El análisis multivariable (Modesto Escobar)

1 Introducción

Como se ha podido deducir en el capítulo anterior, el análisis de datos aplicado está construido sobre los cimientos de dos conceptos básicos: el de caso, referido a la unidad de recogida de información y el de variable o característica susceptible de adquirir distintas modalidades. El primero atañe a la pregunta sobre qué personas, entidades u objetos se observan, mientras el segundo tiene que ver con la cuestión sobre qué características se recopilan en la realidad estudiada. En más complejos análisis que no serán abordados en este capítulo, entraría en consideración el concepto de tiempo, concerniente al interrogante de cuándo se ha obtenido la información de los casos analizados.

El análisis de datos persigue la descripción, comparación, reducción, clasificación, explicación o previsión a través del examen de las informaciones obtenidas de una serie determinada de objetos. Sus elementos básicos son las variables, que pueden ser analizadas en distintos niveles de complejidad: se procede al análisis univariable, si se las estudia una a una y se analiza la distribución de una sola característica en un conjunto de objetos. De este modo, solo podríamos describir la información disponible mediante la media, los porcentajes, la varianza, la desviación típica u otros estadísticos que reflejen características de las distribuciones. Con más complejidad, si se persigue la comparación o el estudio de la asociación existente entre un par de características, se pasaría a los problemas bivariables con otros tipos de medidas como las diferencias de medias o porcentajes y los coeficientes de asociación. Ahora bien, para la comprensión de los fenómenos naturales y especialmente los sociales, es necesaria la introducción conjunta de más de dos características, para lo que se necesitan técnicas que permitan la consideración conjunta de más de dos variables.

Tras esta presentación de análisis multivariable, conviene proceder su definición y a una primera presentación de los distintos procedimientos existentes para el estudio conjunto de más de dos variables. Para ello, se procederá desde las clasificaciones más globales a las más concretas, para describir someramente diversas técnicas de análisis multivariable. Posteriormente, se descubrirán conceptos centrales para realizar estos análisis y finalmente, a través de un ejemplo común, se explicarán con más detalle los análisis básicos que consideramos más empleados en cada categoría: el análisis factorial, la regresión múltiple, la regresión logística y el análisis de segmentación.

2 Definición y clasificaciones del análisis multivariable

Diríase, por tanto, que el análisis multivariable surge de la necesidad de analizar simultáneamente más de dos variables de las distintas unidades de análisis y se podría definir, en un sentido amplio, como aquellos procedimientos que abordan la distribución o la relación existente entre 3 o más variables. Estadísticamente, se tiende a emplear el concepto de multivariante más que el de multivariable, por responder a la idea de que se

trabaja con modelos de variables aleatorias (variantes), en lugar de variables empíricas. Sin embargo, en la literatura sociológica se ha preferido emplear el de multivariable (García Ferrando, 1984; Sánchez Carrión, 1984 y Cea, 2002).

De esta simple introducción al concepto de análisis multivariable se deduce la importancia de su uso, pues quienes redujesen la investigación a las relaciones y comparaciones entre pares de variables estarían simplificando extraordinariamente la realidad. En cualquier caso, la metodología multivariable no es una panacea pues la calidad de sus resultados depende de los datos con los que se trabaje y está constreñida a una serie de modelos de tratamientos de la información.

En primer lugar, es preciso mencionar la clásica clasificación de Kendall (1975), mencionada en la mayor parte de textos como así lo hacen Hair et al. (1999), que está basada en tres juicios que el analista puede adoptar en función de la naturaleza y la utilización de los datos: 1) Si las variables pueden ser clasificadas o no en dependientes e independientes, es decir, si se asume la existencia o no de relaciones causales entre variables; 2) si hay un modelo de dependencia, la determinación de cuántas variables dependientes han sido incluidas en el análisis, y 3) cómo se han medido las variables, distinguiendo al efecto entre variables métricas (cuantitativas) y no métricas (cualitativas). A partir de tales cuestiones, a cada problema le correspondería un determinado tipo de análisis. Cara a la presentación de las distintas modalidades de éstos, vamos a proseguir con la tipología propuesta por Kendall a la que se le añadirá la dimensión de la finalidad del análisis.

