
```

Como podemos ver, el valor absoluto de los residuos es inferior a 2,0 por lo que, según este criterio, el modelo es adecuado.

Multicolinealidad

El coeficiente de correlación de las dos variables explicativas no alcanza un valor muy alto ($r_{distancia\text{-}ingresos}$ =0,265) lo que no sugiere, a *priori*, la existencia de un problema grave de multicolinealidad en nuestro modelo. Tampoco se da la circunstancia simultánea de tener coeficientes no significativos (los dos tienen una probabilidad de ser cero inferior a 0,000) y un alto coeficiente de determinación, inclinándonos de igual forma a rechazar este problema. De igual forma llegamos a la misma conclusión si aplicamos la regla de Klien, como muestra la tabla siguiente donde se presentan los coeficientes de determinación del modelo completo y de cada uno de los modelos compuestos por una única variable explicativa:

Tabla 8.2. *Coeficientes de determinación de la prueba de Klein*

Modelo	Especificación	R^2 ajustado
Completo	$DAP=b_0 + b_1 \cdot DISTANCIA + b_2 \cdot INGRESOS$	0,681
Parcial 1	$DAP=b_0 + b_1 \cdot DISTANCIA$	0,583
Parcial 2	$DAP=b_0 + b_1 \cdot INGRESOS$	0,253

Como muestra la Tabla 8.1, ninguno de los modelos parciales tiene un coeficiente de determinación superior al del modelo global por lo que, según la regla de Klien, descartamos la existencia de multicolinealidad.

Por último, también podemos utilizar el factor de inflación de la varianza (FIV, o su acrónimo anglosajón VIF) y el coeficiente de tolerancia para analizar el modelo. El factor de inflación de la variable i (FIVi) se calcula como sigue (Gujarati, 1995, p. 338):

$$FIV_i = \frac{1}{1 - R_i^2}$$

siendo R_i^2el coeficiente de determinación de la regresión que tiene como variable dependiente X_i y como variables explicativas el resto de las variables explicativas del modelo inicial.

En nuestro ejemplo tenemos sólo dos variables explicativas: DISTANCIA e INGRESOS, por tanto:

- $R^2_{DISTANCIA}$ es el coeficiente de determinación de la regresión auxiliar: *DISTANCIA*=b_0 + b_1·*INGRESOS*,
- $R^2_{INGRESOS}$ es el coeficiente de determinación de la regresión auxiliar: *INGRESOS*=b_0 + b_1·*DISTANCIA*.

En este caso, al tener sólo dos variable explicativas $R^2_{DISTANCIA}$ =$R^2_{INGRESOS}$. Dicha regresión auxiliar da un coeficiente de determinación igual a 0,070 por lo que $VIF_{DISTANCIA}$=$VIF_{INGRESOS}$=1/(1-0,070)=1,075, muy por debajo de 10, valor a partir del cual nos indica que esa variable presenta un problema grave de multicolinealidad. Tampoco la tolerancia de ambas variables (1/1,075=0,93) es inferior a 0,20, por lo que llegamos a la misma conclusión.

No es necesario sin embargo calcular manualmente el valor del factor de inflación de la varianza, OS4 lo genera automáticamente junto con la tolerancia en la salida estándar del análisis de regresión, como muestra el siguiente extracto:

Variable	Beta	B	Std.Error	t	Prob.>t	VIF	TOL
DISTANCIA	0.678	0.009	0.001	16.341	0.000	1.075	0.930
INGRESOS	0.328	0.468	0.059	7.895	0.000	1.075	0.930

El valor que nos da la salida (VIF=1,075, en ambos casos) coincide con el valor calculado con anterioridad. Se incluye de igual forma la tolerancia (Tol=0,930), que no es más que la inversa del valor de inflación de la varianza.

Homogeneidad de la varianza

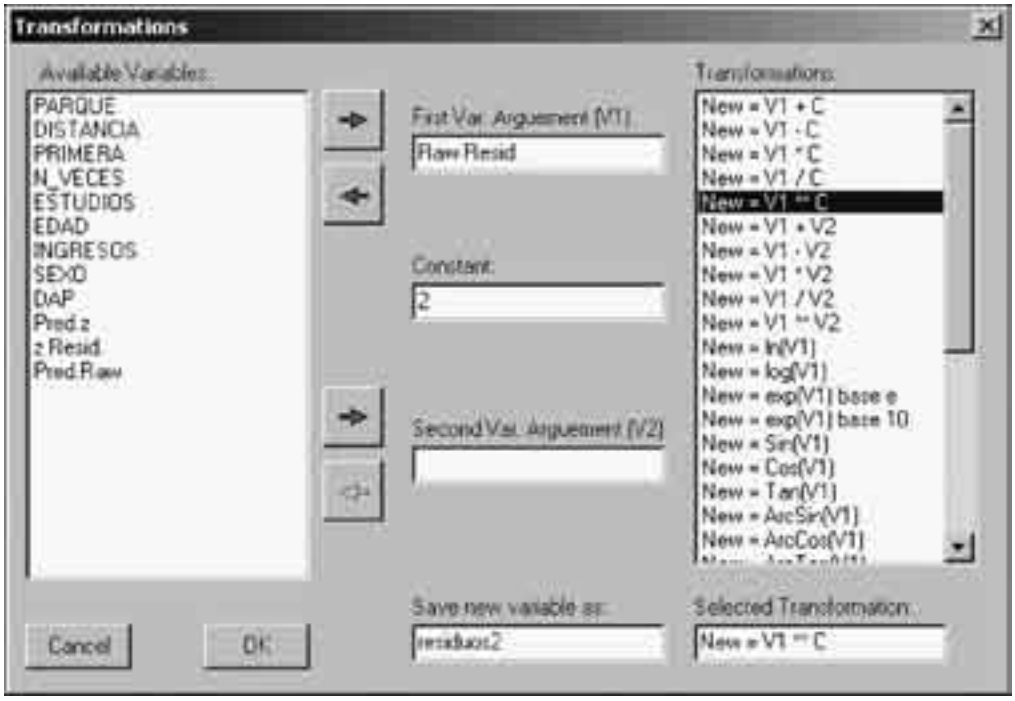
La primera prueba para comprobar la homogeneidad de la varianza (homocedasticidad) que utilizamos es la de **Park**. En ella, la variable dependiente es el cuadrado de los residuos de la regresión inicial. Primero obtenemos los residuos de esta primera regresión:

Analyses → Regression → Block Entry: Dependent Variable: *Dap*; Independent Var's to enter in block: *Distancia, Ingresos* ⊙Predictions, residuals, C.I.'s to grid (para guardar los residuos, esto es, la diferencia entre el valor observado de *DAP* y el valor estimado por el modelo en el panel de datos).

La columna "`Raw Resid.`" (residuos no normalizados) es la que utilizaremos para las pruebas de homocedasticidad, para ello el primer paso consiste en la obtención del cuadrado de los residuos

Variables → Transform; First Var. Argument (V1): *Raw Resid.*; Constant: 2; Save New Variable As: *residuos2*; Select Transformation: *New =V1^C* → Compute → Return

Figura 8.4. *Cálculo del cuadrado de los residuos*



A continuación calculamos cada una de las regresiones auxiliares de la prueba de Park. En la primera regresión la variable dependiente es residuos2 y la independiente *DISTANCIA*. En la segunda cambiamos la variable independiente por *INGRESOS*. En el primer caso tenemos:

```
Block Entry Multiple Regression by Bill Miller

----------------- Trial Block 1 Variables Added ------------------
R            R2      F       Prob.>F     DF1        DF2
0.123        0.015   3.061   0.082       1          198
Adjusted R Squared =0.010

Std. Error of Estimate =0.683

   Variable     Beta        B    Std.Error        t   Prob.>t
  DISTANCIA    0.123    0.001        0.000    1.750     0.082

Constant =0.541
Increase in R Squared =0.015
F = 3.061 with probability =0.082
Block 1 did not meet entry requirements
```

Y en el segundo:

```
Block Entry Multiple Regression by Bill Miller

----------------- Trial Block 1 Variables Added ------------------
          R        R2         F      Prob.>F       DF1        DF2
      0.166     0.028     5.642        0.018         1        198
Adjusted R Squared =0.023

Std. Error of Estimate =0.679

   Variable      Beta         B    Std.Error         t    Prob.>t
   INGRESOS     0.166     0.118        0.050     2.375      0.018

Constant = 0.355
Increase in R Squared = 0.028
F = 5.642 with probability =0.018
Block 1 met entry requirements
```

La Tabla 8.3 resume los resultados anteriores. En ella vemos cómo en la segunda regresión auxiliar el parámetro sí es estadísticamente significativo (prob<0,05), esto es, rechazamos la hipótesis nula de que el parámetro b1 es cero.

Tabla 8.3. *Significación de los coeficientes de las regresiones auxiliares para la prueba de homogeneidad de la varianza*

Regresión auxiliar	Probabilidad *t*
Residuos2=b_0 + b_1·*DISTANCIA*	0,082
Residuos2=b_0 + b_1·*INGRESOS*	0,018

A la vista de estos resultados concluimos que, según la prueba de Park, existe un problema de heterocedasticidad causado por la variable INGRESOS. Antes de aplicar cualquiera de las medidas correctoras para solucionar este problema conviene continuar con las otras dos pruebas para corroborar nuestro diagnóstico.

La siguiente prueba que podemos utilizar para comprobar la homogeneidad de la varianza es la de **Breusch-Pagan-Godfrey**. OS4 nos proporciona automáticamente el valor del estadístico y su probabilidad de esta prueba (en el cuadro de diálogo de la regresión *Block Entry* marcar la opción *BPG Heteroscedasticity Test)*

```
Block Entry Multiple Regression by Bill Miller REGRESIÓN INICIAL

---------------- Trial Block 1 Variables Added ------------------
    SOURCE      DF        SS         MS         F  Prob.>F
Regression       2   260.236    130.118   213.816    0.000
  Residual     197   119.884      0.609
     Total     199   380.120

Dependent Variable: DAP
         R     R2         F    Prob.>F      DF1      DF2
     0.827  0.685   213.816      0.000        2      197
Adjusted R Squared =0.681
Std. Error of Estimate =0.780

  Variable    Beta         B  Std.Error         t  Prob.>t    VIF    TOL
 DISTANCIA   0.678     0.009      0.001    16.341    0.000  1.075  0.930
  INGRESOS   0.328     0.468      0.059     7.895    0.000  1.075  0.930
Constant =-0.410
Increase in R Squared =0.685
F =213.816 with probability =0.000
Block 1 met entry requirements
```

```
============================================================
Breusch-Pagan-Godfrey Test of Heteroscedasticity
============================================================

Auxiliary Regression

    SOURCE      DF        SS         MS       F   Prob.>F
Regression       2     9.006      4.503   3.518     0.032
  Residual     197   252.182      1.280
     Total     199   261.188

Dependent Variable: BPGResid.

         R      R2         F    Prob.>F     DF1       DF2
     0.186   0.034     3.518      0.032       2       197
Adjusted R Squared =0.025

Std. Error of Estimate =1.131

  Variable    Beta         B  Std.Error       t   Prob.>t    VIF    TOL
 DISTANCIA   0.085     0.001      0.001   1.175     0.242  1.075  0.930
  INGRESOS   0.144     0.170      0.086   1.982     0.049  1.075  0.930

Constant =0.581

Breusch-Pagan-Godfrey Test of Heteroscedasticity
Chi-Square =4.503 with probability greater value =0.105
```

El estadístico que buscamos es θ=4,503, con una probabilidad, bajo el supuesto de homocedasticidad (H_0=homocedasticidad), igual a 0,105. Siendo esta probabilidad superior a 0,05 rechazamos la hipótesis nula de homocedasticidad.

Al igual que la prueba de Park, que sugería un ligero problema de heterocedasticidad debido a la variable INGRESOS, la prueba de Breusch-Pagan-Godfrey corrobora esta hipótesis. Sin embargo, teniendo en cuenta que ambas pruebas violan uno de los supuestos en las que se basan (los residuos de las regresiones auxiliares deben seguir una distribución normal[32]), nos lleva a considerar los resultados anteriores con cautela. Para superar esta limitación de la normalidad de los residuos de las regresiones auxiliares podemos recurrir a la prueba de White, que tiene la ventaja de no imponer ninguna restricción a la distribución de los residuos de dicha regresión auxiliar.

La prueba de White, al igual que las otras, comienza con el cálculo de los residuos de la regresión inicial y de sus cuadrados. Seguidamente obtenemos el coeficiente de determinación de la regresión auxiliar:

$residuos^2 = b_0 + b_1 \cdot DISTANCIA + b_2 \cdot INGRESOS + b_3 \cdot DISTANCIA \cdot INGRESOS + b_4 \cdot DISTANCIA^2 + b_5 \cdot INGRESOS^2 + v_i$

```
Block Entry Multiple Regression by Bill Miller

 --------------- Trial Block 1 Variables Added ----------------
          R       R2          F      Prob.>F     DF1        DF2
      0.268    0.072      3.010        0.012       5        194
   Adjusted        R    Squared       =0.048

Std. Error of Estimate =0.670

Variable        Beta          B    Std.Error        t    Prob.>t
  DISTANCIA    0.518      0.004        0.002    1.819      0.070
   INGRESOS    0.858      0.609        0.242    2.515      0.013
  DIST·INGR   -0.039      0.000        0.000   -0.199      0.842
 DISTANCIA²   -0.398      0.000        0.000   -1.579      0.116
  INGRESOS²   -0.751     -0.112        0.051   -2.183      0.030

Constant =-0.201
Increase in R Squared =0.072
F =3.010 with probability =0.012
Block 1 met entry requirements
```

A continuación calculamos el producto R^2*n: 0,072*200=14,40. Teniendo en cuenta que el valor crítico de una Chi-cuadrado con 5 grados de libertad (5 variables explicativas) es igual a 11,07, rechazamos la hipótesis nula de homocedasticidad. Sin embargo no rechazaríamos la hipótesis nula de homocedasticidad si el nivel de significación se sitúa en el 1% (valor crítico igual a 15,12). ¿Qué decisión debemos tomar? Considerando que el grado de incumplimiento no es muy importante (no incumple al 1% de nivel de significación) no resultaría arriesgado aceptar el modelo propuesto, el cual reproducimos a continuación, como válido.

```
Block Entry Multiple Regression by Bill Miller

----------------- Trial Block 1 Variables Added ------------------

        R     R2         F     Prob.>F     DF1       DF2
    0.827  0.685   213.816       0.000       2       197

Adjusted R Squared =0.681

Std. Error of Estimate =0.780

Dependent Variable: DAP

 Variable    Beta         B  Std.Error        t  Prob.>t     VIF     TOL
DISTANCIA   0.678     0.009      0.001   16.341    0.000   1.075   0.930
 INGRESOS   0.328     0.468      0.059    7.895    0.000   1.075   0.930

Constant =-0.410
Increase in R Squared =0.685
F =213.816 with probability =0.000
Block 1 met entry requirements
```

Otro aspecto que apoya nuestra decisión son los valores alcanzados por el estadístico *t* y sus probabilidades correspondientes (16,341 y 7,895 con probabilidad 0,000 en ambos casos). En efecto, la estimación del estadístico *t* se ve afectada por la presencia de heterocedasticidad sin embargo, el hecho de sufrir nuestro modelo de un problema leve de heterocedasticidad y de que las probabilidades de *t* no se aproximan a 0,05 (si exceden esta probabilidad no podríamos rechazar la hipótesis nula de que el coeficiente es cero) nos permite ampliar el margen de confianza en el modelo propuesto.

En el caso de enfrentarnos a un problema grave de heterocedasticidad, por ejemplo si rechazamos con la prueba de White la hipótesis nula de homocedasticidad incluso al 1% de nivel de significación, tenemos dos opciones para tratar de corregir, o al menos reducir, este problema:

Opción 1. Transformación de las variables del modelo

Podemos intentar corregir el problema estimando alguno de los modelos siguientes:

- $DAP = b_0 + b_1 \cdot DISTANCIA + b_2 \cdot \ln(INGRESOS) + u_i$
- $DAP = b_0 + b_1 \cdot \ln(DISTANCIA) + b_2 \cdot INGRESOS + u_i$
- $\ln(DAP) = b_0 + b_1 \cdot \ln(DISTANCIA) + b_2 \cdot \ln(INGRESOS) + u_i$

Nota: En el caso de tener valores nulos en alguna de las variables a las que se le aplica el logaritmo neperiano basta con crear una nueva variable que sea igual a la anterior más una cantidad mínima que no afecte en la práctica al valor de los coeficiente del modelo de regresión.

Opción 2. Buscar una relación entre los residuos (su cuadrado) y alguna de las variables explicativas

Como explicamos en la Figura 8.2, podemos encontrar una relación entre el cuadrado de los residuos de nuestro modelo heterocedástico y alguna de las variables explicativas. Si este es el caso, podemos utilizar esta variable explicativa para transformar todas

las variables del modelo con el objeto de corregir el problema de heterocedasticidad. En nuestro ejemplo el problema de heterocedasticidad al ser leve no permite observar ninguna relación clara entre estas variables, como podemos ver en la figura siguiente a través del diagrama de dispersión:

Analyses → Descriptive → Plot X versus Y: X Axis Variable: *Distancia* Y Axis Variable: *Resid2* ⊙ Plot Regression Line -de igual forma para *Ingresos*-.

Figura 8.5. *Diagrama de dispersión del cuadrado de los residuos y las variables explicativas del modelo*

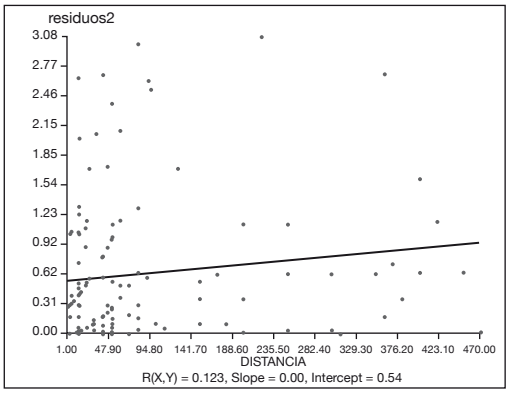


Diagrama de dispersión Distancia-u_i^2

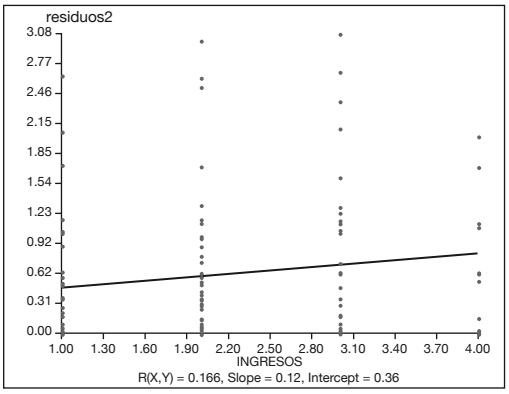


Diagrama de dispersión Ingresos-u_i^2

En el primer caso, la relación entre la variable DISTANCIA y el cuadrado de los residuos del análisis de regresión rechaza la hipótesis de relación entre ambas variables (probabilidad de *t* igual a 0,082 que es mayor que 0,05 por lo que no rechazamos la hipótesis nula de que el coeficiente es cero), como muestra la salida siguiente:

```
Block Entry Multiple Regression by Bill Miller

     -------------- Trial Block 1 Variables Added ---------------

         R      R2       F    Prob.>F     DF1       DF2
     0.123   0.015   3.061      0.082       1       198
Adjusted R Squared =0.010

Std. Error of Estimate =0.683

Variable      Beta       B  Std.Error       t   Prob.>t     VIF    TOL
 DISTANCIA   0.123   0.001      0.000   1.750     0.082   1.000  1.000

Constant =0.541
Increase in R Squared = 0.015
F = 3.061 with probability = 0.082
Block 1 did not meet entry requirements
```

En el segundo caso, en la relación entre los INGRESOS y el cuadrado de los residuos, el coeficiente de correlación de Spearman (porque una de las variables es ordinal) resultó

significativo (probabilidad igual a 0,0059, menor que 0,05, por lo que rechazamos la hipótesis nula H_0: r=0) si bien el grado de correlación (r=0,194) no es muy alto.