A este respecto, pueden distinguirse de acuerdo a Payne y O'Muircheartaigh (1977) dos grandes tipos: Aquellas cuyo fin es la *búsqueda de una estructura*, usadas para el descubrimiento de regularidades o irregularidades en los datos. En estos modelos la teoría sólo determina las variables a incluir; pero no se especifica la precisa fórmula funcional que las relaciona. El segundo tipo son las técnicas de ajuste de modelos, aplicables en aquellas áreas en las que el conocimiento predice un determinado tipo de relaciones entre las variables y encuentra en el análisis un instrumento para probar la bondad de *ajuste de los datos al modelo* establecido.

Similar a esta clasificación es la que presenta Sánchez Carrión (1984) en su *Introducción a las Técnicas de Análisis Multivariable* donde se abordan las técnicas de ajuste de modelos, y se distinguen aquellos procedimientos que persiguen la reducción (eliminación de la información redundante y conservación de lo esencial) de los que permiten la clasificación de los sujetos (agrupamiento de los individuos por la similitud en un conjunto múltiple de características) entre las técnicas de búsqueda de estructuras. Esta tripartita división de los análisis multivariables se puede expresar mediante la finalidad u objetivo de su aplicación. Si el objetivo es explicar acudiremos a aquellos análisis basados en los ajustes de modelos; si el fin es resumir un conjunto amplio de datos, se acudirá a las técnicas de reducción; y si el propósito es agrupar a los objetos por sus similitudes, se hará uso de técnicas con la finalidad de clasificar.

Por su parte, García Ferrando (2004: 384), además de distinguir entre análisis de dependencia y de independencia, diferencia aquellas técnicas que están centradas en las variables, las que se articulan en torno a las unidades de información, es decir los casos, y las que parten de la semejanza entre los objetos.

En el cuadro 1 se clasifican los análisis multivariables más comunes de acuerdo a los criterios acabados de exponer. En las filas aparecen las tres grandes finalidades de los análisis: resumir, explicar, clasificar; en las columnas se distinguen los modelos sin variable respuesta o dependiente, de aquellos en los que sí hay distinción entre resultados o variables dependientes, por un lado, y predictores o variables independientes, por el otro. Finalmente, mediantes símbolos se indican si los modelos de dependencia admiten más de una variable dependiente y si cada una de estas técnicas admiten o no variables no métricas.

Cuadro 1.- Clasificación de los análisis multivariables

Finalidad	Modelos de interdependencia	Modelos de dependencia
Resumir	Análisis factorial Análisis de correspondencias (*) Escalas multidimensionales[2] (*)	
Explicar	Correlaciones canónicas (&)	(=) Ecuaciones estructurales
	Log-lineal (*)	(=) Análisis de (co)varianza (/*) Regresión múltiple
Clasificar		Logit/Probit/Multilogit (*/) Análisis discriminante (*/) Análisis de segmentación[1] (*/*)
	Análisis de conglomerados[1]	

(*) Variables no métricas. (*/) V. dependientes no métricas. (/*) V. independientes no métricas. (*/*) V. dependientes e independientes no métricas. (=) Más de una variable dependiente. (&) Dos conjuntos de variables métricas.

1 Técnicas basadas en las unidades de información. 2 Técnicas basadas en la semejanza de objetos. El resto de técnicas están centradas en las variables

3 Procedimientos del análisis multivariable

Tras estas primeras clasificaciones, se procede a una primera presentación de los distintos procedimientos existentes para el estudio conjunto de más de dos variables. Para ello, se procederá desde las clasificaciones más globales a las más puntuales, deteniéndose en los diferentes tipos de técnicas concretas. Posteriormente, se explicarán con más detalle a través de un ejemplo los análisis que consideramos más empleados en cada categoría: el

análisis factorial, la regresión múltiple, la regresión logística y el análisis de segmentación.

3.1 Análisis de interdependencia

Los modelos de interdependencia se caracterizan por no distinguir tipos de variables. Todas desempeñan el mismo estatus. La primera subdivisión de tales métodos tiene que ver con la finalidad de su aplicación. De este modo, los análisis factoriales y las escalas dimensionales sirven para resumir; los conglomerados para clasificar, y los modelos log-lineales para explicar. Pero también de distinguen por la naturaleza de las variables que intervienen en el análisis: si se dispone de variables de características cualitativas podríamos optar por el análisis factorial de correspondencias o por el análisis multidimensional. El primero, en el caso de que se persiguiera descubrir las pautas de relación entre un conjunto de variables nominales; el segundo, para representar una serie de objetos en función de las evaluaciones o clasificaciones que de ellos hayan realizado un grupo de individuos. Véanse con más detalle:

En primer lugar, se considera por su finalidad explicativa al *análisis lineal-logarítmico*, que es utilizado para descubrir y diferenciar las asociaciones e interacciones presentes en un conjunto de variables nominales (Knoke, 1980 y Latiesa, 1991). Su empleo no requiere la división de las variables en dependientes e independientes, pues su finalidad analítica no es estimar los valores de la variable dependiente, sino expresar las diferentes frecuencias que aparecen en las casillas de un cruce de variables nominales en función de los valores del conjunto de variables analizadas. Por ello, está clasificado en las técnicas de interdependencia. No obstante, este análisis es apropiado para la comprobación de modelos causales de relación, dado que está provisto de procedimientos de contraste de hipótesis sobre la presencia de asociaciones e interacciones significativas y, en consecuencia, su finalidad es básicamente explicativa, habiendo sido ampliamente utilizados en el análisis sociológico para el estudio de la movilidad social.

Ahora bien, más que para explicar, los modelos de interdependencia son especialmente apropiados para clasificar o para resumir información. En este último caso, destacan tres técnicas según se dispongan de variables medidas en una escala nominal (correspondencias), ordinal (escalamiento multidimensional) o métrica (análisis factorial).

La utilización del *análisis factorial de correspondencias* responde a la necesidad de descubrir la estructura de asociaciones entre variables categóricas observadas, tratando de distinguir las categorías de un conjunto de variables (2 en el caso del análisis de correspondencias simple y más de 2 si se emplea su opción múltiple) que están más relacionadas entre sí (Greenacre, 2008). El resultado de la aplicación de este análisis es una representación espacial en la que se ubican las distintas categorías o valores de las que se componen las variables introducidas. Supóngase que se dispone una tabla donde aparece la composición de la población activa (variable A con los valores: Agricultura, Industria y Servicios) cruzada con distintos países (variable B con valores como España,

Francia, Honduras, Paraguay, Argentina, etc.). Esta técnica representaría a todos estos valores en un sistema de coordenadas interpretables por la distribución de los distintos puntos de información introducidos. Así en este caso, se configurarían dos dimensiones: una principal, que distinguiría a los países agrícolas del resto y una segunda que separaría a las naciones según su proporción de personas activas dedicadas a los servicios. Para ello se vale de distancias basadas en la χ^2.

El *escalamiento multidimensional* se utiliza para representar espacialmente una matriz de similitudes o de distancias (Sánchez Carrión, 1985). Sea, por ejemplo, un conjunto de x objetos –un conjunto de líderes políticos- evaluados ordinalmente por una muestra de individuos (casos). El resultado de las similitudes de las evaluaciones permite construir un mapa espacial con la virtud de que los objetos considerados más próximos queden más cercanos espacialmente y de esta forma puedan detectarse los criterios de evaluación implícitos que emplean los individuos para juzgar un conjunto de estímulos. El análisis matemático tiene como función la de convertir una matriz de distancias o similitudes en una representación geométrica. Así si introducimos como datos un conjunto de ciudades y la distancia kilométrica entre ellas, la técnica multidimensional produce una representación espacial de dónde se ubicarían las ciudades en un plano bidimensional. Este procedimiento posee distintos algoritmos según se les introduzcan variables de tipo ordinal o métrico.

Bajo el nombre de *análisis factorial* se encuentran comprendidas un conjunto de diferentes técnicas de las que las más conocidas son el análisis de componentes principales y el análisis factorial clásico (Yela, 1997). Todas ellas están basadas en la matriz de correlaciones calculadas a partir de una serie de variables de naturaleza cuantitativa. Su objetivo es descubrir la estructura de relaciones entre un conjunto numeroso de variables, reduciéndolas a componentes, factores de variación o variables latentes. Por ejemplo, si se introdujeran las notas de un grupo de estudiantes de enseñanza secundaria, las calificaciones podrían aparecer agrupadas en tres apartados: las de ciencias, las literarias y las actividades físicas o manuales.

Una extensión del análisis factorial puede emplearse cuando el analista esté interesado en la relación existente entre dos conjuntos de variables. La técnica que realiza al mismo tiempo un resumen de variables y un cálculo de la asociación entre las variables latentes conformadas es la de las *correlaciones canónicas* (Peña, 2002: 485-505; Hair et al., 1999: 469-488). Mediante ellas se puede: a) determinar si dos grupos de variables están relacionados entre sí y la magnitud de la relación entre ellos, b) encontrar una serie de dimensiones en cada uno de los conjuntos de variables preestablecidos, de tal manera que la relación entre ellas sea la máxima posible. Cada función canónica consiste en un par de variables nuevas, una para cada par de subconjuntos de las variables originales del análisis. Por ejemplo, podíamos relacionar una serie de medidas relacionadas con el rendimiento académico (las notas de los profesores) con otra serie compuesta por puntuaciones en pruebas de aptitudes. Esta técnica es útil y eficiente para explorar las relaciones entre pares de conjuntos de variables. Es un procedimiento que, al mismo tiempo, reduce dos grupos de variables y estima la máxima relación posible entre ellos.

También han de mencionarse los procedimientos de interdependencia que sirven para clasificar los casos, en lugar de centrarse en la reducción de variables. En ellos la finalidad es la de agrupar los casos en función de sus afinidades en un conjunto de variables para lo que emplean algoritmos de aglomeración de casos semejantes

El *análisis de conglomerados* clasifica a los sujetos en función de las distancias entre ellos en una serie de variables de naturaleza cuantitativa (Aldendarfer y Blashfield, 1984 y Everitt, 1984). Emplea técnicas clasificadoras en el sentido de que permiten una simplificación del conjunto de sujetos analizados. Por ejemplo, podemos disponer de datos de un conjunto de municipios de una provincia (habitantes, distancia a la capital, porcentaje de población agrícola, indicadores socioeconómicos) y en función de éstos establecer una tipología de tales municipios agrupando a aquellos con características semejantes y diferenciándolos del resto, que a su vez, formarán bloque con otros municipios con similares datos. Bajo tal técnica se encuentran una gran variedad de procedimientos para la consecución de la finalidad clasificatoria: así pueden utilizarse distintas medidas para juzgar la distancia entre casos y existen varios algoritmos para proceder a la clasificación de los sujetos.

3.2 Análisis de dependencia

La finalidad de todos estos análisis es similar: descubrir la cuantía y la significación de las asociaciones existentes entre unas variables con otras, si bien alguno de ellos, como se verá en cada caso, también pueda cumplir una función reductora o clasificadora. Una característica importante de los análisis estadísticos de la causalidad es que no son capaces de discernir qué variable juega el papel de causa y cuál el de efecto; por lo que ha de ser el investigador -dotado de una serie de teorías e hipótesis congruentes- el que determine de antemano a la realización del análisis el papel desempeñado por las variables seleccionadas. Utilizar una u otra técnica va a depender del tipo de modelo y de las características de las variables que se empleen.

En los *análisis de regresión múltiple* sólo se dispone de una variable dependiente cuantitativa (efecto) y se desea explicar con un más de una variable independiente, también de naturaleza numérica (Guillén, 2014 y Escobar et al., 2012: 271-367). Esta técnica permite descubrir qué variables tienen mayor peso en la determinación de la considerada variable dependiente. Por ejemplo, se podría concebir el rendimiento académico de los alumnos de una determinada área en función de variables como las aptitudes, el nivel económico de su familia y el historial académico. Esta técnica permite, además, cifrar el porcentaje de varianza de la variable dependiente explicado por las que son consideradas como determinantes de ella. Sin embargo, es necesario prever que tal técnica sólo da cuenta de las relaciones lineales entre variables, por lo que es posible que exista otro tipo de asociaciones no reflejadas mediante la ecuación de una recta. Ante esta desventaja siempre se puede recurrir a transformaciones de las variables que convierta la relación lineal en relaciones de tipo logarítmico o exponencial.

Todas estas transformaciones aplicadas a la variable respuesta, dependiente o resultado se consideran bajo lo que se denomina el modelo lineal generalizado, que implica modelos como el exponencial o el logatítmico para relaciones no lineales; el logit y el probit, cuando la respuesta es dicotómica; el multinomial, si es nominal con más de dos valores, o el de Poisson o binomial negativo, en el caso de que el resultado sea una variable discreta que represente un recuento (Long y Freese, 2014 y Escobar et al., 2012: 369-436).