```
Spearman Rank Correlation =0.194
t-test value for hypothesis r =0 is 2.782
Probability > t =0.0059
```

En consecuencia, si quisiéramos corregir el problema de heterocedasticidad la variable instrumental sería INGRESOS ya que es la que ha mostrado una cierta relación con el cuadrado de los residuos. Por tanto, el modelo original se transformaría como sigue:

Modelo original:
$DAP_i = b_0 + b_1 \cdot DISTANCIA_i + b_2 \cdot INGRESOS_i + u_i$

Modelo transformado:
$DAP_i / INGRESOS_i = b_0 / INGRESOS_i + b_1 \cdot DISTANCIA_i / INGRESOS_i + b_2 \cdot + v_i$

Es decir, la nueva regresión tendrá la forma habitual:
$Y_i = b_2 + b_0 X_{1i} + b_1 \cdot X_{2i} + v_i$

Este nuevo modelo debería eliminar o reducir el problema de heterocedasticidad de los residuos (v_i en lugar de u_i). Una vez calculados los coeficientes de esta regresión auxiliar procederíamos a deshacer los cambios hasta obtener de nuevo el modelo original.

Distribución de los residuos

En el modelo de regresión lineal clásico los residuos tienen media y covarianza igual a cero y varianza constante. Si bien el supuesto de normalidad es más restrictivo que los requisitos anteriores, su cumplimiento permite la utilización de las pruebas *t* y *F* independientemente del tamaño de la muestra. En primer lugar obtengamos los residuos de nuestro modelo:

Analyses → Regression → Block Entry: Dependent Variable: *Dap; Ind.* Var's to enter in block: *Distancia, Ingresos*; ⊙Predictions, residuals, C.I.'s to grid -sólo esta opción-.

Cerrando la ventana de resultados podemos ver que se han añadido a la hoja de datos 7 nuevas variables. De ellas nos interesan:

- `Pred.Raw`: es el valor estimado de DAP que se obtiene de sustituir los valores de las variables explicativas DISTANCIA e INGRESOS en la fórmula anterior.
- `Raw.Resid`: es la diferencia entre el valor estimado y el observado.

Para comprobar la normalidad de los residuos utilizamos la función de OS4: Analyses → Descriptive → Normality Tests: Test Normality of: *Raw.Resid* → Apply.

Según los resultados de la prueba rechazamos la hipótesis nula de normalidad de los residuos (*Strong evidence against normality.*). El incumplimiento de este requisito nos deja con dos opciones:

- Recurrir al tamaño de la muestra Si la muestra es superior a 50 casos podríamos justificar la no normalidad de los residuos amparándonos en el teorema central de límite. De cualquier forma, seguiría siendo necesario comprobar que se cumplen los supuesto de residuos con esperanza igual a 0 y varianza constante (homocedasticidad).
- Transformar alguna de las variables explicativas. Mediante los diagramas de dispersión podemos detectar relaciones entre los residuos y alguna de las variables explicativas y recurrir a transformaciones logarítmicas o de raíces de dicha variable para corregir el problema.

En el primer caso, teniendo en cuenta que el número de elementos de nuestra muestra excede ampliamente el punto crítico de los 50 casos, podríamos asumir la eficiencia de los estimadores de mínimos cuadrados asintóticamente, si bien no seríamos capaces de determinar el grado de acercamiento a esa eficiencia. Así, quedaría por demostrar que los residuos tienen una esperanza igual a cero y la varianza es constante.

Respecto a la media: Analyses → Descriptive → Mean,Var,SD,Range,Skew,Kurt → *Raw.Resid.*

```
DISTRIBUTION PARAMETER ESTIMATES

Raw Resid. (N =200) Sum =0.050
Mean =0.000 Variance =0.602 Std.Dev. =0.776
Std.Error of Mean =0.055
Range =2.890 Minimum =-1.750 Maximum =1.140
Skewness =-0.513 Std. Error of Skew = 0.172
Kurtosis =-0.684 Std. Error Kurtosis =0.342
```

Como podemos ver en los resultados anteriores, la media es igual a cero. Teniendo en cuenta además que los residuos, según las pruebas de Park, de Breusch-Pagan-Godfrey y de White realizadas en la sección anterior, presentan un problema de heterocedasticidad asumible podemos concluir con ciertas garantías que el modelo es válido desde el punto de vista de los supuestos del modelo de regresión lineal clásico.

En el caso de no disponer de una muestra con más de 50 casos, o en el supuesto de buscar una mayor precisión en las estimaciones podemos intentar corregir el problema de normalidad de los residuos. Para ello resulta útil visualizar la relación entre los residuos y las variables explicativas métricas[33], como se mostraba en la sección anterior.

Teniendo en cuenta que en la mayoría de las ocasiones no podremos ver una relación clara entre los residuos y las variables explicativas, una estrategia a seguir para la normalización de los residuos consiste en la transformación logarítmica de la variable explicativa métrica que tenga un mayor coeficiente Beta (en valor absoluto) en la regresión. Por tanto, el nuevo modelo podría ser el siguiente:

$DAP = b_0 + b_1 \cdot \ln(DISTANCIA) + b_2 \cdot INGRESOS$

Una vez transformada la variable DISTANCIA mediante su logaritmo (ln_DISTANCIA), los resultados de la regresión son:

```
Block Entry Multiple Regression by Bill Miller
     -------------- Trial Block 1 Variables Added --------------
            R      R2          F    Prob.>F     DF1          DF2
        0.748   0.559    125.109      0.000       2          197
Adjusted R Squared =0.555
Std. Error of Estimate =0.922
Variable         Beta          B  Std.Error       t       Prob.>t
ln_DISTANCIA    0.582      0.664      0.057  11.629         0.000
     INGRESOS   0.317      0.454      0.071   6.345         0.000
Constant =-2.043
F =125.109 with probability =0.000
Block 1 met entry requirements
```

Si aplicamos la prueba de normalidad a los residuos de este modelo comprobaremos que sí siguen una distribución normal (*No evidence against normality*). Por tanto, hemos corregido el problema de la no normalidad de los residuos a cambio de una reducción en la capacidad del modelo para explicar la variabilidad de la variable dependiente, esto es, el coeficiente de determinación ha pasado de 0,689 a 0,559.

Por tanto, si bien es posible asumir el error de la no normalidad de los residuos de la primera regresión ($DAP=b_0+b_1 \cdot DISTANCIA+b_2 \cdot INGRESOS$) ya que la muestra excede de 50 casos, sería aconsejable optar por el segundo modelo, con la variable ln(DISTANCIA) en lugar de DISTANCIA, en el caso de querer ser más precisos en la determinación de los coeficientes del modelo. Sin embargo, como hemos apuntado, cuanto mayor sea la muestra menos error cometeremos obviando la normalidad de los residuos.

Podemos ver las diferencias de los dos modelos a la hora de realizar estimaciones de la disposición a pagar por la visita al parque en función de la distancia recorrida y el nivel de ingresos. En el primer ejemplo ambos modelo estiman una DAP similar, sin embargo en el segundo ejemplo la diferencia excede el 50%, por lo que claramente la elección de un modelo u otro tendrá una influencia decisiva en nuestras predicciones.

Ejemplo 1: DISTANCIA=200 e INGRESOS=2

- Modelo 1, con residuos no normalizados:
  · DAP=-0,410 + 0,009·(200) + 0,468·(2) =2,33 euros
- Modelo 2, con residuos normalizados:
  · DAP=-2,043 + 0,664·ln(200) + 0,454·(2) =2,38 euros

Ejemplo 2: DISTANCIA=100 e INGRESOS=1

- Modelo 1, con residuos no normalizados:
  · DAP=-0,410 + 0,009·(100) + 0,468·(1) =0,96 euros
- Modelo 2, con residuos normalizados:
  · DAP=-2,043 + 0,664·ln(100) + 0,454·(1) =1,47 euros

Todos somos muy ignorantes.
Lo que ocurre es que no todos ignoramos las mismas cosas
(Albert Einstein)

CAPÍTULO 9

REGRESIÓN LOGÍSTICA

Capítulo 9. Regresión logística

9.1. Introducción teórica

Especificación del modelo

En el capítulo anterior hemos analizado el efecto de un conjunto de variables sobre una variable dependiente métrica mediante el modelo de regresión lineal múltiple, sin embargo, es frecuente manejar variables dependientes de tipo ordinal (por ejemplo, el nivel de ingresos o el grado de satisfacción de un cliente medido mediante una escala Likert) o dicotómicas (por ejemplo, la visita o no a un parque determinado o la ocurrencia o no de una enfermedad).

La mayoría de los modelos econométricos (Maddala, 1983) que tratan este tipo de variables pertenecen a una de las siguientes categorías:

- Modelos de variable truncada, donde la muestra sólo cubre una parte de la población y, por tanto, sólo podemos observar y_i cuando $y_i \leq c$ (ó $y_i \geq c$) siendo c una constante determinada;
- Modelos de variable censada, en este caso se dispone de los valores de las variables explicativas para $y_i \leq c$;
- Modelos de variable dicotómica, donde la variable dependiente es binaria, tomando valores cero o uno. El modelo desarrollado en este capítulo pertenece a esta categoría.

Una de las posibles aproximaciones a modelos econométricos de variable dependiente binaria es el modelo lineal probabilístico. Considerando la variable dependiente y_i que toma el valor 1 cuando un individuo cumple una condición (por ejemplo visita el parque Cazorla) y 0 cuando no (visita cualquier otro parque), un conjunto de variables explicativas x_i con sus respectivos parámetros β' y el error u_i, el modelo lineal probabilístico se presenta como: $y_i=\beta' x_i+ u_i$

El modelo anterior de probabilidad condicional, $P(y_i=1|x_i)$, tiene el inconveniente de hacer crecer la probabilidad de forma lineal con los valores de x_i, planteamiento poco realista ya que a partir de ciertos valores de x_i la probabilidad no debería crecer significativamente (Gujarati, 1995, p. 542; Menard, 2002, p. 8). Las dos alternativas más frecuentes a este modelo lineal son los modelos *logit* y *probit* [34]. El primero tiene la siguiente formulación:

$$P_i=E(y_i=1|x_i)=\frac{1}{1+e^{-\beta' x_i}}$$

Para la obtención de los parámetros del modelo (β') el método de los mínimos cuadrados ordinarios no es aplicable por la heterocedasticidad de los residuos y la no distribución normal de los mismos (Cramer, 1991). Por tanto, los parámetros del modelo logístico (*logit*) suelen estimarse mediante el estimador de máxima verosimilitud.

Bondad del ajuste

Podemos evaluar la bondad del ajuste de nuestro modelo logístico a través de tres procedimientos principalmente:

- Estadístico del coeficiente de verosimilitud.
- Pseudo coeficiente de determinación (R^2).
- Capacidad predictiva del modelo.

Estadístico del coeficiente de verosimilitud

La prueba F es tradicionalmente usada para comprobar si la hipótesis nula de que todos los coeficientes son cero es rechazada o no. En el modelo logístico se dispone de una prueba similar (Davidson y Mackinnon, 1984; Engle, 1984) que utiliza la ratio definida como: c=-2(lnL_0- lnL_1). Donde L_1 es el valor de la función de probabilidad en el modelo logístico con todas las variables y L_0 el correspondiente valor en el modelo reducido, esto es, todos los coeficientes cero, excepto el término independiente. El resultado habitual de los paquetes econométricos es el logaritmo de este valor. Este ratio se distribuye aproximadamente como una chi cuadrado con *K-1* grados de libertad (siendo *K* el número de parámetros a comprobar) cuando la hipótesis nula es verdadera.

No es necesario repetir la regresión sólo con la constante para obtener $\ln L_0$, ya que podemos utilizar la siguiente igualdad:

$$\ln L_0 = N_0 * \ln\frac{N_0}{N} + N_1 * \ln\frac{N_1}{N}$$

con N representando el tamaño de la muestra y N_0 y N_1 número de observaciones para Y=0 y Y=1, respectivamente.

Pseudo coeficiente de determinación

En los modelos logísticos podemos medir la bondad del ajuste mediante un coeficiente de determinación conocido como pseudo-R^2 cuya interpretación es similar al coeficiente tradicional R^2 de la regresión lineal múltiple. La formulación de este coeficiente varía según el autor que se consulte, entre ellos destacamos:

- Cragg y Uhler (1970) $\rightarrow$ Pseudo-R^2=($L_1^{2/n}$- $L_0^{2/n}$) / (1- $L_0^{2/n}$)
- McFadden (1974) $\rightarrow$ Pseudo-R^2=1- $\ln L_1$ / $\ln L_0$
- Maddala (1983) $\rightarrow$ Pseudo- R^2=1- $(L_0/L_1)^{2/n}$

En general los valores de este pseudo coeficiente de determinación no son demasiado altos, sin embargo hay que tener en cuenta que en este tipo de modelos el límite superior no es 1 sino 1- $L_0^{2/n}$, por tanto en lugar de estar acotado entre 0 y 1 se cumple la siguiente igualdad: $0 \leq$ pseudo-$R^2 \leq$ 1- $L_0^{2/n}$.

Otra medida intuitiva de la bondad del ajuste está basada en la ratio de la diferencia del coeficiente de verosimilitud del modelo nulo y del propuesto y el coeficiente de verosimilitud del modelo nulo (Hosmer y Lemeshow, 1989).

Comprobación de los supuestos del modelo

Cuando no se cumplen los supuestos del modelo logístico los resultados pueden ser sesgados (los coeficientes son demasiado altos o bajos) o ineficientes (el error estándar es demasiado alto para la dimensión del coeficiente). En el origen de estos problemas podemos señalar (Menard, 2002, p. 67):

- Incorrecta especificación del modelo. Por ejemplo, falta alguna variable explicativa importante.
- Omisión de casos No existen casos en alguna de las celdas de la tabla de contingencia que cruza la variable dependiente (Y=0 ó Y=1) y una variable independiente categórica.
- No normalidad de los residuos. Al igual que en la regresión lineal múltiple, los residuos deben seguir una distribución determinada. Mientras en el primer caso deben seguir una distribución normal, en el segundo, la regresión logística, deben seguir una distribución binomial (que puede aproximarse a una normal para muestras de gran tamaño).
- Multicolinealidad. Alguna de las variables independientes tienen un grado de correlación alto.

El primer supuesto sólo se puede comprobar basándonos en el juicio del investigador ya que asumimos, a *priori*, que las variables explicativas importantes están presentes en el modelo.

La comprobación del número de casos es fácilmente verificable y, en general, con un tamaño muestral suficiente no es probable que todos los individuos de una categoría (por ejemplo nivel de ingresos) presenten la característica analizada (Y=1) o que todos no la presenten (Y=0).

Al contrario que en el modelo de regresión lineal múltiple, la violación del supuesto de normalidad de los residuos no es demasiado problemática en el modelo logístico (Menard, 2002, p. 83).

Por último, el problema de la multicolinealidad, a diferencia del anterior problema, sí debe tenerse en cuenta en todo análisis de regresión logística.

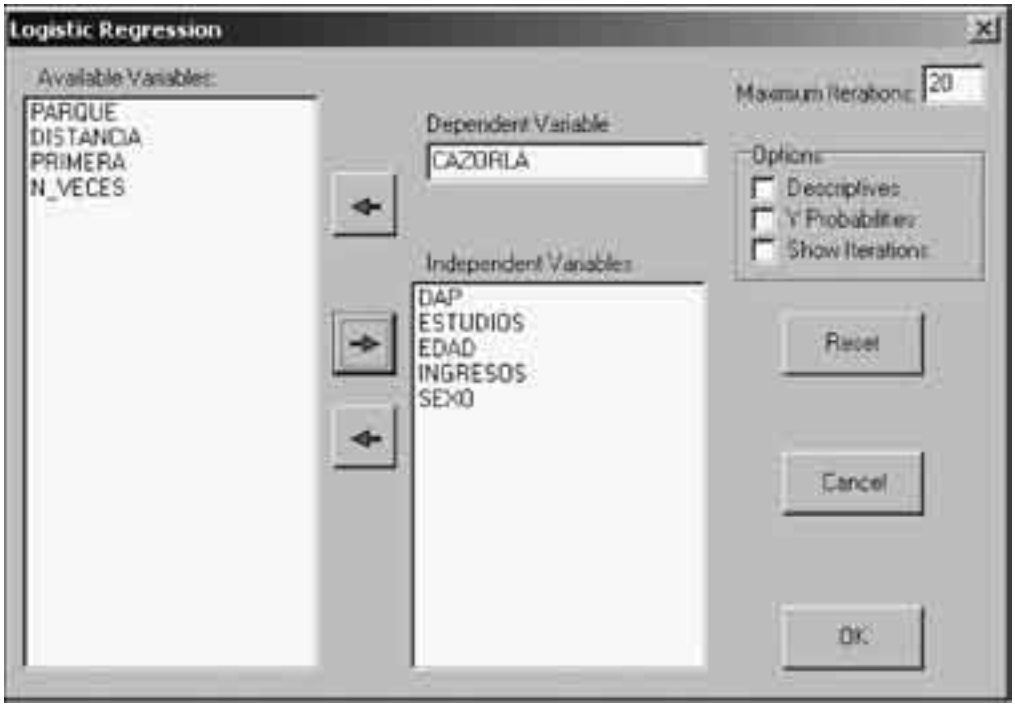
9.2. Ejemplo de regresión logística

Selección de las variables explicativas

Vamos a determinar qué variables explican la elección del parque de Cazorla frente al resto de parques. Nuestra hipótesis de partida incluye a las variables DAP, ESTUDIOS, EDAD, INGRESOS y SEXO como variables explicativas. El modelo se especifica como sigue:

Analyses → Regression → Logistic regression → Dependent Variable: *Cazorla*; Independent Variables: *Dap, Estudios, Edad, Ingresos, Sexo*. -Sin ninguna opción-.

Figura 9.1. *Cuadro de diálogo de la regresión logística*



Tras pulsar el botón “OK” obtenemos el siguiente resultado:

```
Logistic Regression Adapted from John C. Pezzullo
176 cases have Y=0; 24 cases have Y=1.
Variable Label Average Std.Dev.
Converged
Overall Model Fit... Chi Square =53.9918 with df =5 and prob. =0.0000

Coefficients and Standard Errors...
Variable         Label     Coeff.      StdErr          p
          1         DAP     1.4494      0.2682     0.0000
          2   ESTUDIOS      0.2690      0.2938     0.3599
          3        EDAD     0.0750      0.2161     0.7284
          4   INGRESOS     -0.9841      0.3422     0.0040
          5        SEXO     -0.2370      0.6043     0.6949
  Intercept     -3.1802

Odds Ratios and 95% Confidence Intervals...
Variable          O.R.        Low   --     High
       DAP      4.2606     2.5187        7.2071
   ESTUDIOS     1.3087     0.7357        2.3277
      EDAD      1.0779     0.7058        1.6463
   INGRESOS     0.3738     0.1911        0.7310
      SEXO      0.7890     0.2414        2.5791
```

Vemos cómo el número de visitantes al parque de Cazorla (Y=1) es igual a 24, mientras los otros 176 visitaron otros parques (Y=0). La columna p indica la probabilidad de la hipótesis nula de que el coeficiente de la variable sea cero, es decir, que dicha variable no influya en la decisión de visitar Cazorla. Como es habitual, rechazamos aquellos coeficientes con una probabilidad superior a 0,05. En este caso, las variables ESTUDIOS, EDAD y SEXO deben eliminarse del modelo. Repetimos pues los cálculos sólo con las variables DAP e INGRESOS e incluimos las opciones de probabilidad estimada e iteraciones:

Analyses → Regression → Binary Logistic → Independent Variables: *Dap, Ingresos*; Dependent Variable: *Cazorla*; ⊙Y Probabilities; ⊙Show Iterations.

```
Logistic Regression Adapted from John C. Pezzullo
Java program at http://members.aol.com/johnp71/logistic.html

176 cases have Y=0; 24 cases have Y=1.

Iteration History
-2 Log Likelihood = 146.7700 (Null Model)
-2 Log Likelihood = 105.6567
-2 Log Likelihood =  94.2500
-2 Log Likelihood =  93.8363
-2 Log Likelihood =  93.8319
-2 Log Likelihood =  93.8319
Converged

Overall Model Fit... Chi Square = 52.9381 with df = 2 and prob. = 0.0000

Coefficients and Standard Errors...
Variable           Label    Coeff.    StdErr  p
           1         DAP    1.4087    0.2589  0.0000
           2    INGRESOS   -0.8844    0.3187  0.0055
Intercept      -2.7085

Odds Ratios and 95% Confidence Intervals...
Variable      O.R.        Low    --  High
         DAP  4.0907      2.4626    6.7953
    INGRESOS  0.4130      0.2211    0.7713
```

En primer lugar tenemos el valor del coeficiente de verosimilitud multiplicado por –2 para el modelo nulo (no hay variables explicativas) y para el modelo propuesto. En el primer caso este valor es igual a 146,77, en el segundo se sitúa en la última iteración en 93,83. Cuanto menor sea este último valor, el del modelo propuesto, mejor será la capacidad explicativa (y predictiva) de nuestro modelo. La diferencia entre ambos es una medida de esta capacidad, ajustándose a una distribución Chi-cuadrado.

En nuestro caso: 146,77 – 93,83 =52,94. Este valor excede al valor crítico $Chi_{2\ 0,05}$=5,99, por lo que rechazamos la hipótesis nula de que todos los coeficientes de la regresión son cero. De igual manera, la probabilidad del valor 52,94 en la distribución Chi-cuadrado no excede 0,05 (en realidad `prob.=0,0000`) por lo que llegamos a la misma conclusión.

A continuación el programa proporciona los coeficientes del modelo de regresión logística. Así, la probabilidad de visitar el parque de Cazorla se calcula mediante la siguiente expresión:

$$Pi = P(cazorla = 1 \mid b'x) = \frac{1}{1 + e^{-(-2,7085+1,4087*dap-0,8844*ingresos)}}$$

Por ejemplo, para el primer individuo con una DAP=0 y un nivel de ingresos=2 la probabilidad de que visitara Cazorla es 0,0112, inferior a 0,5 por lo que estimamos que no visitó este parque. En este caso la estimación coincide con la realidad (Y=0→No visitó Cazorla). Podemos comprobar que:

$$Pi = P(cazorla = 1 \mid b'x) = \frac{1}{1 + e^{-(-2,7085+1,4087*0-0,8844*2)}} = 0,0112$$

Igualmente, el segundo individuo con una DAP igual a 1,8 y un nivel de ingresos igual a 1 tampoco visitó Cazorla ya que la probabilidad estimada es inferior a 0,5 (exactamente 0,2578).

Bajo la columna O.R. (*Odds Ratios*) tenemos el incremento de probabilidad si la variable independiente aumenta una unidad. Por ejemplo, si la DAP pasa de 0 a 1 € la probabilidad de que el visitante visite el parque de Cazorla se multiplica aproximadamente por 4 (en concreto por 4,09). De igual forma, si el nivel de ingresos sube una categoría la probabilidad se multiplica por 0,4.

Bondad de ajuste

Cuanto mayor sea la diferencia entre los coeficientes de verosimilitud mejor será el ajuste. En nuestro caso esta diferencia alcanza un valor de 52,94, dato obtenido como la diferencia 146,77-93,83. Por tanto, una medida de la bondad del modelo es el cociente 52,94 / 146,77=0,36.

Otras medidas alternativas son el pseudo coeficiente de determinación. OS4, al igual que la mayoría de los paquetes estadísticos, proporciona el valor $-2 \cdot \ln L_0 = 146,77$ y $-2 \cdot \ln L_1 = 93,83$, por lo que $L_0 = 1,35 \cdot 10^{-32}$ y $L_1 = 4,22 \cdot 10^{-21}$.

- Pseudo-R^2 (Cragg y Uhler) = $(L_1^{2/n} - L_0^{2/n}) / (1 - L_0^{2/n}) = [(4,22 \cdot 10^{-21})^{2/200} - (1,35 \cdot 10^{-32})^{2/200}] / [1 - (1,35 \cdot 10^{-32})^{2/200}] = 0,28$
- Pseudo-R^2 (McFadden) $= 1 - \ln L_1 / \ln L_0 = 1 - \ln(4,22 \cdot 10^{-21}) / \ln(1,35 \cdot 10^{-32}) = 0,36$
- Pseudo-R^2 (Maddala) $= 1 - (L_0/L_1)^{2/n} = 1 - (1,35 \cdot 10^{-32}/4,22 \cdot 10^{-21})^{2/200} = 0,23$

Estos valores no son tan bajos como a primera vista parecería ya que el límite superior de R^2 no es 1 como en la regresión lineal múltiple sino $1-(1,35 \cdot 10^{-32})^{2/200} = 0,52$.

Capacidad predictiva del modelo

Al final de la salida de la regresión logística de OS4 se incluye una tabla que clasifica los elementos a partir de los coeficientes estimados de la función logística (para ello marcar la opción *Y Probabilities*). Una vez clasificados se comparan con el valor real observado.