Una alternativa a los modelos logit, probit o multinomial es *el análisis discriminante*, que halla los mejores predictores de la distribución de una característica cualitativa (Gil et al, 2001 y Cea, 2002: 322-425). Por ejemplo, se pueden usar variables como la edad, los ingresos, el número de miembros de la familia como determinantes de poseer un determinado tipo funcional de coche. Este análisis tiene, además de la función explicadora, una posibilidad de tarea clasificadora. En virtud de las regularidades encontradas en los datos, la técnica tiene como misión la de hallar una serie de funciones que permiten averiguar las probabilidades de pertenencia a uno u otro grupo del sujeto analizado. Y estas funciones se pueden utilizar para clasificar incluso a sujetos de los que no se posee el dato nominal. Siguiendo con el ejemplo, si hubiera un sujeto del que dispusiéramos la información de la edad, ingresos y miembros de su familia, pero no supiéramos si posee el tipo de coche en cuestión, podríamos averiguarlo aplicando a las tres primeras variables la ecuación discriminante obtenida en el análisis.

En el caso de *los análisis de varianza y covarianza*, la variable dependiente es de intervalo o razón (cuantitativa o métrica); pero las variables independientes son de naturaleza nominal. Este análisis, como el anterior, es una extensión de la técnica bivariada de su mismo nombre (Tejedor, 1999 y Tejedor, 2003). Un ejemplo de su uso sería la consideración del nivel educativo y de la región de residencia como factores determinantes de la renta de los individuos. Permite no sólo comprobar el factor que más incide en la variable determinada, sino también si existe interacción entre las variables independientes introducidas al explicar aquélla. Por su peculiar tratamiento de los datos no permite la introducción de muchas variables independientes a la vez. Este tipo de análisis posibilita dos extensiones a su estructura básica: una, es la introducción de variables independientes de naturaleza cuantitativa, las llamadas covarianzas; otra, el considerar como variable dependiente más de un dato, operación que se denomina análisis multivariado de la varianza.

Más complejos son los *modelos de ecuaciones estructurales*, en los que se conjugan dos tipos de análisis: un factorial confirmatorio, que permite comprobar si el procedimiento de medida basado en múltiples indicadores es apropiado, por un lado, y un análisis causal con múltiples variables dependientes con la posibilidad de tratamientos no recursivos, que hacen posible diferenciar los efectos de una variable sobre otra y de ésta sobre la primera (Bollen, 1989 y Batista, 2000).

Finalmente, cabe hacer referencia entre los modelos de dependencia a los *árboles de regresión y clasificación*, análisis de segmentación o, inicialmente, denominados

detectores automáticos de la interacción (Breiman, 1984 y Escobar, 2007). Estas técnicas tienen como finalidad dividir jerárquicamente una muestra (clasificar) en segmentos homogéneos en función de la variable criterio, utilizando un conjunto múltiple de predictores. Así si se quiere segmentar a los votantes de un determinado partido, es probable que el mejor predictor sea el de la ideología política y, una vez dividida la muestra en los grupos de derecha, centro y derecha, la edad juegue un papel importante de modo que las personas mayores sean más partidarias de ofertas conservadoras, mientras que los jóvenes opten por partidos más emergentes.

4 Conceptos básicos del análisis multivariable

Antes de estudiar las distintas formas de tratar un conjunto múltiple de variables, resulta conveniente desarrollar una serie de conceptos básicos sobre los que se articulan las distintas técnicas de tratamiento de los datos:

1) *Distribución conjunta*: Del mismo modo que cada variable aleatoria (X) posee una distribución dada marcada por la P(X<=x), se puede estudiar la distribución de más de una variable a través de la sucesión de las probabilidades de cada una de ellas:

$$P(X_1 \leq x_{1,} X_2 \leq x_2, ..., \mathrm{X}_p \leq x_p)$$

O lo que es lo mismo, la probabilidad de que dos o más fenómenos ocurran al mismo tiempo. Los modelos de distribución conjunta básicos en el análisis multivariable son el multinomial, el normal, el de Wishart, equivalente al χ^2 y el de Wilks que se corresponde con la F de Snedecor (Cuadras, 2014).