```
Classification Table
             Predicted
           ---------------
Observed     0        1      Total
           ---------------
    0     |  174  |    2    |  176  |
    1     |   15  |    9    |   24  |
           ---------------   |       |
Total     |  189  |   11    |  200
           ---------------
```

Como podemos observar, el modelo predice correctamente la pertenencia de 174 individuos al grupo 0 (no visitó el parque de Cazorla) y de 9 al grupo 1 (visitó Cazorla). Por el contrario, según el modelo 15 individuos no visitaron el parque de Cazorla cuando en realidad sí lo hicieron. De igual forma clasifica erróneamente a 2 individuos que en realidad no visitaron el parque. En total, el porcentaje de aciertos (174+9) se eleva al 91,5% (183/200).

9.3. Análisis discriminante vs regresión logística

Si la variable dependiente es dicotómica, es decir, hay sólo dos grupos, podemos optar por el análisis discriminante o la regresión logística. A continuación comparamos ambos enfoques con resultados prácticamente idénticos aunque con una mayor economía de datos por parte de la regresión logística.

En nuestro ejemplo vamos a explicar la visita o no al parque de Cazorla (No=0, Sí=1) en función de tres variables explicativas: cantidad que el visitante está dispuesto a pagar por la entrada al parque (DAP), la distancia recorrida (DISTANCIA) y si es la primera vez o no que visita el parque (PRIMERA). En primer lugar procedemos con el análisis multivariante de la varianza como paso inicial del análisis discriminante para determinar si las tres variables son significativas o no:

Analyses → Multivariate → Discriminant Function / MANOVA; Group Var. *Cazorla*; Predictor Var: *Dap, Distancia, Primera*; Options: ⊙One-Way ANOVAS ⊙ Compute.

```
MULTIVARIATE ANOVA / DISCRIMINANT FUNCTION
Reference: Multiple Regression in Behavioral Research
Elazar J. Pedhazur, 1997, Chapters 20-21
Harcourt Brace College Publishers

Total Cases:=200, Number of Groups:=2

UNIVARIATE ANOVA FOR VARIABLE DAP
SOURCE            DF             SS           MS          F  PROB > F
BETWEEN            1         99.147       99.147     69.868     0.000
ERROR            198        280.974        1.419
TOTAL            199        380.120

UNIVARIATE ANOVA FOR VARIABLE DISTANCIA
SOURCE            DF             SS           MS          F  PROB > F
BETWEEN            1    1001475.801  1001475.801    200.634     0.000
ERROR            198     988329.379     4991.563
TOTAL            199    1989805.180

UNIVARIATE ANOVA FOR VARIABLE PRIMERA
SOURCE            DF             SS           MS          F  PROB > F
BETWEEN            1          6.110        6.110     40.538     0.000
ERROR            198         29.845        0.151
TOTAL            199         35.955
```

Como vemos, las tres variables explicativas son estadísticamente significativas (todas con `PROB > F =0,0000`). Omitiendo los resultados de las funciones canónicas y de Fisher, contrastamos la validez global del modelo discriminante. Según el valor alcanzado por el estadístico Wilk's Lambda, y por su equivalente F, no rechazamos la hipótesis nula de validez global del modelo.

```
Corr.s Between Variables and Functions with 200 valid cases.
Variables
                    1
         DAP     0.718
   DISTANCIA     0.997
     PRIMERA     0.579

Wilk's Lambda =0.4936.
F =67.0358 with D.F. 3 and 196. Prob > F =0.0000
Bartlett Chi-Square =138.7471 with 3 D.F. and prob. =0.0000
Pillai Trace =0.5064
```

Repetimos el análisis pero ahora con la opción de clasificación de los casos para determinar la capacidad explicativa de la función discriminante:

Analyses → Multivariate → Discriminant Function / MANOVA; Group Var. *Cazorla*; Predictor Var: *Dap, Distancia, Primera*; Options: ⊙Classify Scores ⊙ Compute.

```
CLASSIFICATION TABLE

PREDICTED GROUP
Variables      Valores estimados
               Cazorla=0    Cazorla=1
                                    1            2      TOTAL
Valores                 1          168            8        176
observados              2            7           17         24
                    TOTAL          175           25        200
```

Según los resultados anteriores, la función discriminante predice correctamente la pertenencia al grupo 1 (se corresponde con el primer valor de la variable Cazorla, esto es, Cazorla=0 → no visitó el parque) de 168 casos. De la misma forma también tiene éxito en 17 casos indicando su pertenencia al grupo 2 (segundo valor de la variable, Cazorla=1 → visitó el parque). Por tanto, la capacidad predictiva global del modelo es igual a (168+17)/200=92,5%.

A continuación procedemos con el análisis logístico.

```
Logistic Regression Adapted from John C. Pezzullo
Java program at http://members.aol.com/johnp71/logistic.html

176 cases have Y=0; 24 cases have Y=1.
Variable Label Average Std.Dev.

Converged

Overall Model Fit... Chi Square =71.4388 with df =3 and prob. =0.0000

Coefficients and Standard Errors...
Variable          Label     Coeff.     StdErr          p
         1          DAP    -0.1768     0.3520     0.6155
         2   DISTANCIA      0.0169     0.0045     0.0002
         3     PRIMERA      0.8186     0.7306     0.2625
Intercept   -4.0448

Odds Ratios and 95% Confidence Intervals...
Variable           O.R.        Low --     High
       DAP       0.8380     0.4203     1.6705
  DISTANCIA      1.0171     1.0081     1.0261
    PRIMERA      2.2674     0.5416     9.4933
```

De acuerdo con el resultado de la regresión, sólo la variable DISTANCIA es estadísticamente significativa (p<0,05), por lo que excluimos el resto del análisis logístico con el siguiente resultado:

```
Logistic Regression Adapted from John C. Pezzullo
Java program at http://members.aol.com/johnp71/logistic.html

176 cases have Y=0; 24 cases have Y=1.
Variable Label          Average     Std.Dev.

Converged

Overall Model Fit... Chi Square =70.1500 with df =1 and prob. =0.0000

Coefficients and Standard Errors...
   Variable           Label     Coeff.     StdErr         p
      1           DISTANCIA     0.0170     0.0028    0.0000
Intercept -3.9806

Odds Ratios and 95% Confidence Intervals...
Variable               O.R.        Low       High
     DISTANCIA       1.0171     1.0116     1.0226
```

Esto es, la probabilidad de que un visitante haya elegido Cazorla tiene la siguiente formulación:

$$Pi = P(cazorla = 1 | b' x) = \frac{1}{1 + e^{-(-3,9806+0,0170*distanc)}}$$

Si comparamos la capacidad predictiva de ambos modelos, el discriminante y el logístico, como aparece en la Tabla 9.1, vemos que tienen un comportamiento similar, en términos de número de individuos clasificados correctamente. En efecto, el análisis discriminante predice correctamente la pertenencia de 185 individuos (168+17) frente a 183 en el análisis logístico (170+13), si bien este último utiliza una única variable explicativa (DISTANCIA) frente a las tres del primero (DAP, DISTANCIA y PRIMERA).

Tabla 9.1. *Comparación de la capacidad predictiva del modelo discriminante y el logístico*

		Función discriminante		Función logística		Total
		0	1	0	1	
Observados	0	168	8	170	6	176
	1	7	17	11	13	24

Aunque los resultados anteriores representan sólo un ejemplo, en general, es preferible utilizar el análisis de la regresión logística en lugar del análisis discriminante, dejando este último para la clasificación de variables dependientes con más de dos grupos.

Aquel que duda y no investiga, se torna no sólo infeliz, sino también injusto
(Blaise Pascal)

CAPÍTULO 10

ANÁLISIS DE LA COVARIANZA

Capítulo 10. Análisis de la covarianza

10.1. Combinación del análisis de la varianza y de regresión

Mediante el análisis de la covarianza (ANCOVA) podemos analizar el efecto de una variable nominal u ordinal (factor) y otra métrica (llamada covariable o cofactor) sobre una variable métrica. En ciencias experimentales se utiliza el término *tratamiento* en lugar de *factor*. Por ejemplo, podemos determinar si es significativo el efecto del abono A, B o C (factor o tratamiento) sobre el rendimiento de una planta (variable dependiente) teniendo en cuenta la cantidad de agua aplicada (covariable) mediante el análisis de la covarianza.

Por tanto, el análisis de la covarianza combina aspectos del análisis de la varianza y del análisis de regresión, teniendo su máxima utilidad cuando la covariable y la variable dependiente están fuertemente correlacionadas. La idea central consiste en separar la variabilidad de la variable dependiente que se deba a la covariable con el fin de potenciar las diferencias entre grupos (o tratamientos).

Para que las conclusiones a las que llegamos mediante la prueba F sobre el efecto de la variable de grupo y la covariable sean válidas es necesario comprobar que se cumplen los supuestos paramétricos correspondientes al análisis de la varianza y del análisis de regresión:

- La covariable es independiente de la variable de grupo o tratamiento. Para medir esta independencia podemos utilizar el análisis de la varianza (ANOVA), cuya hipótesis nula sería la no diferencia de los valores de la covariable entre los grupos (Muñoz Serrano, 2003, p. 649; Field, 2000, p. 294).
- Homogeneidad de la covarianza entre grupos. Si esta hipótesis se cumple, la interacción entre la covariable y el factor no debería tener un efecto significativo sobre la varianza de la variable dependiente.
- Igualdad de pendientes para cada grupo, es decir, la relación entre la variable dependiente y la covariable no varía entre los grupos del factor (líneas paralelas). Si las pendientes de la regresión para cada grupo son diferentes el modelo global no es correcto (Field, 2000, p. 307).
- Los residuos tienen media cero y varianza constante.

10.2. El modelo lineal ANCOVA

Supongamos que tenemos dos tratamientos o factores, el primero (X) con dos categorías y el segundo (Z) con tres. El primero puede ser el *Sexo* (Hombre / Mujer) y el segundo el *Municipio* (Ciudad 1 / Ciudad 2 / Ciudad 3). Cada factor da lugar a n-1 variable ficticias, ya que la última se representa con un cero en todas las anteriores. Por ejemplo, la variable Z da lugar a dos variables: Z_1 (que toma el valor 0 si vive en la Ciudad 2 o 3, y 1 si vive en la Ciudad 1) y Z_2 (que toma el valor 0 si vive en la Ciudad 1 o 3, y 1 si vive en la Ciudad 2). Los habitantes de la Ciudad 3 se representan por $Z_1=Z_2=0$.

Finalmente tenemos la covariable Edad y la variable dependiente *Gasto en bienes de lujo* (Y). El modelo lineal completo ANCOVA incluye todas las variables ficticias y las interacciones entre factores:

Modelo completo 1: $Y=b_0 + b_1 \cdot X_1 + b_2 \cdot Z_1 + b_3 \cdot Z_2 + b_4 \cdot X_1 \cdot Z_1 + b_5 \cdot X_1 \cdot Z_2 + b_6 \cdot Edad + u$

Con el respectivo coeficiente de determinación $R^2_{\text{Modelo completo 1}}$

Si queremos por ejemplo comprobar si el efecto del factor *Sexo* es o no significativo tendríamos primero que estimar el modelo siguiente:

Modelo restringido 1: $Y=b_0 + b_1 \cdot Z_1 + b_2 \cdot Z_2 + b_3 \cdot X_1 \cdot Z_1 + b_4 \cdot X_1 \cdot Z_2 + b_5 \cdot Edad + u$

Con el respectivo coeficiente de determinación $R^2_{\text{Modelo restringido 1}}$

Seguidamente, con ambos coeficientes de determinación, calculamos el estadístico F, cuya prueba tendrá como hipótesis nula la no influencia del tratamiento, como sigue:

$$F = \frac{R^2_{Modelocompleto1} - R^2_{Modelorestringido1}}{1 - R^2_{Modelocompleto1}} * \frac{N - K_{c1} - 1}{K_{c1} - K_r}$$

Donde *N* es el número de casos, K_{c1} es el número de regresores en el modelo completo (en nuestro caso X_1, Z_1, Z_2, $X_1 \cdot Z_1$, $X_1 \cdot Z_2$ y Edad, por tanto es igual a 6) y K_r en el modelo restringido (igual a 5).

Los grados de libertad del estadístico F son (K_{c1}-K_r) y (N-K_{c1}-1). Si el estadístico calculado es inferior al valor crítico de F, esto es, su probabilidad es inferior a 0,05, rechazamos la hipótesis nula de no efecto de la variable de grupo sobre la variable dependiente, y por tanto aceptamos que el factor, *Sexo* en el modelo restringido 1, sí tiene un efecto estadísticamente significativo sobre el gasto en bienes de lujo.

El análisis de la covarianza asume la homogeneidad de la covarianza entre grupos. Para comprobar esta hipótesis (H_0: el efecto covariable*factor no es significativo) comparamos el modelo completo anterior con otro que incluya las interacciones de la covariable con las variables de grupo, en este caso:

Modelo completo 2: $Y=b_0 + b_1 \cdot X_1 + b_2 \cdot Z_1 + b_3 \cdot Z_2 + b_4 \cdot X_1 \cdot Z_1 + b_5 \cdot X_1 \cdot Z_2 + b_6 \cdot Edad + b_7 \cdot Edad \cdot X_1 + b_8 \cdot Edad \cdot Z_1 + b_9 \cdot Edad \cdot Z_2 + u$

Con el estadístico F calculado como: $F = \frac{R^2_{Modelocompleto2} - R^2_{Modelocompleto1}}{1 - R^2_{Modelocompleto2}} * \frac{N - K_{c2} - 1}{K_{c2} - K_{c1}}$

No es necesario calcular este estadístico ya que OS4 lo proporciona al principio del análisis de la covarianza, así como el correspondiente a la prueba de igualdad de pendientes. Tanto en un caso como en el otro, si rechazamos la hipótesis nula (homogeneidad de la covarianza o igualdad de pendientes) no podemos aplicar el análisis de la covarianza.

10.3. ANCOVA frente ANOVA

Para comprender la utilidad de esta técnica veamos un ejemplo donde hemos recogido datos sobre la producción por hectárea de un conjunto de fincas con distintos niveles de abonado y dosis de riego. En el primer grupo de fincas no hay riego (dosis=0), en el segundo se han utilizados 1.000 m^3/ha (dosis=1), por último, en el tercer grupo de fincas la cantidad utilizada fue de 4.000 m^3/ha (dosis=2). Los datos aparecen en la tabla siguiente:

Tabla 10.1. *Ejemplo de análisis de la covarianza*

Rendimiento	Abonado	Dosis
51	31	0
49	49	0
60	60	0
60	71	0
30	10	0
48	19	1
40	20	1
60	40	1
80	60	1
60	50	1
55	66	2
60	50	2
66	66	2
60	50	2
60	60	2

Si no dispusiéramos de la información sobre la cantidad de abono utilizado en cada finca tendríamos que explicar el rendimiento en función de la dosis de riego. Para ello aplicaríamos un análisis de la varianza donde el factor sería la dosis de riego y la variable dependiente el rendimiento.

Analyses → Analyses of Variance → 1, 2 or 3 way ANOVA → Dependent Variable: *Rendimiento*; Factor 1 Variable: *Dosis.*

Los resultados aparecen a continuación:

```
ONE WAY ANALYSIS OF VARIANCE RESULTS

Dependent variable is: rendimiento, Independent variable is: dosis
-------------------------------------------------------------------------
SOURCE     D.F.    SS         MS      F       PROB.>F   OMEGA SQR.
-------------------------------------------------------------------------
  BETWEEN       2    280.93  140.47    1.07      0.37     0.01
   WITHIN      12   1578.00  131.50
    TOTAL      14   1858.93
-------------------------------------------------------------------------

MEANS AND VARIABILITY OF THE DEPENDENT VARIABLE FOR LEVELS OF THE
INDEPENDENT VARIABLE
-------------------------------------------------------------------------
    GROUP    MEAN  VARIANCE      STD.        N
                                  DEV.
-------------------------------------------------------------------------
        1   50.00    150.50   12.27        5
        2   57.60    228.80   15.13        5
        3   60.20     15.20    3.90        5
-------------------------------------------------------------------------
    TOTAL   55.93    132.78   11.52       15
-------------------------------------------------------------------------

TESTS FOR HOMOGENEITY OF VARIANCE
-------------------------------------------------------------------------
Hartley Fmax test statistic =15.05 with deg.s freedom: 3 and 4.
Cochran C statistic =0.58 with deg.s freedom: 3 and 4.
Bartlett Chi-square =5.29 with 2 D.F. Prob. > Chi-Square =0.071
```

Según los resultados de ANOVA no podemos rechazar la hipótesis nula de no diferencia entre grupos, esto es, no relación entre la dosis de riego y el rendimiento (probabilidad de independencia igual 0,37). Podemos comprobar que la prueba es correcta ya que no se rechaza tampoco la hipótesis nula de homogeneidad de la varianza, requisito necesario en el análisis de la varianza (prob=0,071). Siendo evidente que la dosis de riego sí tendría que haber resultado con un efecto significativo sobre el rendimiento, ¿por qué ANOVA no lo ha detectado? La respuesta hay que buscarla en el efecto enmascarador de la cantidad de abono utilizado. Utilizando el análisis de la covarianza podemos discernir entre ambos efectos:

Analyses → Analyses of Variance → Analysis of Covariance; Dependent Variable: *Rendimiento*; Fixed Factors: *Dosis*; Covariates: *Abonado.*

Centrándonos en la descomposición de la varianza de la variable dependiente (`Test for Each Source of Variance`) tenemos:

```
Test for Each Source of Variance
     SOURCE     Deg.F.         SS         MS          F    Prob>F
          A          2     312.90     156.45      4.067    0.0476
 Covariates          1    1122.90    1122.90     29.192    0.0002
      Error         11     423.13      38.47
      Total         14    1858.93
```

Por lo que podemos afirmar que tanto la dosis de riego como el abonado tienen un efecto estadísticamente significativo sobre el rendimiento[35]. En este caso el análisis de la covarianza ha podido diferenciar ambos efectos.

10.4. Ejemplo ANCOVA

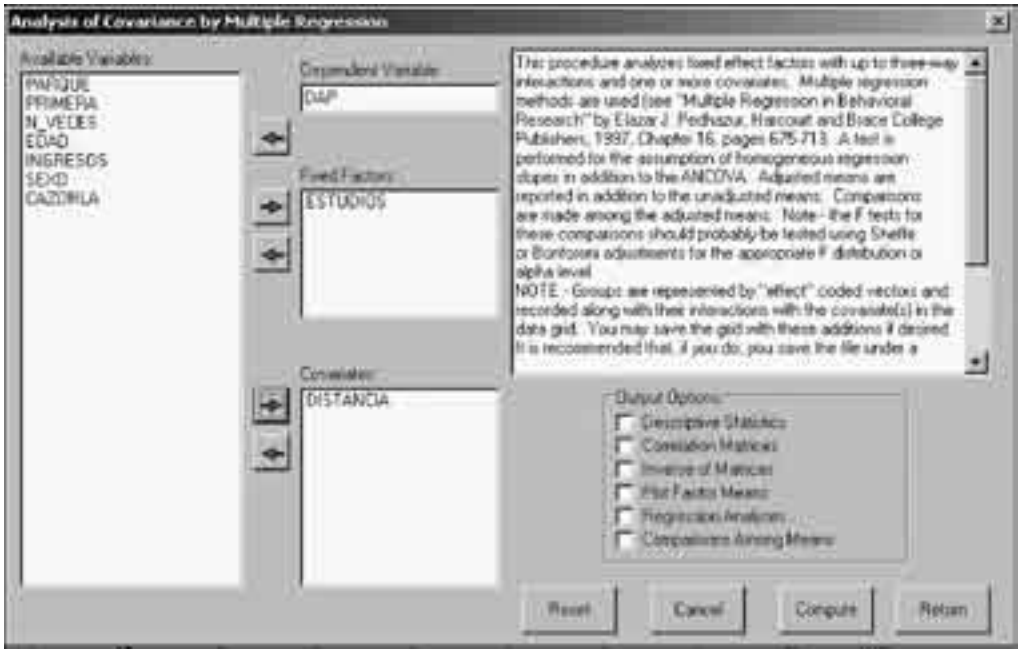
Volviendo a los datos de las encuestas realizadas en cinco parques, imaginemos que queremos determinar el efecto del nivel de estudios (ESTUDIOS) sobre la disposición a pagar (DAP). Suponemos que la distancia recorrida para visitar el parque influye sobre la disposición a pagar (coeficiente de correlación igual a 0,77) y, a su vez, puede interferir en la relación entre DAP y ESTUDIOS. Para aislar y cuantificar estos efectos utilizaremos el análisis de la covarianza, si bien en primer lugar hay que comprobar que la covariable es independiente del factor. Para ello:

Analyses → Analyses of Variance →1, 2 or 3 way ANOVA → Dependent Variable: *Distancia*; Factor 1 Variable: *Estudios*.

Los resultados muestran un valor del estadístico F igual a 1,36, con una probabilidad asociada de 0,26, por lo que no rechazamos la hipótesis nula de independencia entre ambas variables y podemos proceder con el análisis de la covarianza.

Analyses → Analyses of Variance → Analysis of Covariance; Dependent Variable: *Dap*; Fixed Factors: *Estudios*; Covariates: *Distancia*.

Figura 10.1. *Cuadro de diálogo del análisis de la covarianza (ANCOVA)*



Con el siguiente resultado:

```
ANALYSIS OF COVARIANCE USING MULTIPLE REGRESSION

Model for Testing Assumption of Zero Interactions with Covariates

Analysis of Variance for the Model to Test Regression Homogeneity
      SOURCE     Deg.F.         SS         MS          F    Prob>F
   Explained          7     235.57      33.65     44.699    0.0000
       Error        192     144.55       0.75
       Total        199     380.12

Model for Analysis of Covariance

Test for Homogeneity of Group Regression Coefficients
Change in R2 =0.0152. F =2.557 Prob.> F =0.0565 with d.f. 3 and 192

Analysis of Variance for the ANCOVA Model
      SOURCE     Deg.F.         SS         MS          F    Prob>F
   Explained          4     229.79      57.45     74.519    0.0000
       Error        195     150.33       0.77
       Total        199     380.12

Unadjusted Group Means for Group Variables ESTUDIOS

Means with 200 valid cases.

Variables
             1.426      0.901      1.557      0.794

Intercepts for Each Group Regression Equation for Variable: ESTUDIOS

Inercepts with 200 valid cases.

Variables       Group 1    Group 2    Group 3    Group 4
                  0.557      0.365      0.716      0.148

Adjusted Group Means for Group Variables ESTUDIOS

Means with 200 valid cases.

Variables       Group 1    Group 2    Group 3    Group 4
                  1.270      1.077      1.428      0.861
```

En primer lugar OS4 comprueba la relación entre la variable dependiente y la covariable. Como sabemos, el análisis de la covarianza se basa en un alto grado de correlación entre ambas. Así, con un valor de F igual a 44,699 y una probabilidad de 0,0000 rechazamos la hipótesis nula de cero correlación entre ambas variables.

Respecto a los supuestos del análisis de la covarianza, OS4 analiza la homogeneidad de la covarianza y la igualdad de las pendientes entre grupos (*Test for Homogeneity of Group Regression Coefficients*). En el primer caso, el incremento del coeficiente de determinación (0,0152) que se produce al incluir en el modelo completo las interacciones entre la covariable (DISTANCIA) y la variable de grupo (ESTUDIOS) es ínfimo, por lo que parece razonable descartar del modelo dichas interacciones y aceptar la homogeneidad de la covarianza. En la segunda prueba, el valor del estadístico F igual a 2,557, con una probabilidad asociada de 0,0565, no nos hace rechazar la hipótesis nula de igualdad de las pendientes.

A continuación tenemos el análisis de la varianza para el modelo ANCOVA (*Analysis of Variance for the ANCOVA Model*) donde podemos ver que dicho modelo explica el 60,5% (229,79/380,12) de la variabilidad de la variable dependiente. No rechazamos el modelo global ya que la probabilidad de que los coeficientes del modelo sean cero es igual a 0,0000.

También se incluyen los puntos de intersección de las regresiones $DAP=b_0 + b_1 \cdot DISTANCIA$ con el eje de ordenadas (Adjusted Group Means for Group Variables ESTUDIOS). En este caso existe un coeficiente b0 por cada uno de los grupos de ESTUDIOS (1,27, 1,08, 1,43 y 0,86).

OS4 continua con la comparación entre grupos de ESTUDIOS y la composición de la varianza del modelo ANCOVA.

```
Multiple Comparisons Among Group Means

Comparison of Group 1 with Group 2
F =1.162, probability =0.282 with degrees of freedom 1 and 195

Comparison of Group 1 with Group 3
F =0.769, probability =0.382 with degrees of freedom 1 and 195

Comparison of Group 1 with Group 4
F =3.276, probability =0.072 with degrees of freedom 1 and 195

Comparison of Group 2 with Group 3
F =5.427, probability =0.021 with degrees of freedom 1 and 195

Comparison of Group 2 with Group 4
F =1.278, probability =0.260 with degrees of freedom 1 and 195

Comparison of Group 3 with Group 4
F =8.056, probability =0.005 with degrees of freedom 1 and 195

Test for Each Source of Variance
     SOURCE      Deg.F.          SS          MS          F   Prob>F
          A           3        7.49        2.50      3.238   0.0233
 Covariates           1      222.30      222.30    288.364   0.0000
      Error         195      150.33        0.77
      Total         199      380.12
```

Como los resultados anteriores muestran, tenemos sólo dos diferencias significativas, las correspondientes a las comparaciones del grupo 2 con el 3 y del grupo 3 con el 4. El análisis termina con el estudio de la contribución del factor y de la covariable a la explicación de la variabilidad total (Test for Each Source of Variance). En primer lugar vemos cómo ambas variables ("A" es el factor y "Covariate" la covariable) influyen en la variable dependiente DAP (ESTUDIOS con F=3,238→Prob>F=0,0233; y DISTANCIA con F=288,364→Prob>F=0,0000). En este caso, si bien la influencia de la variable ESTUDIOS sobre la DAP es estadísticamente significativa, la importancia de este efecto queda minimizada por la magnitud de la relación DAP-DISTANCIA, como lo demuestra la diferencia de los valores de las sumas de los cuadrados (columna SS). Cuantificando estas diferencias tenemos:

- Variabilidad explicada por el factor: 7,49 / 380,12= 2%
- Variabilidad explicada por la covariable: 222,30 / 380,12=58%
- Variabilidad no explicada por el modelo 150,33 / 380,12=40%

Como OS4 genera nuevas variables para llevar a cabo el análisis de la covarianza procedemos a recuperar el fichero original y utilizamos el botón Save and Return en el menú Variables→Define para volver a las variables iniciales y descartar las variables auxiliares.

El mal investigador usa la estadística como un borracho una farola: más como apoyo que como iluminación
(Andrew Lang)

CAPÍTULO 11

MODELO LINEAL GENERAL

Capítulo 11. Modelo lineal general

11.1. Introducción teórica

El desarrollo de esta técnica tiene por origen la teoría de los invariantes algebraicos de principios del siglo XIX. En síntesis, esta teoría cuantifica aquellas relaciones entre variables que permanecen inalteradas tras aplicar transformaciones lineales a dichas variables. Por ejemplo, si *X* está correlacionada con *Y*, esta relación se mantiene entre *X* y *Z* si $Z=a\cdot Y+b$, siendo a y b dos constantes cualesquiera.

En el modelo lineal general el vector *Y* con *n* observaciones de una única variable dependiente del análisis de regresión se substituye por una matriz con *n* observaciones por *m* variables. De igual forma, el vector de coeficientes *b* de las variables que explican la única variable dependiente se transforma en una matriz donde cada vector de coeficientes se corresponde con una de las *m* variables dependientes[36].

¿Cuándo es útil esta técnica? Existen dos circunstancias claras donde el modelo lineal general prueba su flexibilidad y sus posibilidades. El primer caso aparece en el análisis de la varianza ANOVA 2,3,..n con problemas de cumplimiento de los requisitos paramétricos y/o diseño desequilibrado de datos. Se excluye ANOVA 1 ya que en este caso no tiene sentido hablar de diferente número de casos por factor (hay sólo uno) y ante problemas relacionados con los requisitos paramétricos donde existe la alternativa de análisis de la varianza mediante la prueba no paramétrica de Kruskal-Wallis.

El segundo caso hace referencia a la limitación del análisis de regresión múltiple frente al problema de la multicolinealidad (las variables independientes están correlacionadas). En este caso, el modelo lineal general permite abordar el análisis obviando este limitante, así como la posibilidad de analizar más de una variable dependiente de forma simultánea. En este sentido, en lugar de plantear dos modelos de regresión con dos variables dependientes y las mismas variables independientes sería mejor incluir ambas variables dependientes en un modelo lineal general, explotando así la posible correlación entre las variables que tratamos de explicar.

A la hora de implementar el modelo lineal general debemos considerar el número de variables dependientes. Si analizamos sólo una seguiremos el procedimiento de regresión múltiple. En el caso de analizar más de una variable dependiente utilizaremos el análisis de correlación canónico[37].

En resumen, con el modelo lineal general podemos hacer todo lo que hemos visto en los capítulos anteriores ya que nos permite estudiar el efecto de un conjunto de variables de cualquier naturaleza sobre otro conjunto de variables también de cualquier naturaleza. En la práctica, el uso de esta técnica se reduce al estudio de una variable métrica dependiente y dos o tres nominales independientes.

11.2. Relación entre las distintas técnicas estadísticas

La siguiente tabla resume las diferentes técnicas estadísticas analizadas hasta ahora en relación con la naturaleza de las variables dependiente(s) e independientes.

Tabla 11.1. *Comparación de técnicas estadísticas según el tipo de variables*

	ANOVA/ MANOVA	ANCOVA	Análisis discriminante	Regresión lineal múltiple	Regresión logística	Modelo Lineal General
Naturaleza de la variable dependiente (Y)	Métrica	Métrica	Nominal	Métrica	Dicotómica	Cualquiera, normalmente 1 o más métricas
Naturaleza de las variables explicativas (X)	Nominal	Nominal y métrica	Cualquiera	Cualquiera	Cualquiera	Cualquiera, normalmente nominales

Fuente: Elaboración propia y Malhotra y Birks (1999, p. 552).

La mayoría de las técnicas estadísticas son casos especiales del modelo lineal general. Así, la prueba t es un caso especial de ANOVA 1, ésta un caso especial de la regresión múltiple y éste del modelo lineal general.

11.3. Comparación ANOVA, regresión y MLG

Podemos entender las posibilidades del modelo lineal general comparándolo con dos técnicas más simples: el análisis de la varianza y el modelo de regresión lineal múltiple, que no son más que dos casos particulares del primero. Comenzaremos analizando el posible efecto del esfuerzo del estudiante y de su capacidad intelectual sobre el resultado de un examen. Ambos factores fijos están codificados como variables dicotómicas (también conocidas como ficticias o *dummy*) –ver el apartado *Enfoque MLG* para aclarar este punto- con el valor 1 para aquellos estudiantes que se esfuerzan mucho y 0 para aquellos que no lo hacen. De igual forma la variable que hace referencia a su capacidad toma el valor 1 en el caso de ser un estudiante con un rendimiento escolar por encima de la media y 0 en caso de estar por debajo. La tabla siguiente presenta los datos obtenidos en una clase de 30 alumnos.

Tabla 11.2. *Ejemplo de resultados de un examen según esfuerzo y capacidad del estudiante*

Estud.	Nota	Esfuerzo	Capacidad
1	100	1	1
2	100	1	1
3	100	1	1
4	100	1	1
5	98	1	1
6	97	1	1
7	93	1	0
8	83	1	0
9	82	1	0
10	81	1	0
11	77	0	1
12	77	0	1
13	76	0	1
14	72	0	1
15	70	1	0

Estud.	Nota	Esfuerzo	Capacidad
16	65	1	1
17	62	1	0
18	61	1	1
19	60	1	0
20	55	1	0
21	52	1	0
22	52	1	0
23	51	1	0
24	50	1	0
25	42	0	1
26	41	0	0
27	41	0	0
28	40	0	0
29	36	0	0
30	25	0	0

Enfoque ANOVA

Comenzaremos con ANOVA II: Analyses → Analyses of Variance → 1, 2 or 3 way ANOVA → Dependent Variable: *Nota*; Factor 1 Variable: *Esfuerzo*; Factor 2 Variable: *Capacidad*.

```
Two Way Analysis of Variance

Variable analyzed: nota
Factor A (rows) variable: esfuerzo (Fixed Levels)
Factor B (columns) variable: capacidad (Fixed Levels)
```

SOURCE	D.F.	SS	MS	F	PROB.>F	Omega Squared
Among Rows	1	3496.067	3496.067	16.117	0.000	0.222
AmongColumns	1	4468.514	4468.514	20.600	0.000	0.288
Interaction	1	936.594	936.594	4.318	0.048	0.049
Within Groups	26	5639.792	216.915			
Total	29	14540.967	501.413			

```
Omega squared for combined effects = 0.559

Note: Denominator of F ratio is MSErr

Descriptive Statistics
```

GROUP	Row	Col.	N	MEAN	VARIANCE	STD.DEV.
Cell	0	0	5	36.600	46.300	6.804
Cell	0	1	5	68.800	228.700	15.123
Cell	1	0	12	65.917	232.811	15.258
Cell	1	1	8	90.125	282.696	16.814
Row	0		10	52.700	410.233	20.254
Row	1		20	75.600	386.989	19.672
Col	0		17	57.294	361.221	19.006
Col	1		13	81.923	357.744	18.914
TOTAL			30	67.967	501.413	22.392

```
TESTS FOR HOMOGENEITY OF VARIANCE
-------------------------------------------------------------------------
Hartley Fmax test statistic = 6.11 with deg.s freedom: 4 and 4.
Cochran C statistic = 0.36 with deg.s freedom: 4 and 4.
Bartlett Chi-square statistic = 7.22 with 3 D.F.
-------------------------------------------------------------------------
```

Vemos cómo ambas variables tienen un efecto estadísticamente discernible sobre NOTA (con una $F_{esfuerzo}$=16,12, y $F_{capacidad}$=20,60, ambas con PROB.>F=0,000 → un 0% de probabilidad de no relación con la variable dependiente). El efecto combinado de ambas (interacción) es estadísticamente significativo al 5% ($F_{interacción}$=4,32, Prob=0,048) y además muy importante (0,56), atribuyéndose casi por igual a ambos factores (0,22 y 0,29, respectivamente).

En la segunda parte podemos comprobar cómo el estadístico de Hartley, cuya hipótesis nula es la homocedasticidad, tiene un valor de 6,11, inferior al valor crítico del estadístico $F_{4, 4\ 0,05}$ =6,39 (para obtenerlo: Simulation → Distribution Plots and Critical Values → ⊙Central F distribution; Deg.Freedom (1): *4* Deg.Freedom (2): *4*). De

igual manera, el estadístico de Barlett con un valor de 7,22 no excede el valor crítico Chi $_{3\ 0,05}$=7,78, por lo que también se llega a la misma conclusión. Ambas pruebas implican que no rechazamos la hipótesis nula de varianzas homogéneas (requisito de ANOVA) por lo que nuestro análisis es correcto.

Enfoque regresivo

Planteamos el siguiente modelo lineal de regresión múltiple: nota=b_0 + b_1·esfuerzo + b_2·capacidad.

Analyses → Regression → Block Entry: Dependent Variable: *Nota*; Ind. Var's to enter in block: *Esfuerzo, Capacidad*.

```
Block Entry Multiple Regression by Bill Miller

--------------- Trial Block 1 Variables Added ----------------
        R      R2         F    Prob.>F     DF1        DF2
    0.778   0.605    20.671      0.000       2         27
Adjusted R Squared =0.576

Std. Error of Estimate = 14.587

Variable      Beta          B  Std.Error       t     Prob.>t
  esfuerzo   0.548    25.595      5.675   4.510      0.000
 capacidad   0.606    26.945      5.399   4.991      0.000

Constant = 39.227
Increase in R Squared = 0.605
F =20.671 with probability = 0.000
Block 1 met entry requirements
```

La probabilidad de que nuestro modelo sea incorrecto es igual a 0,000 (`Prob.>F =0,000`), con una capacidad explicativa del 57,6%. Ambas variables ficticias son significativas (`Prob.>t =0,000`) y tienen un efecto similar (Nota =39,2 + 25,6·esfuerzo + 26,9·capacidad).

Es decir, un estudiante que no se esfuerza mucho (ESFUERZO=0) y no tiene gran capacidad (CAPACIDAD=0) tendría una nota estimada de 39 puntos. Si se esfuerza (ESFUERZO=1) pero no tiene gran capacidad (CAPACIDAD=0) su nota estimada sube hasta 39,2+25,6=64,8. Por el contrario, si no se esfuerza (ESFUERZO=0) pero tiene gran capacidad (CAPACIDAD=1) su nota estimada es 39,2+26,9=66,1. Finalmente, el mejor estudiante (trabaja y es capaz) obtiene una puntuación estimada de 39,2+25,6+26,9=91,7.

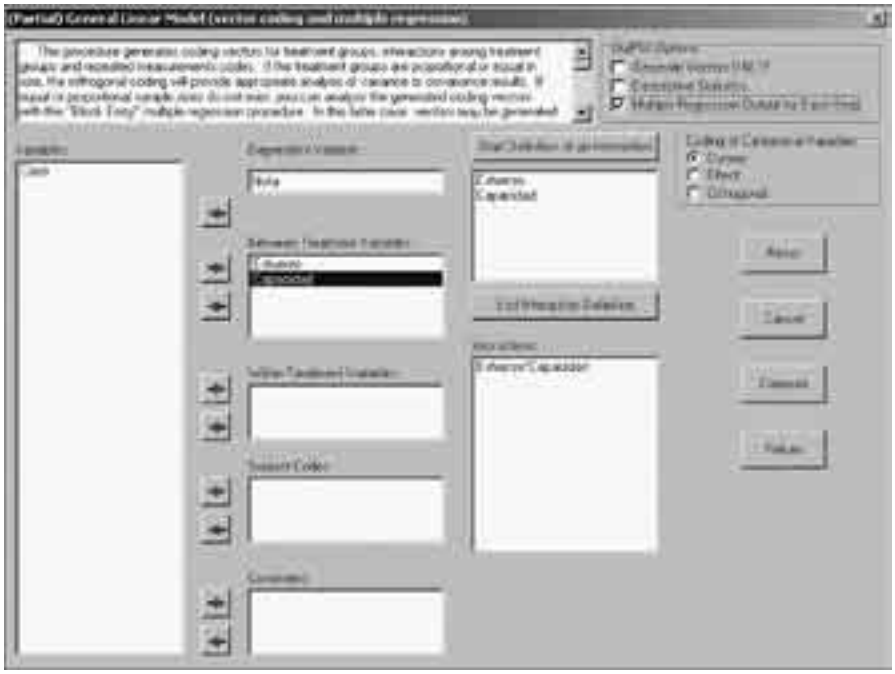
Mientras que con ANOVA II hemos podido determinar que los factores ESFUERZO y CAPACIDAD afectan a la nota alcanzada y que el primero tiene un efecto ligeramente inferior que el segundo ($\text{Omega Squared}_{esfuerzo}=0,222 < \text{Omega Squared}_{capacidad}=0,288$), con el análisis de regresión hemos llegado a la misma conclusión planteando un modelo lineal ($\text{beta}_{esfuerzo}=0,548 < \text{beta}_{capacidad}=0,606$).

Enfoque MLG

Finalmente, podemos analizar el efecto de las dos variables dicotómicas sobre la nota mediante el modelo lineal general:

Analyses → Multivariate → Sums of Squares by Regression → Dependent Variable: *Nota*; Between Treatment Variables: *Esfuerzo, Capacidad.*:

Figura 11.1. *Cuadro de diálogo del Modelo Lineal General*



En el cuadro de diálogo del modelo lineal general deben tenerse en cuenta dos aspectos:

A. Formato de las variables categóricas

Tabla 11.3. *Tipos de codificación de las variables categóricas (variables tratamiento o factores)*

Codif.	Caso	Código A	Código B	Descripción	Grupo
Efecto (*effect coding*)	1	1	0	Recibió el tratamiento A pero no el B	Grupo 1
	2	0	1	Recibió el tratamiento B pero no el A	Grupo 2
	3	-1	-1	No recibió ningún tratamiento	Grupo 3
Ortogonal (*orthogonal coding*)	1	1	1	Recibió el tratamiento A pero no el B	Grupo 1
	2	-1	1	Recibió el tratamiento B pero no el A	Grupo 2
	3	0	-2	No recibió ningún tratamiento	Grupo 3
Ficticia (*dummy coding*)	1	1	0	Recibió el tratamiento A pero no el B	Grupo 1
	2	0	1	Recibió el tratamiento B pero no el A	Grupo 2
	3	0	0	No recibió ningún tratamiento	Grupo 3

En nuestro caso la codificación seguida es la tercera (la primera en el cuadro de diálogo del modelo lineal general), esto es, la codificación ficticia. Esta codificación es la más habitual tanto en diseños experimentales como en estudios de tipo sociológico.

B. El tipo de efecto de la variable independiente sobre la variable dependiente:

Dentro del diseño del modelo lineal que analiza el efecto de uno o varios factores o tratamientos sobre una variable dependiente podemos clasificar dichos efectos en dos categorías[38]:

- Factores de efectos fijos (conocidos en los paquetes estadísticos como *Fixed effects o Between Treatment Variables*). En estos casos, el investigador está interesado en comparar los valores medios de los diferentes grupos (en ciencias experimentales los que recibieron el tratamiento y los que no lo hicieron / en ciencias sociales, los que presentan la característica del factor frente a los que no la tienen). Son factores cuyos niveles recogen todos los valores posibles (por ejemplo, en el caso del factor sexo: hombre y mujer) o factores cuyos niveles, si bien no cubren todo el universo de valores posibles, cuyos efectos no se quieren extrapolar al resto de valores de los niveles no incluidos en la investigación. Un factor puede ser de efecto fijo o aleatorio dependiendo de la inferencia de los resultados, así, por ejemplo, se estudia el factor ciudad en los ingresos medios de los ciudadanos en cinco ciudades, si no extrapolamos los resultados al resto de ciudades del país, el factor ciudad es fijo, si se extrapola, el efecto es aleatorio.
- Factores de efectos aleatorios (*Random effects o Within Treatment Variables*). Los niveles del factor (o factores) utilizados en el experimento representan una muestra del total de posibles valores. En estos casos, el investigador centra su atención en el efecto del factor aleatorio A sobre la variabilidad de la variable dependiente Y para cada uno de los niveles del factor fijo B. Por ejemplo, si estudiamos el efecto de la posesión de un título universitario (factor fijo) sobre el nivel de ingresos (variable dependiente) en cinco ciudades (factor aleatorio) para su extrapolación al resto del país, el objetivo del estudio consiste en determinar el efecto de la posesión del título sobre los ingresos una vez separado el efecto aleatorio del factor ciudad. De igual forma, en experimentación, la repetición del experimento con los mismos niveles de los factores en otras parcelas -o cualquier otra unidad de experimentación- trata de aislar el efecto aleatorio de la respuesta individual de cada parcela, independientemente de que reciba el tratamiento o no, del efecto del tratamiento sobre la variable dependiente. En este caso, el factor aleatorio sería la parcela (o bloque de ensayo) ya que sólo consideramos en el experimento algunas de las numerosas parcelas posibles, si bien se pretende extrapolar el efecto del factor fijo (por ejemplo, el herbicida A) sobre la variable dependiente (por ejemplo, el rendimiento).

La siguiente tabla resume las características principales de los factores de efectos fijos y de efectos aleatorios.

Tabla 11.4. *Diferenciación de factores de efectos fijos y de efectos aleatorios*

Característica	Factor de efecto fijo	Factor de efecto aleatorio
Objetivo principal	Diferencias entre las medias de los diferentes grupos	Varianza de la variable dependiente con respecto a los niveles del factor fijo una vez separado el efecto del factor aleatorio
Número de niveles del factor	Todos los posibles (o sólo aquellos para los cuales queremos obtener resultados)	Una muestra de los valores posibles de la población de niveles

Continuando con nuestro ejemplo, y como apuntamos anteriormente, los factores ESFUERZO y CAPACIDAD son factores de efectos fijos por contener todos los niveles posibles (esfuerzo y capacidad por encima o por debajo de la media). En OS4 los factores de efectos fijos se incluyen en la caja `Between Treatment Variables`, los factores de efectos aleatorios en `Within Treatment Variables` y las covariables, aquellas variables de naturaleza métrica (a diferencia de los factores, de naturaleza categórica) que están correlacionadas con la variable dependiente, en la caja `Covariates`.

Además del los efectos individuales de los factores, hemos decidido analizar la posible interacción de los dos factores, es decir, que la nota estimada de un estudiante trabajador y capaz es mayor que la suma de la nota estimada de estudiante trabajador más la nota estimada de un estudiante capaz. Para introducir esta interacción en el cuadro de diálogo, una vez los factores están en su caja correspondiente, pulsamos el botón "`Start Definition of an Interaction`", y después sobre las variables ESFUERZO y CAPACIDAD consecutivamente. Para finalizar la definición de la interacción pulsamos "`End Definition of an Interaction`". La interacción aparece en el cuadro "`Interactions`" como "ESFUERZO*CAPACIDAD". En el caso de que hubiera más variables se podrían incluir otras interacciones (incluidas las interacciones entre tres variables), para lo cual bastaría con repetir el procedimiento con cada bloque de nuevas variables.

Con objeto de estudiar cada una de los modelos lineales alternativos marcamos la opción `Multiple Regression Output for Each Step`. A continuación pulsamos Compute para obtener los resultados del modelo lineal general.

1. En primer lugar OS4 nos muestra los resultados del modelo completo, el cual incluye los dos factores y la intereacción entre los mismos:

```
Sum of Squares Total = 14540.9667

Dependent variable: Nota

  Variable     Beta          B  Std.Err.        t   Prob.>t     VIF     TOL
 Esfuerzo1   -0.457    -21.325     8.396   -2.540     0.017   2.167   0.462
Capacidad1   -0.545    -24.208     6.722   -3.601     0.001   1.535   0.652
      C1E1    0.135      7.992    11.487    0.696     0.493   2.535   0.395
 Intercept    0.000     90.125     5.207   17.308     0.000

SOURCE           DF         SS        MS        F   Prob.>F
Regression        3   8901.175  2967.058   13.678    0.0000
Residual         26   5639.792   216.915
Total            29  14540.967

R2 = 0.6121, F = 13.68, D.F. = 3 26, Prob>F = 0.0000
Adjusted R2 = 0.5674

Standard Error of Estimate = 14.73
```

Nota: Es posible que, debido al tipo de codificación de los factores (efecto, ortogonal o ficticia), el signo de la matriz de correlaciones y de los coeficientes del modelo de regresión tenga un signo opuesto al real (tal y como ocurre en este caso, ya que si un alumno se esfuerza y es capaz tiene una nota superior, y no inferior como sugiere el resultado anterior). Por ello, una vez seleccionado el mejor modelo lineal (fijándonos en la columna `Prob.>t`) es aconsejable repetir el análisis de correlación y/o de regresión mediante los módulos correspondiente en OS4.

En el primer modelo de regresión con los dos factores (`Esfuerzo1 y Capacidad1`) y la interacción entre ambos (`C1E1`) comprobamos cómo la interacción no es estadísticamente significativa (la probabilidad de la hipótesis nula de un coeficiente igual a cero es igual a 49,3%, muy por encima del valor crítico del 5%). Este modelo tiene una bondad de ajuste del 61,21%. A continuación OS4 nos proporciona los modelos mixtos factor 1 + interacción y factor 2 + interacción.

```
Dependent variable: Nota

  Variable    Beta          B  Std.Err.        t  Prob.>t     VIF    TOL
Capacidad1   0.360     16.006     6.464    2.476    0.020   1.181  0.847
      C1E1   0.496     29.317     8.595    3.411    0.002   1.181  0.847
 Intercept   0.000     81.923     4.478   18.294    0.000

    SOURCE      DF         SS        MS       F  Prob.>F
Regression       2   7501.927  3750.963  14.388   0.0001
  Residual      27   7039.040   260.705
     Total      29  14540.967

R2 = 0.5159, F = 14.39, D.F. = 2 27, Prob>F = 0.0001
Adjusted R2 = 0.4801

Standard Error of Estimate = 16.15
```

Tanto el modelo global (F=14,388 con probabilidad=0,0001<0,005) como los coeficientes individuales (Prob.>t = 0,020 y 0,002, respectivamente) no son rechazados. La capacidad explicativa del modelo es del 51,59%. A continuación OS4 prueba el otro factor y la interacción entre ambos:

```
Dependent variable: Nota

   Variable   Beta          B  Std.Err.       t  Prob.>t    VIF    TOL
  Esfuerzo1  0.146      6.800     8.847   0.769    0.449  1.667  0.600
       C1E1  0.545     32.200    11.190   2.877    0.008  1.667  0.600
  Intercept  0.000     75.600     3.956  19.108    0.000

     SOURCE     DF         SS        MS       F  Prob.>F
 Regression      2   6088.167  3044.083   9.723   0.0007
   Residual     27   8452.800   313.067
      Total     29  14540.967

R2 = 0.4187, F = 9.72, D.F. = 2 27, Prob>F = 0.0007
Adjusted R2 = 0.3756

Standard Error of Estimate = 17.69
```

El resultado es un modelo que rechaza la variable ESFUERZO (como estadísticamente significativa (Prob.>t = 0,449, muy por encima de la probabilidad máxima admisible de 0,05) y una bondad de ajuste inferior (41,87%). Finalmente, OS4 muestra los resultados del modelo con sólo los dos factores sin la interacción entre ambos:

```
Dependent variable: Nota

    Variable   Beta          B  Std.Err.       t  Prob.>t    VIF    TOL
   Esfuerzo1  0.548     25.595     5.675   4.510    0.000  1.009  0.991
  Capacidad1  0.606     26.945     5.399   4.991    0.000  1.009  0.991
   Intercept  0.000     91.767     4.597  19.963    0.000

     SOURCE     DF         SS        MS       F  Prob.>F
  Regression     2   8796.189  4398.094  20.671   0.0000
    Residual    27   5744.778   212.770
       Total    29  14540.967

R2 = 0.6049, F = 20.67, D.F. = 2 27, Prob>F = 0.0000
Adjusted R2 = 0.5757

Standard Error of Estimate = 14.59
```

El modelo planteado se acepta globalmente (Prob.>F = 0.0000) así como cada uno de los factores de forma individual (Prob.>t = 0.0000 en ambos casos). El coeficiente de determinación nos indica una bondad de ajuste del modelo del 60,49%. La varianza no explicada por el modelo es el cociente 5744,78 / 14540,97 = 0,3951, valor que también se obtiene a partir de la varianza explicada: Varianza No Explicada = Varianza Total – Varianza Explicada = 1 – 0,6049 = 0,3951. Una vez analizados todos los modelos lineales posibles mediante el análisis de regresión, OS4 procede

a resumir la aportación de cada factor y de todas las interacciones (en nuestro caso sólo una, esfuerzo*capacidad) a la varianza de la variable dependiente, como vemos a continuación:

```
SUMS OF SQUARES AND MEAN SQUARES BY REGRESSION
TYPE III SS - R2 = Full Model - Restricted Model

      VARIABLE  SUM OF SQUARES      D.F.
     Esfuerzo1       1399.248         1
    Capacidad1       2813.008         1
          C1E1        104.986         1
         ERROR       5639.792        26
         TOTAL      14540.967        29
```

Así pues, vemos cómo el factor CAPACIDAD tiene el mayor poder explicativo (2813 / 14541 = 0,193) con algo más del 19% de la variabilidad. Le sigue el factor ESFUERZO, con un porcentaje cercano al 10% (1399 / 14541 = 0,096). La interacción entre los factores apenas supone un 1% de la varianza.

La última pantalla resume los modelos mixtos factor-interacción de factores ya descritos anteriormente, por lo que obviamos su comentario.

Suficiente tortura, y los datos confesarán cualquier cosa
(Gregg Easterbrook)

CAPÍTULO 12

ANÁLISIS FACTORIAL

Capítulo 12. Análisis factorial

12.1. Objeto del análisis factorial

El principal objetivo del análisis factorial es la reducción del número de variables mediante la obtención de factores que explican la variabilidad común de estas variables. Estos factores pueden considerarse como ejes o dimensiones que agrupan variables altamente correlacionadas. De esta forma, las variables iniciales se transforman en combinaciones lineales de dichos factores. Matemáticamente:

$X_1=a_{11}\cdot F_1 + a_{12}\cdot F_2 +... + a_{1m}\cdot F_m + u_1$

$X_2=a_{21}\cdot F_1 + a_{22}\cdot F_2 +... + a_{2m}\cdot F_m + u_2$

...

$X_i=a_{i1}\cdot F_1 + a_{i2}\cdot F_2 +... + a_{im}\cdot F_m + u_i$

De este modo, la varianza de la variable X_i se explica mediante la varianza común (*comunalidad*) con el resto de variables a través de los factores F_m más la varianza específica no común u_i (*unicidad*). Idealmente, cada variable se explica por un conjunto de factores y descarta el resto, lo cual se traduce en coeficientes a_{ij} (*factores de carga*) próximo a 1 en el primer caso y prácticamente 0 en el segundo. Teniendo en cuenta el proceso de creación de estos factores, obviamente, el análisis factorial tiene sentido sólo cuando existen variables que están correlacionadas.

Cuando las puntuaciones X_i están estandarizadas (media igual a 0 y varianza igual a 1) los coeficientes a_{ij} del modelo anterior representan el coeficiente de correlación de *Pearson* de la variable con el factor. El cuadrado de estos coeficientes mide el porcentaje de varianza de la variable explicada por el factor correspondiente. Por tanto, $a_{11}^2+ a_{12}^2 +... + a_{1m}^2$ representa el porcentaje de la varianza de la variable X_1 coincidente con los factores comunes F_1, F_2,...F_m, denominada *comunalidad*.

Teniendo en cuenta que el análisis factorial genera tantos factores como variables iniciales es necesario determinar cuáles deben retenerse en el modelo. Para ello podemos utilizar la regla de Kaiser, sugerida por Guttman y adaptada por Kaiser (Kaiser, 1958; Nunnally, 1978): Retener aquellos factores cuyo valor característico (*eigenvalue*) sea superior a 1. También es posible decidir el número de factores a partir del diagrama de valores característicos: Seleccionar el número de factores antes del principal cambio en la pendiente del gráfico. El criterio de Kaiser se recomienda cuando tenemos menos de 30 variables y la comunalidad media es superior a 0,70, o cuando el número de casos es mayor que 250 y con una comunalidad media mayor que 0,60 (Stevens, 1992).

La fiabilidad del análisis factorial está supeditada al tamaño de la muestra. Si bien no existe un consenso respecto al tamaño mínimo podemos apuntar las recomendaciones de algunos autores: Tabachnick y Fidell (1996) y Comrey y Lee (1992) recomiendan al menos una muestra con 300 casos; por otro lado, Guadagnoli y Velicer (1988) opinan

que si la mayoría de factores importantes tienen más de cuatro coeficientes mayores que 0,6 la muestra es representativa independientemente del tamaño; McCallum *et al.* (1999) afirman que una muestra en torno a 150 casos puede ser perfectamente válida si todos los valores comunes exceden el valor de 0,6; por último, Gorsuch (1983), Hatcher (1994) y Bryant y Yarnold (1995) sugieren un mínimo de cinco casos por cada variable y no menos de 100 casos para el análisis.

12.2. Análisis factorial y análisis de componentes principales

Al igual que en el análisis factorial, el principal objetivo del análisis de componentes principales es la reducción del número de variables explicativas. Sin embargo, mientras el análisis de componentes principales no está basado en ningún modelo estadístico, el análisis factorial parte de un modelo matemático específico (Manly, 1994, p. 93). En este sentido, el análisis de componentes principales tiene como elemento principal la transformación de los datos sin necesidad de satisfacer ningún supuesto matemático. Frente a esto, al análisis factorial asume que los datos provienen de un modelo específico donde los factores subyacentes responden a una seria de supuestos (Mardia *et al.*, 1989, p. 275).

Desde un punto de vista práctico, ambas técnicas representan dos alternativas estadísticas al mismo problema, las cuales se diferencian más por el procedimiento de cálculo que por los resultados que producen (Guadagnoli y Velicer, 1988). Así, mientras el análisis de componentes principales transforma los datos tomando como partida las variables iniciales para obtener los factores, el análisis factorial comienza con los factores y construye el modelo en función de las variables iniciales[39]. Sin embargo, siempre es posible obtener la transformación del espacio factorial a partir de la transformación de componentes principales y viceversa[40].

Una muestra de la interconexión de ambas técnicas queda reflejada por el hecho de que el método de extracción de factores más utilizado es precisamente el de componentes principales (Harman, 1976), el cual determina el mínimo número de factores que explican el máximo de la varianza de las variables para su posterior uso en subsiguientes análisis multivariantes, en este caso, la rotación de estos factores (Malhotra y Birks, 2000, p. 582). Así, una vez obtenidos los factores mediante el análisis de componentes principales se procede a obtener otros factores ortogonales –siendo el método varimax el más utilizado- u oblicuos que son combinaciones lineales de los iniciales y facilitan la interpretación de los mismos (Manly, 1994, p. 96).

12.3. Ejemplo de análisis factorial

Imaginemos una prueba psicotécnica con preguntas sobre los rasgos de la personalidad de 15 individuos. Si bien el tamaño de la muestra hace inviable el análisis factorial utilizaremos estos datos con el propósito de describir esta técnica. En la muestra, cada variable se mide en una escala de 0 a 100, como aparece en la tabla siguiente.

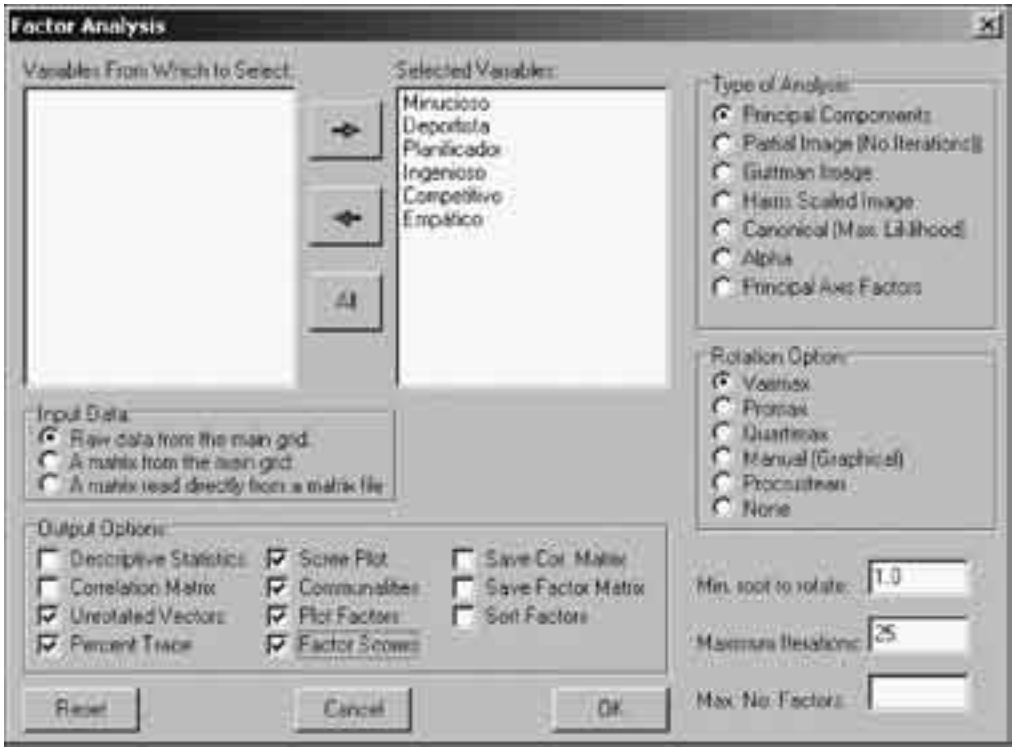
Tabla 12.1. *Ejemplo de una prueba psicotécnica para el análisis factorial*

Minucioso	Deportista	Planificador	Ingenioso	Competitivo	Empático
95	86	65	35	85	29
95	68	52	36	91	53
93	78	51	45	78	26
85	94	65	46	56	38
86	68	35	56	61	46
78	70	56	60	70	38
77	70	46	65	56	39
73	89	65	68	64	41
60	53	45	70	70	52
50	46	35	71	58	62
25	49	38	78	79	84
23	90	34	81	81	80
31	91	26	83	87	82
36	94	19	86	86	83
25	93	34	92	93	90

Utilizando los datos anteriores procedemos al análisis factorial de las variables:

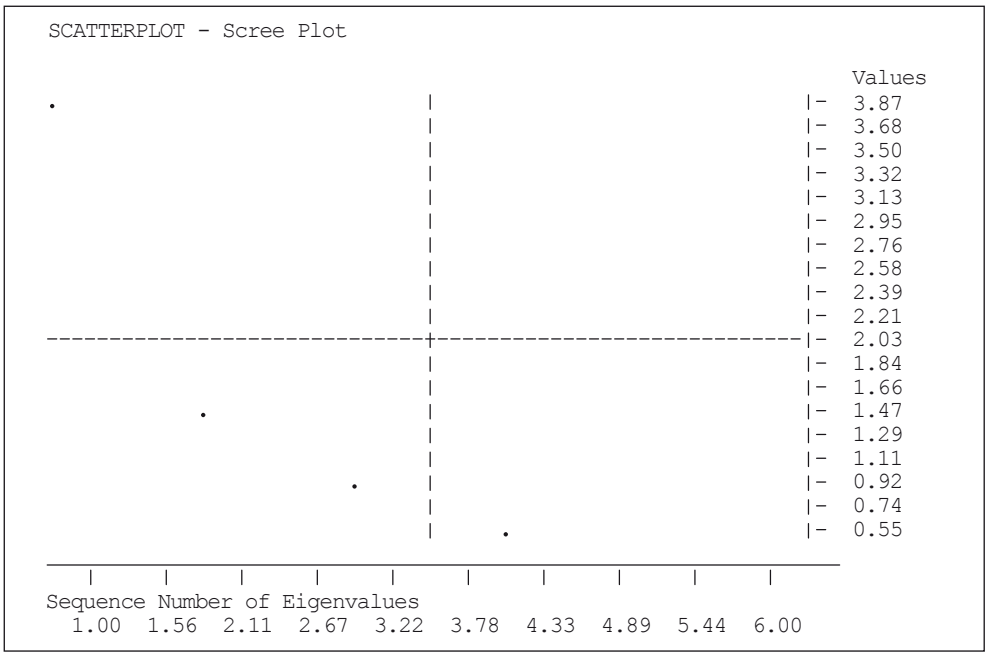
Analyses → Multivariate → Factor Analysis: *Todas las variables.*, mediante el método de las componentes principales (*Principal Components*) y la rotación *Varimax*, así como las siguientes opciones:

Figura 12.1. *Análisis factorial de los datos psicotécnicos*



En la casilla `Min. Root size to rotate` (valor mínimo del valor característico para rotar) el valor por defecto es 1. Así, el espacio factorial sólo retendrá aquellos factores con un valor característico (*eigenvalue*) superior a 1,00. Sin embargo puede ser interesante incluir un factor con un valor ligeramente inferior o incluso fijar el valor a partir de 0,70, como sugieren algunos autores. Para ellos podemos cambiar el valor por defecto: `Min. Root size to rotate` `0.70`

Figura 12.2. *Valores característicos del análisis factorial*



Esta primera pantalla presenta el diagrama de los valores característicos (*eigenvalues*), el cual nos servirá para determinar el número de factores que incluiremos en el modelo factorial. Como hemos visto anteriormente, siguiendo el criterio de Kaiser deberíamos incluir todos aquellos factores con un valor próximo a 1,00. En este caso sólo hay dos factores que superan este valor. También podríamos determinar en qué punto la curva se suaviza, lo que, en este caso, ocurre a partir del tercer valor característico. Así, al igual que antes, retendríamos dos factores.

```
Factor Analysis
Principal Components Analysis
Original matrix trace =6.00
Roots (Eigenvalues) Extracted:

            1  3.683
            2  1.241
            3  0.672
            4  0.289
            5  0.086
            6  0.028

Unrotated Factor Loadings
FACTORS with 15 cases.