2) *Distribución condicionada*: A partir de la distribución conjunta, se puede obtener las distribuciones condicionadas si se supone fijo un(os) valor(es) de otra(s). Una distribución condicionada, por tanto, es la distribución que tiene una(s) variable(s) en el caso de que mantengamos constante otra(s).

$$P(X_1 \leq x_1 \mid X_2 = x_2, X_3 = x_3, ..., X_p = x_p)$$

3) *Asociación*: Dos variables se consideran asociadas en el caso de que haya una variación conjunta de sus valores a través de los casos de los que obtenemos información. Un ejemplo muy conocido de este tipo de relación entre variables es el de la renta y el consumo: a medida que aumenta la primera, la segunda también seguirá la misma tendencia. Las variables precio y demanda de un producto también estarían asociadas; pero en relación inversa, pues la tendencia es que a medida que sube el primero, la segunda disminuye. El término asociación está básicamente inserto en los modelos bivariables a través de la covarianza (**S**) y los coeficientes de correlación (**C**), que se obtienen multiplicando por sí misma la matriz de datos diferenciados $\boldsymbol{X} = \mathbf{X} - \mathbf{1}\bar{\mathbf{x}}'$ o reducidos: $\boldsymbol{Z} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\,\mathrm{diag}(\mathbf{S})^{-1/2}$.

$$\mathbf{S} = \frac{\boldsymbol{X}'\boldsymbol{X}}{n} \quad ; \quad \mathbf{C} = \frac{\boldsymbol{Z}'\boldsymbol{Z}}{n}$$

4) *Control*: Consiste en estudiar el comportamiento de dos variables, manteniendo constante los valores de otras terceras que pudieran estar deformando los valores o la relación de las primeras. Este concepto procede de la terminología experimental en la que el investigador tiene la posibilidad de controlar una serie de incidencias. Piénsese, por ejemplo, en el estudio de la influencia que tiene la aplicación de calor a un gas en su volumen. Para establecer tal pauta de relación el experimentador tendría que mantener constantes otras variables que podrían estar influyendo en el volumen del gas, especialmente la presión. El análisis estadístico permite el control de las terceras variables por procedimientos matemáticos: separando el estudio de la asociación por los distintos valores de la variable independiente (tablas de contingencia), transformando los valores de la variable de control de forma que sean constantes para los distintos casos analizados (regresión y análisis de varianza) o suponiendo una asociación nula entre la variable independiente y la de control, como así lo hacen el análisis de senderos o el proceso de elaboración de tablas propuesto por Lazarsfeld (1955).

5) *Interacción*: Es una extensión del concepto de asociación al comportamiento de más de dos variables. Su comprensión más inmediata se obtiene al definirla diciendo que se produce interacción cuando la relación entre dos variables depende de los valores de una tercera. Por ejemplo, la influencia que ejerce el tener un título universitario en los ingresos obtenidos por el trabajo puede ser más alto entre los hombres que entre las mujeres. Sin embargo, estrictamente hablando, tenemos que la interacción se produce cuando más de dos variables presentan una asociación conjunta de sus valores.

6) *Factorización*: Consiste en crear una variable en función de una combinación lineal de un conjunto de otras variables. En consecuencia, un factor sería una variable representada en distinto grado por una serie de variables simples. Pensemos, por ejemplo, en una serie de test de inteligencia: las variables originales serían las puntuaciones de los sujetos en las distintas pruebas a las que se les someten; un factor, sea el razonamiento abstracto, sería una suma particular de los puntos obtenidos en las diferentes pruebas. Particular, por el hecho de que existirían puntuaciones más relacionadas con el susodicho factor de la inteligencia y otras que tendrían menos que ver con aquél. Por tanto, para la obtención del factor deseado sumaríamos las puntuaciones ponderadas por su contribución a éste: si tienen poco que ver, se multiplicarán por una cantidad cercana a 0, por lo que su peso en la suma total sería muy bajo, y si tienen importancia en la configuración del factor serían multiplicadas por una constante tanto más alta como su contribución a la definición de la nueva variable fuera mayor.

7) *Representación espacial*: Los factores pueden ser considerados como dimensiones subyacentes en un conjunto de variables y son capaces de expresar con mayor parsimonia el conjunto de variaciones de los datos; dicho de otra manera, podríamos reducir la información de veinte variables a un subconjunto de factores que expresen las características de los sujetos. La reducción que conlleva la factorización permite una

mejor presentación gráfica de las pautas de las respuestas. La representación espacial es una técnica que permite expresar la posición bien de las variables, bien de los casos en un conjunto reducido de dimensiones con el objeto de comprender la pauta de distribución de una serie más extendida de datos.