Variables
              Factor 1 Factor 2 Factor 3 Factor 4 Factor 5  Factor 6
    MINUCIOSO    0.957    0.104   -0.104   -0.211   -0.039    -0.125
   DEPORTISTA   -0.124    0.858    0.485   -0.106    0.032     0.006
 PLANIFICADOR    0.862    0.190    0.091    0.459    0.014    -0.026
    INGENIOSO   -0.909   -0.178    0.301    0.107   -0.192    -0.065
  COMPETITIVO   -0.490    0.652   -0.567    0.068   -0.092     0.002
     EMPÁTICO   -0.970   -0.001   -0.079    0.079    0.196    -0.087

Percent of Trace In Each Root:
1 Root:= 3.683 Trace:= 6.000 Percent:= 61.390
2 Root:= 1.241 Trace:= 6.000 Percent:= 20.682
3 Root:= 0.672 Trace:= 6.000 Percent:= 11.208
4 Root:= 0.289 Trace:= 6.000 Percent:= 4.816
5 Root:= 0.086 Trace:= 6.000 Percent:= 1.437
6 Root:= 0.028 Trace:= 6.000 Percent:= 0.467
```

En primer lugar aparecen los valores característicos (*Eigenvalues*) de los factores extraídos. Sólo los dos primeros superan el valor 1,00 ($Factor_1$=3,68 y $Factor_2$=1,24), por lo que OS4 seleccionará estos dos factores para la rotación posterior. El método de las componentes principales considera la totalidad de la varianza (la común con los factores o *comunalidad* más la específica de las variables o *unicidad*) por lo que la suma de los valores característicos es igual a 6 (1 por cada variable), es decir 3,683 +1,241 +0,672 +0,289 +0,086 +0,028 =5,999. A partir de estos valores podemos determinar el porcentaje de la varianza total explicada por cada factor, así, el Factor1 explica el 61,39% (3,683/5,999), como aparece en la sección `Percent of Trace In Each Root`.

La sección `Unrotated Factor Loadings` (factores de carga no rotados) presenta el coeficiente de correlación de Pearson de cada par variable-factor. Vemos como el $Factor_1$ tiene una alta correlación con las variables MINUCIOSO, PLANIFICADOR, INGENIOSO y EMPÁTICO y baja con el resto. Por el contrario, el $Factor_2$ tiene una alta correlación con DEPORTISTA y COMPETITIVO y baja con el resto. Esta es la situación idónea ya que cada variable "cargará" mucho sobre un factor y poco sobre el otro, facilitando la interpretación de los factores.

A partir de los factores de carga del modelo factorial tenemos no sólo el coeficiente de correlación del par variable-factor, como hemos apuntado anteriormente, sino también el porcentaje de la varianza de la variable explicada por el factor. Así, el Factor1 consigue explicar el $0{,}957^2$=0,916 (o el 91,6%) de la varianza de la variable MINUCIOSO. La suma de los cuadrados de cada columna nos proporciona los valores característicos iniciales, los cuales indican la capacidad explicativa global de cada factor. Por ejemplo, $0{,}957^2 + 0{,}124^2 + 0{,}862^2 + 0{,}909^2 + 0{,}490^2 + 0{,}970^2 = 3{,}68$, que es el valor característico del primer factor. OS4 continúa el análisis con los dos factores seleccionados:

```
Proportion of variance in unrotated factors
1 61.390
2 20.682

Communality Estimates as percentages:
     1  92.751
     2  75.226
     3  77.985
     4  85.721
     5  66.572
     6  94.179

Variables
                  Factor 1     Factor 2
MINUCIOSO           -0.954       -0.129
DEPORTISTA          -0.085        0.863
PLANIFICADOR        -0.883       -0.022
INGENIOSO            0.925        0.045
COMPETITIVO          0.319        0.751
EMPÁTICO             0.942        0.232

Percent of Variation in Rotated Factors
Factor 1 59.047
Factor 2 23.025

Total Percent of Variance in Factors: 82.072
Communalities as Percentages
     1 for          MINUCIOSO   92.751
     2 for         DEPORTISTA   75.226
     3 for       PLANIFICADOR   77.985
     4 for          INGENIOSO   85.721
     5 for        COMPETITIVO   66.572
     6 for           EMPÁTICO   94.179
```

La segunda pantalla comienza con los dos factores seleccionados y, como vimos en la pantalla anterior, el porcentaje de la varianza total explicada por ambos (61,39 y 20,68), que en total supone el 82,07%.

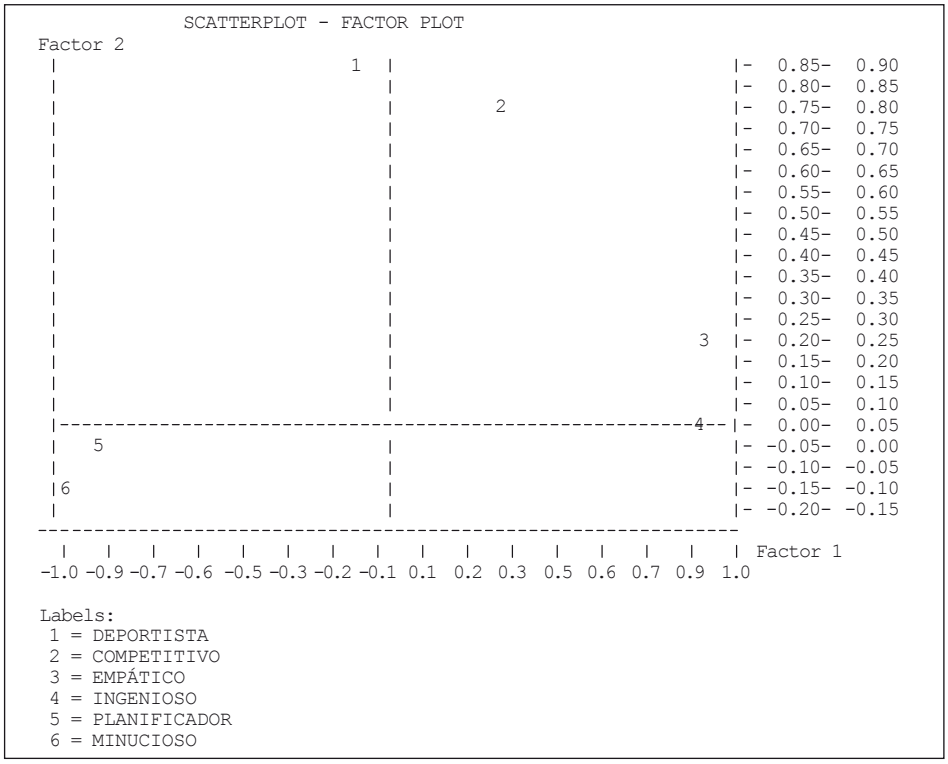
La sección *Communality Estimates as percentages* indica el porcentaje de la varianza de cada variable que es explicada por los dos factores seleccionados. En el caso de la variable MINUCIOSO vemos cómo ambos factores explican el 92,75% de su varianza. La variable con menor varianza explicada es COMPETITIVO, con el

66,57%. Puede comprobarse como el promedio de varianza explicada por los dos factores (92,751+75,226+77,985+85,721+66,572+94,179)/6 es igual al total explicado por ambos factores (82,07%).

Una vez producida la rotación de los factores (*Percent of Variation in Rotated Factors*) vemos cómo el Factor1 reduce el porcentaje de la varianza total explicada (pasando de 61,39 a 59,05) mientras el $Factor_2$ aumenta su porcentaje (de 20,68 a 23,03). Si bien globalmente el modelo factorial no cambia el porcentaje de varianza total explicada, la rotación facilita su interpretación por "acercar" las variables a los dos ejes, como aparece en la última pantalla.

En este diagrama vemos claramente las seis variables cerca de los ejes agrupadas dos a dos: DEPORTISTA-COMPETITIVO sobre el $Factor_2$, y EMPÁTICO-INGENIOSO y PLANIFICADOR-MINUCIOSO sobre el $Factor_1$, aunque con distinto signo. A la vista del tipo de variables agrupada en cada factor podríamos asignar al $Factor_1$ un sentido de EXTROVERSIÓN, ya que valores positivos se corresponden con EMPÁTICO e INGENIOSO y valores negativos con PLANIFICADOR y MINUCIOSO, por corresponder estas dos últimas características a individuos más introvertidos. El eje de ordenadas, $Factor_2$, podemos denominarlo ACTIVIDAD, por agrupar las variables DEPORTISTA y COMPETITIVO. Así, podemos definir a un individuo por sus coordenadas en el espacio factorial en lugar de utilizar las seis variables iniciales. Esta simplificación tiene un precio, una pérdida de 17,93% de la varianza inicial (100,00 – 82,07).

Figura 12.3. *Localización de las variables psicotécnicas en el plano factorial*



Finalmente, tras guardar o no las nuevas variables auxiliares utilizadas en el análisis factorial, obtenemos la puntuación de cada variable sobre los dos ejes, las cuales permiten situar a un individuo en el espacio factorial y describirlo en función de su extroversión y nivel de actividad.

```
SUBJECT FACTOR SCORE RESULTS:
Regression Coefficients with 15 cases.

Variables
                  Factor 1   Factor 2
MINUCIOSO           -0.272      0.019
DEPORTISTA          -0.133      0.680
PLANIFICADOR        -0.264      0.092
INGENIOSO            0.274     -0.080
COMPETITIVO          0.003      0.542
EMPÁTICO             0.256      0.062
```

Para localizar un individuo en el espacio factorial basta con sustituir las puntuaciones obtenidas en las seis variables en las siguientes ecuaciones:

$$Factor_1 = -\ 0{,}272 \cdot MINUCIOSO\ -\ 0{,}133 \cdot DEPORTISTA\ -\ 0{,}264 \cdot PLANIFICADOR + 0{,}274 \cdot INGENIOSO\ +\ 0{,}003 \cdot COMPETITIVO + 0{,}256 \cdot EMPÁTICO$$

$$Factor_2 = +\ 0{,}019 \cdot MINUCIOSO\ +\ 0{,}680 \cdot DEPORTISTA\ +\ 0{,}092 \cdot PLANIFICADOR - 0{,}080 \cdot INGENIOSO\ +\ 0{,}542 \cdot COMPETITIVO + 0{,}062 \cdot EMPÁTICO$$

¿A quién va a creer usted, a mí, o a sus propios ojos?
(Groucho Marx)

CAPÍTULO 13

ANÁLISIS DE CONGLOMERADOS

Capítulo 13. Análisis de conglomerados

13.1. Introducción teórica

El análisis de conglomerados (o *cluster* análisis en la terminología anglosajona) tiene como objetivo principal la clasificación de los casos en grupos relativamente homogéneos a partir de un conjunto de variables clasificatorias. Al igual que en el **análisis factorial**, todas las variables son analizadas simultáneamente sin distinción entre variables dependientes e independientes. En cierta medida, el análisis de conglomerados representa un técnica inversa al **análisis discriminante**: mientras el primero agrupa casos en función de un conjunto de variables, la segunda pondera las variables para que expliquen la agrupación ya existente.

El proceso de agrupamiento parte de una definición de distancia (generalmente se utiliza la distancia euclídea) entre los casos y el "centro" del grupo (conocido como centroide). Al comienzo del proceso de agregación, cada individuo se considera incluido en un grupo propio. En cada iteración se unen los dos grupos de mayor similitud, reduciéndose en uno el número de grupos de forma sucesiva, hasta obtener finalmente un único agregado. Este procedimiento de agregación por etapas puede representarse gráficamente a través de un dendrograma o árbol de semejanzas. Es el propio investigador quien decide por dónde "cortar" y llegar a una solución de compromiso entre la operatividad de la clasificación, que tiende a reducir el número de grupos, y la homogeneidad interna de cada grupo, la cual crece a medida que se incrementa el número de grupos[41].

En relación con la ejecución del análisis de conglomerados, es importante introducir exclusivamente las variables relevantes para la formación de grupos. La inclusión de variables irrelevantes puede distorsionar los resultados dando lugar a grupos no tan homogéneos como cabría esperar a *priori*. En el caso de tener variables medidas en diferentes unidades es necesaria la estandarización de las mismas, opción que permita realizar OS4 de forma automática.

13.2. Aplicación del análisis de conglomerados

Para la aplicación de esta técnica utilizaremos los datos sobre la actitud de una muestra de consumidores ante el hecho de ir de compras. Cada entrevistado, a través de una escala Likert de 7 categorías, mostró su total desacuerdo (valor mínimo de la escala = 1) o total acuerdo (valor máximo de la escala = 7) con cada una de las siguientes afirmaciones:

Tabla 13.1. *Ejemplo de variables consideradas en la actitud de los consumidores*

Afirmación en el cuestionario	Variable
Ir de compras es divertido	Diversión
Ir de compras es malo para el presupuesto	Gasto
Combino ir de compras con almorzar fuera de casa	Ocio
Intento conseguir las mejores ofertas	Ofertas
No pongo mucha atención cuando voy de compras	Indiferencia
Se puede ahorrar mucho dinero comparando precios	Comparación

Los datos recogidos en la encuesta aparecen en la tabla siguiente:

Tabla 13.2. *Datos de actitud de los consumidores ante el hecho de ir de compras*

	Puntuación 1-7 de cada una de la afirmaciones					
Individuo	**Diversión**	**Gasto**	**Ocio**	**Ofertas**	**Indiferencia**	**Comparación**
1	6	4	7	3	2	3
2	2	3	1	4	5	4
3	7	2	6	4	1	3
4	4	4	6	4	5	6
5	1	3	2	2	6	4
6	6	4	6	3	3	4
7	5	3	6	3	3	4
8	7	3	7	4	1	4
9	2	4	3	3	6	3
10	3	5	3	6	4	6
11	1	3	2	3	5	3
12	5	4	5	4	2	4
13	2	2	1	5	4	4
14	4	6	4	6	4	7
15	6	5	4	2	1	4
16	3	5	4	6	4	7
17	4	4	7	2	2	5
18	3	7	2	6	4	3
19	4	6	3	7	2	7
20	2	3	2	4	7	2

En OS4, para ejecutar el análisis de conglomerados seguimos la siguiente línea de comandos:

Analyses → Multivariate → Cluster Analysis → Veldman's Hierarchical Cluster Analysis: Variables Selected for Analysis: *Diversión, Gasto, Ocio, Ofertas, Indiferencia, Comparación*

No es necesario marcar la opción "`Options: Standardize Variables`" (ver Figura 13.1) porque todos las variables se miden en la misma unidad, esto es, una escala Likert desde 1 hasta 7.

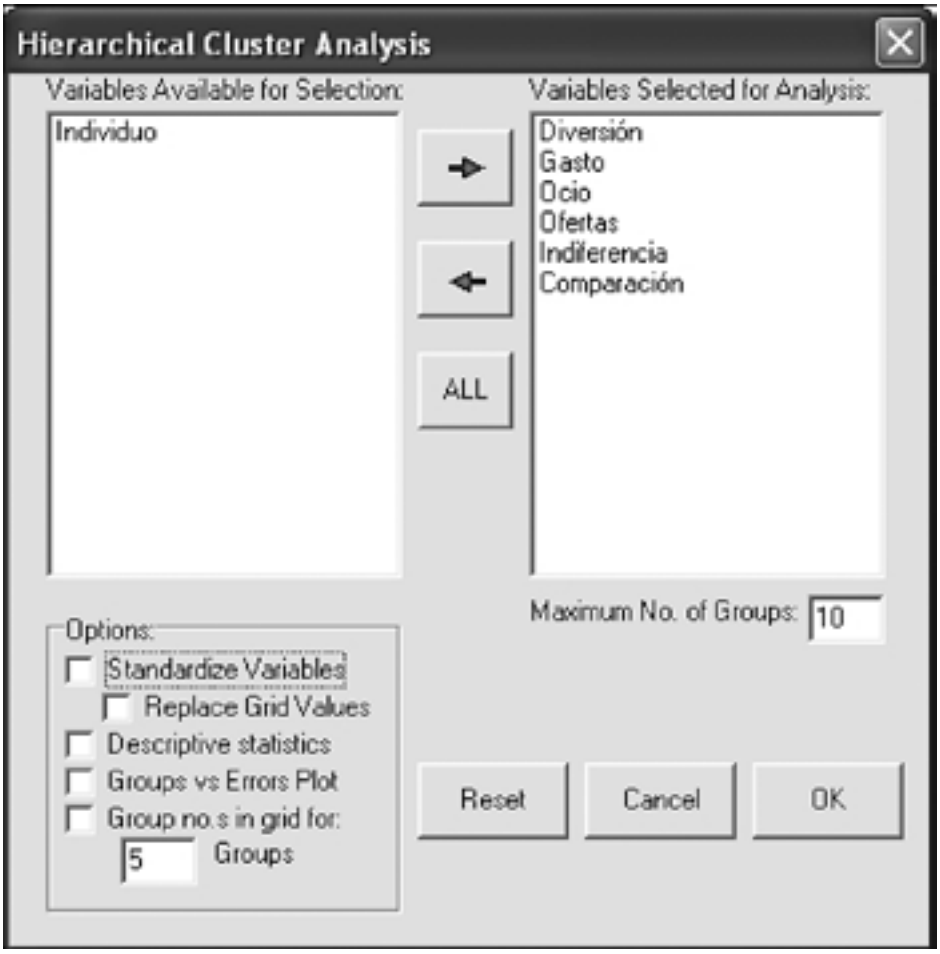
La opción "`Group no.in grid for:` `-n-` `Groups`" nos permite asignar en el panel de datos el grupo al que pertenece cada caso una vez especificado el número de grupos (*n*). Por ahora no haremos uso de esta opción ya que no hemos decidido aún el número de grupos en que vamos a dividir la muestra.

En el caso de tener un número elevado de variables (por ejemplo, más de diez) es aconsejable su reducción mediante el análisis factorial. Esta técnica tiene dos ventajas:

- Se estandarizan las unidades en las que se miden las variables,
- Se reduce el número de variables (por ejemplo pasando de diez a tres o cuatro).

Una vez obtenidos los ejes del análisis factorial se proceda a realizar el análisis cluster de la misma forma que se ha procedido en la línea de comandas anterior, sin embargo, en lugar de la variables originales se utilizan los factores como variables clasificatorias. A continuación se muestra el cuadro de diálogo del análisis cluster mediante las variables originales:

Figura 13.1. *Cuadro de diálogo del análisis cluster inicial*



En varias pantallas OS4 nos muestra el error que se comete dependiendo del número de conglomerados (cuanto mayor es el número de conglomerados menor será error pero menos útil es la clasificación) junto con los casos que contienen cada grupo (a partir de 10 porque en el cuadro de diálogo indicamos que el número máximo de grupos era precisamente ese número). Centrándonos en el valor de los errores en función del número de grupos tenemos:

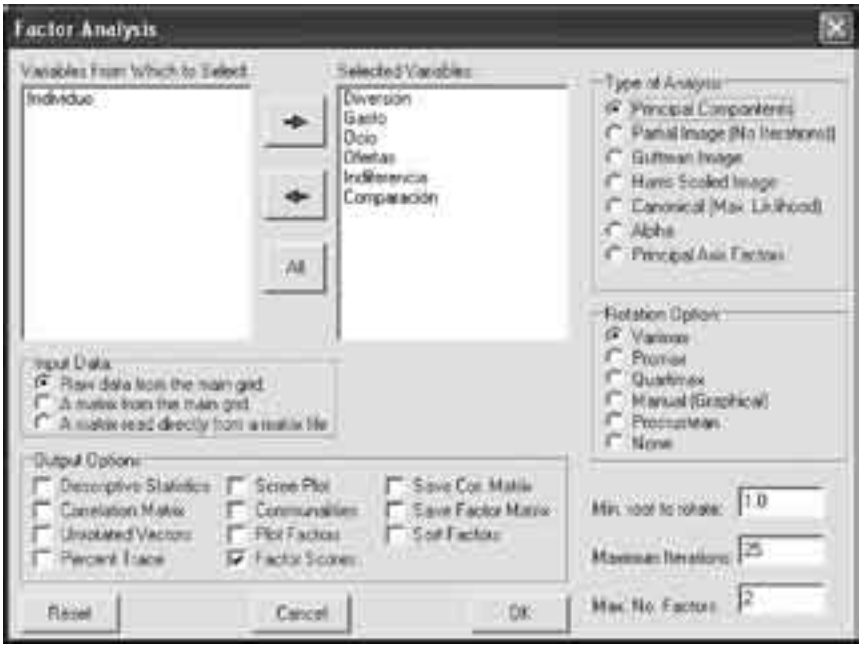
```
Hierarchical Cluster Analysis
Number of object to cluster:=20 on 6 variables.