8) *Distancia*: Es una medida de cuán semejantes puedan ser dos individuos en un conjunto de variables, o dos de éstas para una serie de datos. La medida más simple de distancia se aplica cuando sólo se pretende comparar en una sola faceta (variable) y viene dada por la diferencia. Así la distancia en ingresos entre dos sujetos que ganan respectivamente 100 y 150 euros mensuales sería de 50. Generalmente, las distancias interesan en términos absolutos, por lo que bien se prescinde del signo de la diferencia, bien se eleva al cuadrado con lo cual aumentan las diferencias de los casos más lejanos. La medida más común de distancia para más de dos variables es la euclidiana que es la raíz cuadrada del sumatorio de las distintas diferencias entre variables elevadas al cuadrado. Un importante problema de las distancias complejas es cuando se dispone de variables con distintas unidades de medida (euros y horas, por ejemplo) en cuyo caso habría que estandarizar los valores de aquéllas (restarle la media aritmética y dividir el resultado por la desviación típica). Las distancias más comunes en el análisis multivariable son la euclidiana y la de Mahalanobis.

9) *Clasificación*: Es una técnica cuya finalidad consiste en agrupar objetos homogéneos. En el análisis multivariado se puede aplicar esta técnica bien a las variables, bien a los casos. Los medios para tal fin se pueden clasificar en dos grandes grupos: aquellos que se basan en las distancias o similitudes entre los objetos clasificables, y los que se fundamentan en las probabilidades de pertenencia a uno u otro grupo, en función de una serie de características.

5 Ejemplos de análisis multivariables

Vistos los principales conceptos y operaciones sobre los que se asienta el análisis multivariable, se procede a continuación a una primera aproximación a cuatro de ellos a través de ejemplos. El objetivo perseguido en las próximas páginas es el de que el lector sepa interpretar los análisis multivariables presentes en la literatura de las ciencias sociales. Por ello, se presentan las tablas no como se producen en las aplicaciones estadísticas más usuales, sino como suelen aparecer en los artículos. Además de ello, se prestará más atención en el significado de los resultados que en el algoritmo de su obtención.

5.1 Resumen de variables: el análisis factorial

Como se dijo anteriormente, el análisis factorial es una técnica multivariada cuya finalidad es la de identificar una estructura en un conjunto de variables observadas.

El punto de partida es una matriz de correlaciones compuesta de un conjunto de m variables obtenidas de un conjunto de sujetos. El objetivo es encontrar una serie de k factores (k<m) que expliquen al máximo posible las asociaciones de las que se parte. De ahí que este análisis tenga dos funciones distintas:

- Reducir el número de variables a analizar, de forma tal que se pierda el mínimo de información posible.

- Encontrar una serie de dimensiones latentes en la pauta de asociación de un conjunto de variables analizadas.

Los conceptos analíticos que se utilizan en esta técnica reductiva son los siguientes:

- Factor o componente (**Z**): Es el resultado de una transformación de las variables con distintos pesos. Ello implica que son variables latentes elaboradas a partir de las observadas.

- Autovalor (**A**): Es la cantidad de varianza de la que da cuenta cada factor. La suma de autovalores debe ser igual al número de variables.

- Saturación (**V**): Es la contribución de cada variable a cada factor; por tanto, es una descomposición del autovalor, implicando que la suma de todas las saturaciones al cuadrado de un factor es igual a su autovalor.

Siendo **X** los datos originales, la relación entre **Z** y **V** en el análisis factorial de componentes principales es la siguiente:

$$\mathbf{Z} = \mathbf{XV}$$

- Comunalidad: Es la cantidad total de información que suministran los factores sobre una determinada variable. Su valor inicial depende de una decisión del investigador. Si se utiliza el método de componentes principales se asume que el conjunto de factores explican por completo las variables. En el caso del análisis factorial de ejes principales, sólo se intenta explicar con los factores la varianza de cada variable que esté recogida por el resto de variables.

- Representación gráfica: Consiste en ubicar las variables en las dimensiones correspondientes a los factores de acuerdo a las saturaciones que aquéllas presenten en éstos. Generalmente sólo se representan simultáneamente dos dimensiones.

- Rotación: Consiste en mover los ejes que representan a los factores, de forma tal que se configure una estructura de más fácil interpretación. Hay varios criterios de rotación; el más útil y empleado en la investigación es el de varimax cuyo fin es el de maximizar la varianza entre factores (engendrar saturaciones muy heterogéneas entre las distintas columnas).