19 groups after combining group 10 (n=1 ) and group 16 (n:=1) error:= 1.000
18 groups after combining group 6 (n=1 ) and group 7 (n:=1) error:= 1.000
17 groups after combining group 5 (n=1 ) and group 11 (n:=1) error:= 1.500
16 groups after combining group 3 (n=1 ) and group 8 (n:=1) error:= 1.500
15 groups after combining group 2 (n=1 ) and group 13 (n:=1) error:= 1.500
14 groups after combining group 10 (n=2 ) and group 14 (n:=1) error:= 1.667
13 groups after combining group 6 (n=2 ) and group 12 (n:=1) error:= 2.333
12 groups after combining group 5 (n=2 ) and group 9 (n:=1) error:= 2.500
11 groups after combining group 1 (n=1 ) and group 6 (n:=3) error:= 2.917
10 groups after combining group 10 (n=3 ) and group 19 (n:=1) error:= 4.500
9 groups after combining group 5 (n=3 ) and group 20 (n:=1) error:= 4.833
8 groups after combining group 1 (n=4 ) and group 17 (n:=1) error:= 5.350
7 groups after combining group 1 (n=5 ) and group 15 (n:=1) error:= 8.233
6 groups after combining group 2 (n=2 ) and group 5 (n:=4) error:= 10.500
5 groups after combining group 1 (n=6 ) and group 4 (n:=1) error:= 12.667
4 groups after combining group 1 (n=7 ) and group 3 (n:=2) error:= 13.300
3 groups after combining group 10 (n=4 ) and group 18 (n:=1) error:= 17.200
2 groups after combining group 2 (n=6 ) and group 10 (n:=5) error:= 84.917
```

El resultado muestra la variabilidad entre casos dentro de los conglomerados. Por ejemplo, si decidimos dividir la muestra en dos grupos este error alcanza un valor de 84,92. Si optamos por agruparlos en tres conglomerados el error baja hasta 17,20. Si se incrementa el número de grupos hasta 4, si bien reducimos el error cometido (baja hasta 13,30), esta reducción del error es porcentualmente mucho menor. Teniendo en cuenta estos resultados, la mejor solución de compromiso, entre un número reducido de conglomerados y una relativa homogeneidad entre objetos dentro de un mismo grupo, parece ser la elección de 3 grupos.

Como apoyo a la decisión sobre el número de grupos podemos recurrir al análisis factorial, el cual reduce el número de variables a ejes que son combinaciones lineales de las mismas. Si deseamos visualizar gráficamente la localización de los casos es necesario limitar el número de ejes a dos:

Analyses → Multivariate → Factor Analysis: *Diversión, Gasto, Ocio, Ofertas, Indiferencia, Comparación*; Output Options: *Factor Scores*; Type of analysis: *Principal Components*; Rotation Option: *Varimax*; Max. No. Factors: *2*..

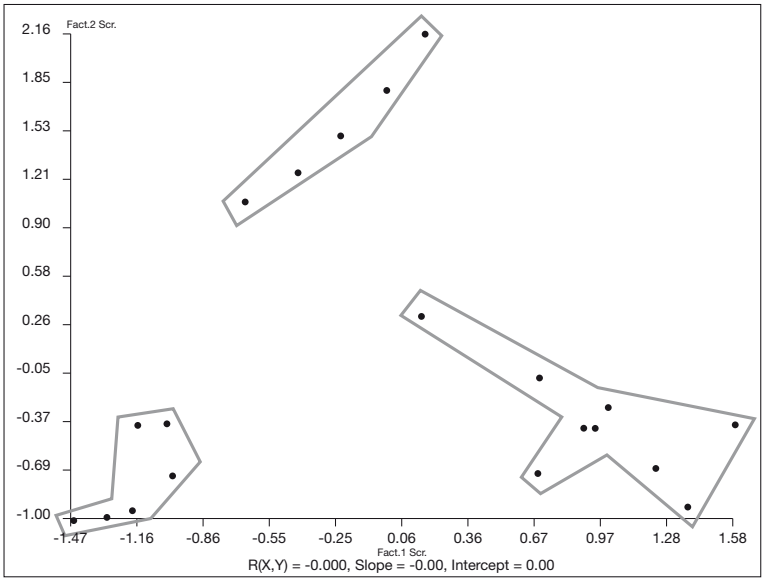
Figura 13.2. *Cuadro de diálogo del análisis factorial*



Una vez asignado en el panel de datos el valor que cada caso alcanza en los dos factores mediante dos nuevas variables (`FACT. 1 SCR. y FACT. 2 SCR.`) podemos proceder a la localización de todos los casos en el plano mediante un gráfico de dispersión:

Analyses → Descriptive → X versus Y Plot: *Fact. 1 Scr. y Fact. 2 Scr.*

Figura 13.3. *Representación gráfica de los conglomerados con OS4*



En el gráfico anterior se pueden apreciar claramente tres agrupaciones de casos, lo cual corrobora nuestra clasificación inicial basada en los errores calculados en el análisis cluster. Así pues, una vez determinado el número óptimo de conglomerados, repetimos el análisis indicando que el número máximo de clusters sea 3:

Analyses → Multivariate → Cluster Analysis → Veldman's Hierarchical Cluster Analysis: Variables Selected for Analysis: *Diversión, Gasto, Ocio, Ofertas, Indiferencia, Comparación*; Maximum No. of Groups: *3*.

OS4 nos proporciona todas las agrupaciones con un número de grupos igual o inferior al número marcado en la caja "`Maximum No. of Groups`". En nuestro caso, nos proporciona todos los casos distribuidos en 3 y en 2 grupos. Centrándonos en la clasificación de 3 conglomerados tenemos:

Tabla 13.3. *Clasificación de los casos según el análisis de conglomerados*

Caso	Grupo
Object :=Case 01	1
Object :=Case 03	1
Object :=Case 06	1
Object :=Case 07	1
Object :=Case 08	1
Object :=Case 12	1
Object :=Case 15	1
Object :=Case 17	1
Object :=Case 02	2
Object :=Case 05	2
Object :=Case 09	2
Object :=Case 11	2
Object :=Case 13	2
Object :=Case 20	2
Object :=Case 04	3
Object :=Case 10	3
Object :=Case 14	3
Object :=Case 16	3
Object :=Case 18	3
Object :=Case 19	3

Para analizar las características de cada grupo es necesario trasladar la información de la salida de OS4 sobre la pertenencia de cada caso a un grupo (Tabla 13.3) al panel de datos, repetimos el análisis marcando la opción "`Group no.in grid for: 3 Groups`":

Analyses → Multivariate → Cluster Analysis → Veldman's Hierarchical Cluster Analysis: Variables Selected for Analysis: *Diversión, Gasto, Ocio, Ofertas, Indiferencia, Comparación*; Maximum No. of Groups: *3*; Group no.in grid for: *3* Groups.

Obteniéndose en el panel de datos la asignación de cada caso al grupo correspondiente, como muestra la tabla siguiente:

Tabla 13.4. *Valor de las variables y pertenencia de los casos a cada grupo*

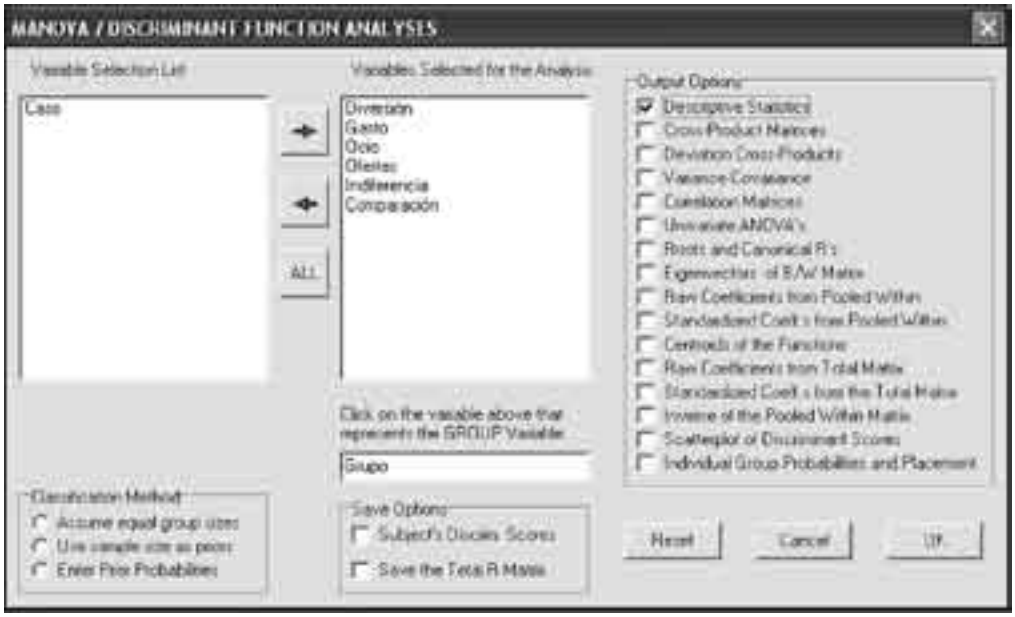
Caso	Diversión	Gasto	Ocio	Ofertas	Indiferencia	Comparación	Grupo
1	6	4	7	3	2	3	1
2	2	3	1	4	5	4	2
3	7	2	6	4	1	3	1
4	4	4	6	4	5	6	3
5	1	3	2	2	6	4	2
6	6	4	6	3	3	4	1
7	5	3	6	3	3	4	1
8	7	3	7	4	1	4	1
9	2	4	3	3	6	3	2
10	3	5	3	6	4	6	3
11	1	3	2	3	5	3	2
12	5	4	5	4	2	4	1
13	2	2	1	5	4	4	2
14	4	6	4	6	4	7	3
15	6	5	4	2	1	4	1
16	3	5	4	6	4	7	3
17	4	4	7	2	2	5	1
18	3	7	2	6	4	3	3
19	4	6	3	7	2	7	3
20	2	3	2	4	7	2	2

Una vez asignado cada caso a su grupo correspondiente en el panel de datos, el siguiente paso consiste en obtener el valor medio de cada variable clasificatoria por grupo y poder así sacar las conclusiones pertinentes sobre la caracterización de cada uno de los tres conglomerados. Para ello podemos aprovechar una de las opciones de la prueba multivariante siguiente:

Analyses → Multivariate → Discriminant Function / MANOVA. Group Variable: *Grupo*; Predictor Variables: *Diversión, Gasto, Ocio, Ofertas, Indiferencia, Comparación*; GROUP Variable: *Grupo; Options: Descriptives.*

Nota: primero seleccionamos todas la variables, incluyendo la variable de grupo (GRUPO), y después con un click en la variable de grupo “bajamos” esta variable al cuadro inferior, tal y como aparece en la figura siguiente:

Figura 13.4. *Uso de MANOVA para caracterización de los clusters*



Centrándonos en la primera parte del resultado y reorganizando las medias tenemos:

Variables	Diversión	Gasto	Ocio	Ofertas	Indiferencia	Comparación
Group 1	5.556	3.667	6.000	3.222	2.222	4.111
Group 2	1.667	3.000	1.833	3.500	5.500	3.333
Group 3	3.400	5.800	3.200	6.200	3.600	6.000
TOTAL	3.850	4.000	4.050	4.050	3.550	4.350

Recordando que el valor 1 significa total desacuerdo y el 7 total acuerdo, vemos cómo Grupo 1 tiene un sentido lúdico de la compra. Este grupo tiene una media muy superior en la actitud hacia la compra como una actividad divertida (5.56 frente a 3.85) y en la combinación de la compra con el almuerzo fuera de casa (6.00 frente a 4.05). No se sienten indiferentes ante esta actividad (2.22 frente a 3.55). Podríamos calificar este grupo como compradores compulsivos.

El Grupo 2 tiene una concepción totalmente opuesta al Grupo 1. Para estos consumidores comprar no es divertido (1.67 frente a una media de diversión de 3.85) ni aprovechan para comer fuera de casa. Esta actitud negativa hace que los califiquemos como compradores hostiles.

Por último, el Grupo 3 se caracteriza por considerar la compra como un gasto (5.80 frente a 4.00), buscar ofertas (6.20 frente a 4.05) y comparar precios (6.00 frente a 4.35). Claramente podemos considerarlos como consumidores ahorradores.

En primer lugar acabemos con Sócrates,
porque ya estoy harto de este invento de que
no saber nada es un signo de sabiduría
(Isaac Asimov)

REFERENCIAS
ANEJOS
NOTAS

Referencias

AKAIKE, H. (1973). Information theory and an extension of the maximum likehood principle. En: *Second International Symposium on Information Theory*. B.N. Petrov y F. Cszaki (Editores). Akademiai Kiado, Budapest, 267-281.

BARNETT, V. (1991). *Sampling survey*. Principles and methods. Arnold, Londres.

BERRY, W.D. (1993). *Understanding regression assumptions*. Series: Quantitative Applications in the Social Sciences 92. Sage Publications, Londres.

BINDER, A. (1984). Restrictions on statistics imposed by method of measurement: Some reality, some myth. *Journal of Criminal Justice* **12**, 467-481.

BISHOP, Y., FIENBERG, S. y HOLLAND, P. (1974). Discrete multivariate analysis, theory and practice. MIT Press, Cambridge.

BOROOAH, V. K. (2002). *Logit and probit*. Sage Publications, Londres.

BOWERMAN, B.L. y O'CONNELL, R.T. (1990). *Linear statistical models: an applied approach*. Duxbury, Belmont, CA.

BRAY, J.H. y MAXWELL, S.E. (1985). *Multivariate analysis of variance*. Series: Quantitative Applications in the Social Sciences 54. Sage Publications, Londres.

BOSCH, J.L. y TORRENTE, D. (1993). *Encuestas telefónicas y por correo.* Cuadernos Metodológicos 9. CIS, Madrid.

BREUSCH, T. y PAGAN, A. (1979). A simple test for heterocedasticity and random coefficient variation. *Econometrica* **47**, 1287-1294.

BRYANT, F.B. y YARNOLD, P.R. (1995). Principal Component Analysis and Exploratory and Confirmatory Factor Analysis. En: *Reading and Understanding Multivariate Statistics*. L. G. Grimm y P. R. Yarnold (Editores). American Psychological Association Books.

BRYMAN, A. y CRAMER, D. (1997). *Quantitative data analysis with SPSS for Windows*. *A guide for social scientists.* Routledge, Londres.

CAMPBELL, Y.Y., LO, A.W. y MACKINLAY, A.C. (1997). *The Econometrics of Financial Markets*. Princeton University Press, New Jersey.

CATTELL, R.B. (1966). The meaning and strategic use of factor analysis. En: *Handbook of Multivariate Experimental Psychology*. R.B. Cattell (Editor). Rand McNally, Chicago.

COMREY, A.L. y LEE H.B. (1992). A first course in factor analysis. Erlbaum, New Jersey.

CRAGG J.G. y UHLER, R. (1970). The demand for automobiles. *Canadian Journal of Agricultural Economics* **3**, 386-406.

CRAMER, J.S. (1991). *The Logit Model for Economists*. Edward Arnold Publishers, Londres.

CRASK, M.R. y FOX, R.J. (1987). An exploration of the interval properties of three commonly used marketing research scales: A magnitude estimation approach. *Journal of Marketing Research Society* **29**, 317-339.

DAVIDSON, R. y MACKINNON, J.G. (1981). Several tests for model specification in the presence of alternative hypothesis. *Econometrica* **49**, 781-793.

DAVIDSON, R. y MACKINNON, J.G. (1984). Convenient Specification Tests for Logit and Probit Models. *Journal of Econometrics* **25**, 241-262.

DILLMAN, D.A. (1978). Mail and telephone surveys: The total design methods. John Wiley and Sons, Nueva York.

DUFOUR, J.M., FARHAT, A., GARDIOL, L. y KHALAF, L. (1998). Simulation-based finite sample normality tests in linear regressions. *The Econometrics Journal* **1**, 154–173.

DUFOUR, J.M. y KHALAF, L. (2002). Simulation based finite and large sample tests in multivariate regressions. *Journal of Econometrics* **111**, 303-322.

DUNTEMAN, G.E. (1989). Principal components analysis. Sage university paper series on quantitative applications in the social sciences, 07-069. Sage, Newbury Park, CA.

ENGLE, R. (1984). Wald, likelihood ratio, and Lagrange multiplier test in econometrics. En: *Handbook of Econometrics*. Vol. 2. Z. Griliches y M. Intrilligator (Editores). Amsterdam.

FERRÁN ARANAZ, M. (1996). *SPSS para Windows. Programación y análisis estadístico*. McGraw Hill, Madrid.

FIELD, A. (2000). *Discovering statistics using SPSS for Windows*. SAGE Publications, Londres.

FOX, J. (1997). *Applied regression analysis, linear models, and related methods*. Sage Publications, Londres.

GARCÍA FERRANDO, M. (1995). *Socioestadística. Introducción a la estadística en sociología*. Alianza Universidad Textos, Madrid.

GIBBONS, J.D. (1993). *Nonparametric measures of association*. Series: Quantitative Applications in the Social Sciences 91. Sage Publications, Londres.

GIVON, M.M. y SHAPIRA, Z. (1984). Response to rating scales: A theoretical model and its application to the number of categories problem. *Journal of Marketing Research* **21**, 410-419.

GODFREY, L. (1978). Testing for multiplicative heterocedasticity. *Journal of Econometrics* **8**, 227-236.

GORSUCH, S.J. (1983). *Factor analysis*. Lawrence Erlbaum, Hillsdale, New Jersey.

GREENE, W.H. (1997). *Econometric analysis*. Prentice-Hall, New Jersey.

GUADAGNOLI, E. y VELICER, W. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin* **86**, 1255-1263.

GUJARATI, D.N. (1995). *Basic econometrics*. McGraw Hill, Londres.

HAGENAARS, J.A. (1993). *Loglinear models with latent variables*. Series: Quantitative Applications in the Social Sciences 94. Sage Publications, Londres.

HAIR, J.F., ANDERSON, R.E. y TATHAM, R.L. (1987). *Multivariate data analysis with readings*. MacMillan Publishing Company, Nueva York.

HAIR, J.F.; Anderson, R.E.; Tatham, R.L. y Black, W.C. (1998): *Multivariate Data Analysis*. Prentice Hall International, Upper Saddle River, New Jersey.

HARMAN, H.H. (1976). *Modern factor analysis*. University of Chicago Press, Chicago.

HARRIS, R.J. (1985). *A primer of multivariate statistics*. Academic Press, Orlando.

HARVEY, A.C. (1976). Estimating regression models with multiplicative heteroscedasticity. Econometrica 44, 460-465

HATCHER, L. (1994). *A step-by-step approach to using the SAS system for factor analysis and structural equation modeling*. Cary, NC: SAS Institute. Focus on the CALIS procedure.

HOSMER, D.W. y LEMESHOW, S. (1989). *Applied logistic regression*. Wiley, Nueva York.

HOWELL, D.C. (1997). *Statistical methods for psychology*. Duxbury, Belmont, C.A.

ISRAEL, G.D. (1992). *Determining Sample Size*. Fact Sheet PEOD-6. Program Evaluation and Organizational Development Series. University of Florida.

JACCARD, J. y WAN, C.K. (1996). *LISREL approaches to interaction effects in multiple regression*. Sage Publications, Thousand Oaks, CA.

KAISER, H.F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **23**, 187-200.

KHAMIS, H.J. (2000). The two-stage delta-corrected Kolmogorov-Smirnov test. *Journal of Applied Statistics* **27**, 439-450.

KIM, J.O. (1975). Multivariate analysis of ordinal variables. *American Journal Society* **81**, 261-298.

KLIEN, R. (1962). *An introduction to econometrics*. Prentice-Hall, New Jersey.

LITTELL, R.C., STROUP, W.W. y FREUND, R. (2002). SAS for linear models. Wiley-SAS, Londres.

LOBOVITZ, S. (1967). Some observations on measurement and statistics. *Social Forces* **46**, 151-160.

LOBOVITZ, S. (1970). The assignment of numbers to rank order categories. *American Social Review* **35**, 515-524.

MADDALA, G.S. (1983). *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University press. New York.

MALHOTRA, N.K. y BIRKS, D.F. (1999). *Marketing research. An applied Approach.* Prentice Hall, Londres.

MANLY, B.F.J. (1994). *Multivariate statistical methods*. *A primer*. Chapman & Hall, Londres.

MARDIA, K.V., KENT, J.T. y BIBBY, J.M. (1989). *Multivariate analysis*. Academic Press, Londres.

McCALLUM, RC., WIDAMAN, K.F., ZHANG, S. y HONG, S. (1999). Sample size in factor analysis. *Psychological Methods* 4 (1), 84-99.

McFADDEN, D. (1974). Conditional Logit Analysis of Qualitative Choice Behaviour. En: *Frontiers in econometrics*. P. Zarembka (Editor). Academic Press, Nueva York, 105-142.

MENARD, S. (2002). *Applied logistic regression analysis*. Series: Quantitative Applications in the Social Sciences 106. Sage Publications, Londres.

MENDENHALL, W. y SINCICH, T. (2003). *A second course in statistics: Regression Analysis*. Sexta edición. Pearson Education International, New Jersey.

MUÑOZ SERRANO, A. (2003). *Estadística aplicada uni y multivariante*. Consejería de Agricultura y Pesca, Junta de Andalucía, Sevilla.

MYERS, R. (1990). *Classical and modern regression with applications*. Duxbury, Boston.

NUNNALLY, J.C. (1978). *Psychometric theory*. McGraw-Hill, Nueva York.

PAMPEL, F.C. (2000). *Logistic regression*. *A primer*. Series: Quantitative Applications in the Social Sciences 132. Sage Publications, Londres.

PARK, R.E. (1966). Estimation with heterocedastic error terms. *Econometrica* **34**, 888.

RAMSEY, J.B. (1969). Tests for specification errors in classical linear least squares regression analysis. *Journal of Royal Statistical Society* **32**, 350-371.

RODRÍGUEZ OSUNA, J. (1991). *Métodos de muestreo. Cuadernos metodológicos* 1, CIS, Madrid.

RODRÍGUEZ OSUNA, J. (1993). *Métodos de muestreo. Casos prácticos*. Cuadernos metodológicos 6, CIS, Madrid.

ROMESBURG, H.C (1984). *Cluster analisys for researchers*. Lifetime Learning Publications, Belmont.

RUIZ-MAYA PÉREZ, L., MARTÍN-PLIEGO, J., LÓPEZ ORTEGA, J., MONTERO LORENZO, J.M. y URIZ TOME, P. (1990). *Metodología estadística para el análisis de datos cualitativos*. Centro de Investigaciones Sociológicas, Madrid.

SCHWARZ, C.A. (1978). Estimating the dimension and reality of a model. *Annals of Statistics* **6**, 461-464.

SEARLE. S.R. (1971). Linear models. John Wiley & Sons, Nueva York.

SHANKEN, J. (1996). Statistical methods in tests of portfolio efficiency: A synthesis. En: Handbook of Statistics 14: Statistical Methods in Finance. G.S. Maddala y C.R. Rao (Editores). North-Holland, Amsterdam, 693–711.

SHAW, R.G. y MITCHELL-OLDS, T. (1993). ANOVA for unbalanced data: an overview. *Ecology* **74**, 1638-1645.

SIEGEL, S. (1985). *Estadística no paramétrica aplicada a las ciencias de la conducta*. Trillas, México.

STEVENS, J. (1992). *Applied multivariate statistics for the social sciences*. Lawrence Erlbaum, New Jersey, Hillsdale.

TABACHNICK, B.G. y FIDELL, L.S. (1996). *Using multivariate statistics*. Harper and Row, Nueva York.

UPTON, G.J.G. (1980). *The analysis of cross-tabulated data*. John Wiley and Sons, Chichester.

VERBEEK, M. (2000). *A guide to modern econometrics*. Wiley, Chichester.

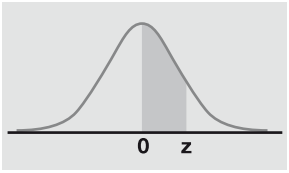
WHITE, H. (1980). A heterocedasticity consistent covariance matrix estimator and a direct test of heterocedasticity. *Econometrica* **48**, 817-818.

WONNACOTT, T.H. y WONNACOTT, R.J. (1990). *Introductory statistics for business and economics*. John Wiley & Sons, Londres.

YAMANE, T. (1967). *Statistics, an introductory analysis*. Harper and Row, Nueva York.

ZUMBO, B.D. y ZIMMERMAN, D.W. (1993). Is the selection of statistical methods governed by the level of measurement? *Canadian Psychology* **34**, 390-399.

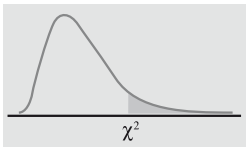
Anejo 1. Distribución normal



	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	**0,4750**	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

Por ejemplo, si z=1,96 el área es igual a 0,4750. Así, entre +1,96 y −1,96 se encuentran el 95% de los casos (0,4750+0,4750=0,9500).

Anejo 2. Distribución Chi-Cuadrado



df\area	**0,995**	**0,99**	**0,975**	**0,95**	**0,9**	**0,75**	**0,5**	**0,25**	**0,1**	**0,05**	**0,025**	**0,01**	**0,005**
1	0,000	0,000	0,001	0,004	0,016	0,102	0,455	1,323	2,706	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	0,575	1,386	2,773	4,605	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	0,584	1,213	2,366	4,108	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	1,923	3,357	5,385	7,779	**9,488**	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	2,675	4,351	6,626	9,236	11,071	12,833	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	3,455	5,348	7,841	10,645	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	2,833	4,255	6,346	9,037	12,017	14,067	16,013	18,475	20,278
8	1,344	1,647	2,180	2,733	3,490	5,071	7,344	10,219	13,362	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	4,168	5,899	8,343	11,389	14,684	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	4,865	6,737	9,342	12,549	15,987	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	5,578	7,584	10,341	13,701	17,275	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	6,304	8,438	11,340	14,845	18,549	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	7,042	9,299	12,340	15,984	19,812	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	7,790	10,165	13,339	17,117	21,064	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	8,547	11,037	14,339	18,245	22,307	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	9,312	11,912	15,339	19,369	23,542	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	10,085	12,792	16,338	20,489	24,769	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	10,865	13,675	17,338	21,605	25,989	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	11,651	14,562	18,338	22,718	27,204	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	12,443	15,452	19,337	23,828	28,412	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	13,240	16,344	20,337	24,935	29,615	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	14,041	17,240	21,337	26,039	30,813	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	14,848	18,137	22,337	27,141	32,007	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	15,659	19,037	23,337	28,241	33,196	36,415	39,364	42,980	45,559
25	10,520	11,524	13,120	14,611	16,473	19,939	24,337	29,339	34,382	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	17,292	20,843	25,336	30,435	35,563	38,885	41,923	45,642	48,290
27	11,808	12,879	14,573	16,151	18,114	21,749	26,336	31,528	36,741	40,113	43,195	46,963	49,645
28	12,461	13,565	15,308	16,928	18,939	22,657	27,336	32,620	37,916	41,337	44,461	48,278	50,993
29	13,121	14,256	16,047	17,708	19,768	23,567	28,336	33,711	39,087	42,557	45,722	49,588	52,336
30	13,787	14,953	16,791	18,493	20,599	24,478	29,336	34,800	40,256	43,773	46,979	50,892	53,672

Por ejemplo, para una función de distribución chi-cuadrado con 4 grados de libertad, el valor 9,49 deja a la derecha un área igual a 0,05. Así, la probabilidad de encontrar al azar un valor mayor que 9,49 es inferior al 5%.

Anejo 3. Datos de los ejemplos

EJEMPLO 1: Datos de pacientes con ataques cardiacos (variable dependiente “Cardio”)

Caso	Cardio	Altura	Peso	Sexo	Edad	Alimenta	Ejercicio
1	0	170	75	1	25	1	3
2	0	172	78	1	22	2	4
3	0	179	81	1	26	2	3
4	0	183	79	1	41	1	3
5	0	165	56	2	28	1	2
6	0	152	52	2	29	1	3
7	0	172	65	2	24	2	3
8	0	165	56	2	54	1	3
9	0	170	61	2	58	3	2
10	0	151	48	2	66	1	2
11	0	176	67	2	62	1	2
12	0	194	87	1	36	2	3
13	0	185	77	1	58	1	2
14	0	186	99	1	57	3	1
15	0	179	78	1	34	2	2
16	0	177	79	1	29	2	2
17	0	174	79	1	25	1	3
18	0	176	75	1	27	2	3
19	0	175	72	1	31	1	3
20	0	181	82	1	26	1	4
21	0	191	82	1	29	1	4
22	0	169	49	2	27	3	2
23	0	176	63	2	29	2	4
24	0	178	67	2	31	2	4
25	0	169	69	2	36	1	3
26	0	165	66	2	41	1	2
27	0	164	77	2	29	3	4
28	0	177	64	2	33	1	4
29	0	175	59	2	19	2	4
30	0	173	78	2	38	3	2
31	0	171	77	2	51	1	4
32	0	171	98	2	47	3	1
33	0	176	69	2	35	1	3
34	0	174	73	2	40	1	3
35	1	173	78	2	42	1	3
36	1	170	79	2	38	3	2
37	1	165	92	1	58	3	1
38	1	155	75	1	29	3	2
39	1	169	86	1	55	3	1
40	1	192	110	1	52	3	2
41	1	172	83	2	31	3	2
42	1	161	81	1	45	3	2
43	1	161	79	2	46	3	1
44	1	180	87	1	47	2	1
45	1	171	84	1	33	2	1
46	1	191	84	1	65	2	2
47	1	169	81	2	58	3	1
48	1	172	89	2	27	3	1
49	1	185	78	1	71	1	1
50	1	168	63	2	57	1	2

EJEMPLO 2: Disposición a pagar una entrada al parque (Dap)

Caso	Dap	Parque	Distancia	Primera	N_veces	Estudios	Edad	Ingresos	Sexo	Cazorla
1	0,00	1	50	0	1	1	3	2	1	0
2	1,50	1	80	1	1	3	2	2	2	0
3	0,90	5	12	0	4	2	4	2	1	0
4	0,00	5	12	0	30	2	4	2	1	0
5	1,80	1	110	1	1	3	2	2	2	0
6	0,00	3	3	0	20	3	3	3	2	0
7	1,20	4	40	0	2	1	6	1	2	0
8	1,80	3	51	0	1	3	3	3	2	0
9	3,01	1	80	0	2	1	2	2	1	0
10	3,61	3	150	0	1	3	4	4	1	0
11	3,01	1	92	1	1	2	3	2	1	0
12	1,20	4	4	0	15	2	2	2	1	0
13	1,20	4	70	0	4	1	4	2	2	0
14	1,80	1	32	0	5	1	3	1	1	0
15	0,00	1	45	0	2	4	3	1	2	0
16	3,01	1	95	0	1	3	3	2	1	0
17	3,01	3	250	1	1	3	5	4	1	0
18	0,60	4	30	0	2	2	2	2	1	0
19	3,01	4	40	0	4	1	3	3	2	0
20	1,20	5	12	0	3	3	2	1	2	0
21	6,01	1	470	0	2	1	5	4	1	0
22	1,80	3	51	1	1	3	4	3	2	0
23	0,00	1	25	0	3	3	2	2	1	0
24	0,00	5	12	0	2	2	3	1	2	0
25	0,00	4	70	0	2	4	6	1	1	0
26	0,00	3	50	0	1	3	2	2	2	0
27	1,20	4	13	0	15	1	2	2	1	0
28	0,00	4	7	0	10	4	6	2	2	0
29	1,80	2	150	0	1	2	4	1	1	1
30	0,00	3	51	1	1	3	2	2	1	0
31	0,00	4	28	0	3	1	4	1	1	0
32	1,20	5	12	0	4	2	4	2	1	0
33	1,20	1	40	0	4	4	6	1	1	0
34	6,01	2	400	1	1	2	2	3	2	1
35	1,20	5	12	0	3	2	6	1	2	0
36	0,00	4	7	0	1	1	5	2	2	0
37	0,00	5	12	0	3	2	4	2	1	0
38	3,61	2	370	1	1	3	5	3	2	1
39	1,50	3	50	1	1	1	3	3	2	0
40	0,00	5	12	0	1	4	2	1	2	0
41	0,00	1	12	0	4	3	3	2	2	0
42	0,00	4	70	0	2	4	6	1	2	0
43	0,60	5	12	0	1	2	2	1	1	0
44	0,00	1	40	0	3	1	6	1	1	0
45	3,01	4	80	1	1	3	2	2	2	0
46	3,01	2	200	1	1	4	3	2	1	1
47	3,01	3	51	1	1	3	2	4	1	0
48	0,00	4	50	1	1	4	6	1	2	0
49	1,80	3	51	1	1	3	4	4	2	0
50	0,00	3	3	0	30	2	3	2	1	0

Si bien en el texto hemos utilizado una muestra de 200 casos, para facilitar la introducción de los mismos por parte del lector, en el caso de que decida practicar por él mismo con OS4, hemos seleccionado una submuestra aleatoria de 50 casos (valor a partir del cual consideramos que tenemos una muestra de tamaño "suficiente"). Desde el punto de vista práctico no existe gran diferencia entre ambas muestras, como podemos ver por ejemplo a partir de los resultados del análisis de regresión de cada una:

Regresión lineal con los 200 casos:

```
Dependent variable: DAP

Variable     Beta    B         Std.Err.   t        Prob.>t  VIF     TOL
  DISTANCIA  0.678   0.009     0.001     16.341    0.000    1.075   0.930
   INGRESOS  0.328   0.468     0.059      7.895    0.000    1.075   0.930
  Intercept  0.000   -0.410    0.131     -3.142    0.002

SOURCE          DF        SS        MS         F  Prob.>F
Regression       2   260.236   130.118   213.816   0.0000
Residual       197   119.884     0.609
Total          199   380.120

R2 = 0.6846, F = 213.82, D.F. = 2 197, Prob>F = 0.0000
Adjusted R2 = 0.6814

Standard Error of Estimate = 0.78
F = 213.816 with probability = 0.000
Block 1 met entry requirements
```

Regresión lineal con una submuestra aleatoria de 50 casos:

```
Dependent variable: DAP

Variable    Beta    B        Std.Err. t        Prob.>t VIF     TOL
  DISTANCIA 0.671   0.010    0.001     7.465   0.000   1.281   0.780
   INGRESOS 0.279   0.458    0.148     3.104   0.003   1.281   0.780
  Intercept 0.000   -0.302   0.295     -1.024  0.311

SOURCE          DF      SS        MS        F Prob.>F
Regression       2  79.544    39.772   55.788  0.0000
Residual        47  33.507     0.713
Total           49 113.051

R2 = 0.7036, F = 55.79, D.F. = 2 47, Prob>F = 0.0000
Adjusted R2 = 0.6910

Standard Error of Estimate = 0.84
F = 55.788 with probability = 0.000
Block 1 met entry requirements
```

LISTA DE NOTAS

[1] Una exposición clara, a la cual nos remitimos para un desarrollo más extenso, puede encontrarse en Rodríguez Osuna (1991 y 1993).

[2] Barnett (1991) explica con detalle el desarrollo de las diferentes técnicas probabilísticas de muestreo, así como las ventajas e inconvenientes de cada una de ellas.

[3] Basta con sustituir z=1,96 por z=2,57 en la fórmula anterior.

[4] Ver por ejemplo García Ferrando (1995, p. 145).

[5] En el caso de un muestreo estratificado imaginemos que se seleccionó al individuo *i* del estrato *j*. Una vez se localiza a dicho individuo se descubre que no pertenece al estrato *j* sino al *k*, sin embargo, debido al coste ya incurrido, se decide entrevistarlo. Si no se consigue localizar a otro individuo del estrato *j* estaremos desvirtuando las frecuencias teóricas de la muestra. De igual forma, en el caso de un muestreo por cuotas, los encuestadores pueden tener instrucciones precisas sobre qué tipo de entrevistados deben buscar, sin embargo, es posible que transcurrido un tiempo les resulte difícil completar alguna cuota. Así pues, el incremento del coste que supone obtener exactamente las cuotas preestablecidas puede no verse compensado con la mejora de la representatividad de la muestra.

[6] Ver Bosch y Torrente (1993).

[7] Ver por ejemplo Bosch y Torrente (1993, p. 29).

[8] Es frecuente que el lector busque en los manuales de estadística cifras exactas que le permitan tomar una decisión. En la mayoría de las ocasiones los autores prefieren dejar a la imaginación del lector lo que él o ella entiende por "suficientemente bajo". Siendo este un trabajo eminentemente práctico, hemos procurado, cuando ha sido posible, dar alguna indicación que sirva de orientación.

[9] Ver Malhotra y Birks (1999, p. 277) o Hair *et al*. (1987, p. 9).

[10] Entre otras razones porque incluyen las técnicas que exigen escala métrica más las técnicas propias de las ordinales: Siempre podemos convertir una variable métrica en ordinal mediante intervalos.

[11] Para un mayor detalle consultar el excelente texto de introducción a la estadística de Wannacott y Wannacott (1990).

[12] Ver por ejemplo Ferrán (1996, p. 230).

[13] Encuestas realizadas por el Departamento de Economía y Sociología Agraria del CIFA "Alameda del Obispo" del Instituto Andaluz de Investigación y Formación Agraria, Pesquera y Agroalimentaria de la Junta de Andalucía (IFAPA).

[14] El *Type I Error Rate*: 0,05 es la probabilidad (5%) de rechazar un valor que sea cierto. En efecto, en la distribución existe un 5% de posibilidades de obtener un valor superior a 2,40, sin embargo, si obtenemos un valor superior (en este caso 8,40) optamos por descartar la hipótesis nula de homogeneidad de la varianza. ¿Por qué no entonces disminuir el nivel de significación del 5% al 1% por ejemplo? Porque entonces incrementamos el error de Tipo II, es decir, aceptar una hipótesis que es falsa.

[15] El cálculo e interpretación de estos coeficientes puede encontrarse en Gibbons (1993).

[16] La idea es utilizar Kendall tau cuando un gran número de casos se distribuyen en un número pequeño de categorías (Howell, 1997; Malhotra y Birks, 1999, p. 521).

[17] Ver Muñoz Serrano (2003, p. 223).

[18] Para profundizar en la teoría que respalda MANOVA consultar la monografía de Bray y Maxwell (1985).

[19] Si bien lo habitual es considerar factores que sean variables nominales (por ejemplo tratamiento A, B o C, hombre-mujer, etc.), con este ejemplo hemos querido poner de manifiesto que no existe un único enfoque a la hora de estudiar relaciones o diferencias entre variables.

[20] Ver por ejemplo Malhotra y Birks (1999, p. 557) o Manly (1994, p. 117).

[21] Ver por ejemplo Berry (1993), Gujarati (1995, p. 59) o Menard (2002, p. 4).

[22] La estimación de los parámetros por el método de mínimos cuadrados ordinarios (MCO) teniendo en cuenta el problema de heterocedasticidad da lugar a estimadores que, si bien siguen siendo lineales e insesgados, dejan de ser eficientes, es decir, no tienen la mínima varianza. En el peor de los casos, estimación de los parámetros por MCO sin tener en cuenta la heterocedasticidad, los estimadores no son insesgados, es decir, una muestra de mayor tamaño no da una estimación más próxima al valor verdadero.

[23] El incumplimiento del requisito de distribución normal de los residuos para muestras grandes se justifica por el teorema central de límite según el cual, la suma de variables con igual distribución, sea cual sea, se distribuye como una normal. Para más detalle consultar por ejemplo Gujarati (1995, p. 317), Greene (1997, pp. 280, 341), Fox (1997, p. 295) o Verbeek (2000, p. 34).

[24] Esta condición sí sería suficiente para un modelo con sólo dos variables explicativas. No lo es con tres o más variables porque una de ellas puede ser, por ejemplo, combinación lineal de las otras dos y sin embargo las correlaciones dos a dos presenten valores bajos.

[25] Ver Gujarati (1995, p. 330).

[26] Recordar que: suma de cuadrados total =suma de cuadrados de la regresión + suma de cuadrados de los residuos ($SC_{total}=SC_{reg}+SC_{res}$), donde el primer sumando es la varianza explicada por nuestro modelo y el segundo la varianza residual que no podemos explicar.

[27] Por ejemplo, en el caso de un modelo con dos variables explicativas tendríamos: $Y_i=b_0 + b_1X_1 + b_2X_2 + u_i$, como regresión inicial, y $u_i^2=b_0 + b_1X_1 + b_2X_2 + b_3X_1 \cdot X_2 + b_4 \cdot X_1^2 + b_5 \cdot X_2^2 + v_i$ como regresión auxiliar.

[28] Para una exposición clara del método de mínimos cuadrados ponderados con una aplicación práctica ver por ejemplo Mendenhall y Sincich (2003, p.438) o Gujarati (1995, pp. 362 y 381).

[29] Puede utilizarse el método de mínimos cuadrados con residuos heterocedásticos en el caso de que la varianza de los residuos no esté correlacionada con ninguna de las

variables independientes y dispongamos de una muestra de tamaño suficiente (Greene, 1997, p. 547).

[30] Aunque seguiría siendo necesaria la comprobación de residuos con media y covarianzas igual a cero y varianza constante.

[31] Si bien no es fácil determinar qué es un coeficiente de determinación aceptable, podemos apuntar que en las encuestas con datos socioeconómicos suelen aceptarse como buenos aquellos modelos con un coeficiente de determinación mayor que 0,50. Sin embargo un coeficiente de 0,20 puede ser aceptable en algunas aplicaciones mientras 0,95 puede ser rechazable en otras (Verbeek, 2000, p. 21).

[32] Basta con obtener los residuos de las regresiones residuos2=b_0 + b_1·*DISTANCIA* y residuos2=b_0 + b_1·*INGRESOS*, para la prueba de Park, y de la regresión P=b_0 + b_1·*DISTANCIA* + b_2·*INGRESOS* para la prueba de Breusch-Pagan-Godfrey y después aplicar la prueba de normalidad a la columna Raw Resid.

[33] Ya que la transformación, logarítmica o de raíz cuadrada, que pretende corregir la no normalidad de los residuos se aplica a una variable continua y no a una variable categórica (nominal u ordinal).

[34] Para una descripción más detallada de ambos modelos el lector puede remitirse a Hagenaars (1993), Gujarati (1995), Pampel (2000), Menard (2002) y Borooah (2002).

[35] Ya que ambas probabilidades del estadístico F son inferiores a 0,05 (0,0476 y 0,0002, respectivamente) rechazamos la hipótesis nula de no relación entre las variables.

[36] No confundir regresión lineal múltiple y regresión lineal multivariante (o modelo lineal general). En el primer caso tenemos una variable dependiente (Y) y múltiples variables independientes (X_1, X_2,... X_n). En el segundo, múltiples variables dependientes (Y_1, Y_2,... Y_n) y múltiples variables independientes (X_1, X_2,... X_n).

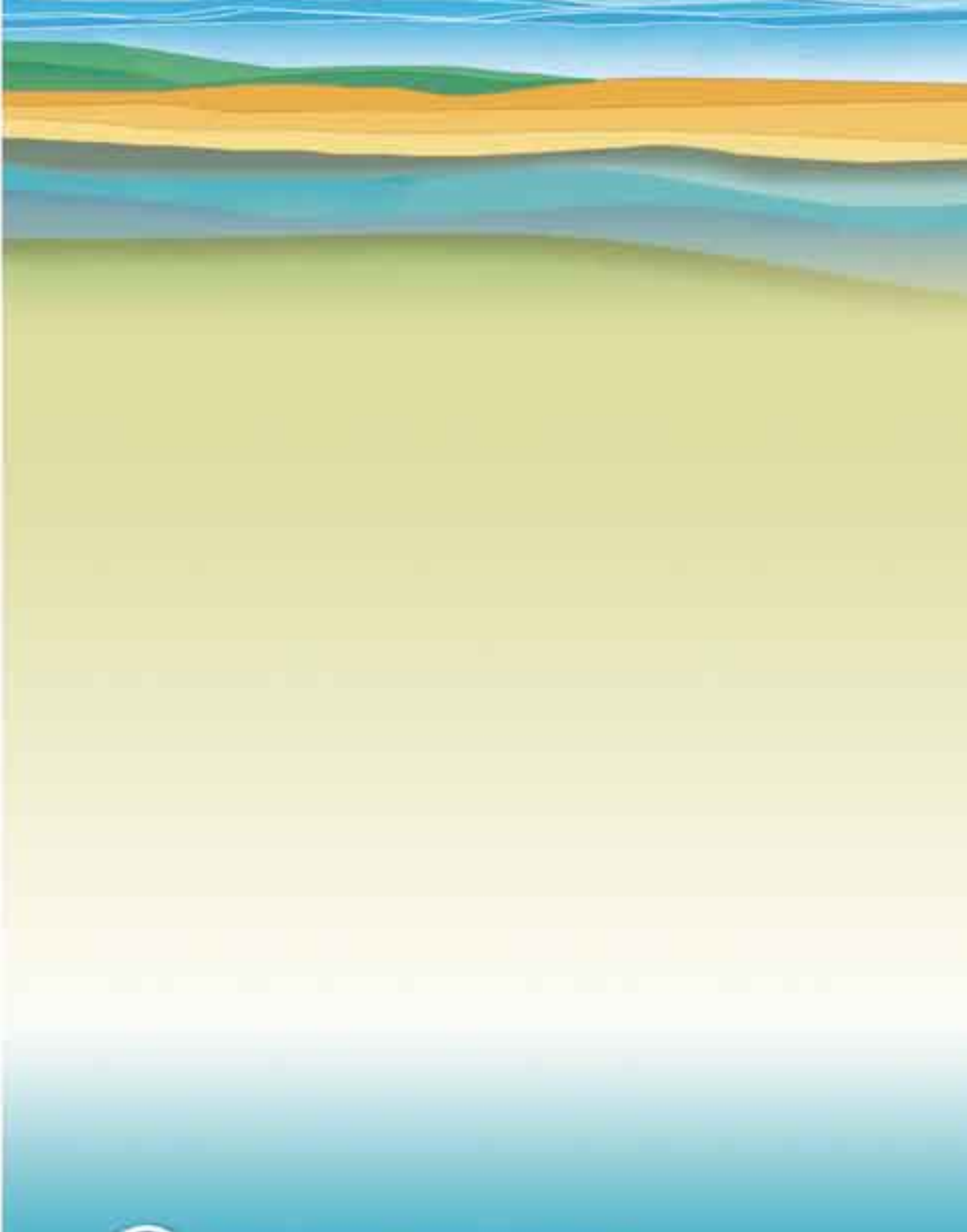
[37] Ver por ejemplo Manly (1994) y Harris (1985).

[38] Ver por ejemplo Ferrán Aranaz (1996, p. 253) o Littell (2002, p. 110).

[39] Dunteman (1989, capítulo 8) presenta una buena descripción de ambos procedimientos.

[40] Ver por ejemplo Muñoz Serrano (2003, p. 761).

[41] Ver por ejemplo Romesburg (1984) o Hair *et al*. (1998).



Instituto de Investigación y Formación Agraria y Pesquera
CONSEJERÍA DE INNOVACIÓN, CIENCIA Y EMPRESA