El procedimiento de interpretación de una salida de ordenador de esta técnica ha de seguir los siguientes pasos:

a) Buscar la solución inicial, fijarse en los autovalores con objeto de descubrir de todos los factores aquéllos que explican una sustancial cantidad de varianza (la regla más simple es la de tener sólo en cuenta factores con autovalor superior a la unidad).

b) Examinar brevemente la matriz de saturaciones (autovectores) de los factores iniciales: Esta matriz presenta generalmente una estructura peculiar basada en altas saturaciones en el primer factor y en las restantes saturaciones medianas de ambos signos (positivas y negativas).

c) Evaluar las estadísticas finales observando la cantidad de varianza de la que informan los factores seleccionados y la comunalidad de las variables. Si en esta última observación se detecta alguna cantidad baja - por debajo de .20- se debería modificar el análisis bien introduciendo más factores, bien eliminando las variables con baja comunalidad.

d) Analizar la matriz de factores rotados, subrayar para cada factor aquellas variables que más saturan en él e interpretarlo en función del significado común de aquellas variables que le son más propias.

Como ejemplo numérico se utilizan seis variables del estudio de PISA relacionadas con pruebas de rendimiento escolar: dos modos de puntuar el rendimiento en matemáticas (M1 y M2), dos modos de puntuar el rendimiento en ciencias (C1 y C2) y otras dos puntuaciones para valorar la habilidad lectora de los estudiantes de 16 años (L1 y L2). Los datos proceden de la muestra del estudio que la OCDE realiza cada par de años. Las medias de estas variables son las siguientes:

Tabla 1.- Medias y desviaciones típicas de las variables de medidas de rendimiento

Variable	Media	Desv. típica
1ª puntuación en Matemáticas	484.6	87.3
2ª puntuación en Matemáticas	484.5	87.9
1ª puntuación en Ciencias	496.5	85.8
2ªpuntuación en Ciencias	496.5	86.2
1ª puntuación en Lectura	488.3	91.7
2ª puntuación en Lectura	488.0	92.0

Fuente: http://pisa2012.acer.edu.au/downloads.php

Habría que aclarar que estas puntuaciones son discrecionales. La media de todas las variables en el conjunto de la muestra de la OCDE es de 500 puntos, mientras que la desviación típica se ha fijado en el valor 100. Observando la tabla de medias y desviaciones típicas de las variables, se ve que en todas las pruebas las medias españolas están por debajo de los 500 puntos. De igual modo, la desviación típica es menor del centenar. El primer dato quiere decir que el rendimiento escolar en el sistema educativo español es menor que en la media de la OCDE. En cambio, una desviación típica más pequeña indica que hay menor dispersión en los rendimientos que en el conjunto de los países evaluados.

El siguiente paso sería el cálculo de los coeficientes de correlación. Como era de esperar, presentan valores altos, especialmente, si se correlacionan las mismas pruebas (primera puntuación en Matemáticas con segunda; primera en Ciencias con segunda en la misma materia, y primera en Lectura con la segunda también en la misma), pues en este caso

rondan el 0,90. En ningún caso, este valor baja del 0,77 en el caso de correlaciones entre las puntuaciones en ciencias y las de la lectura.

Tabla 2.- Matriz de correlaciones entre las variables de rendimiento.

Variable	Mat. 1	Mat. 2	Cien. 1	Cien. 2	Lect. 1	Lec. 2
1ª puntuación en Matemáticas	1					
2ª puntuación en Matemáticas	0.924	1				
1ª puntuación en Ciencias	0.880	0.826	1			
2ªpuntuación en Ciencias	0.826	0.879	0.900	1		
1ª puntuación en Lectura	0.825	0.779	0.823	0.774	1	
2ª puntuación en Lectura	0.781	0.827	0.778	0.828	0.891	1

Fuente: http://pisa2012.acer.edu.au/downloads.php

El primer resultado del análisis factorial es la lista de autovalores, compuesta por tantos como variables se hayan introducido en el análisis. Las cantidades de los autovalores siempre aparecerán ordenadas de mayor a menor. Con estos datos, el valor del primer autovalor es altísimo (5.2 sobre 6), pues refleja que hay un alto componente común en el rendimiento de los estudiantes, sea en matemáticas, ciencias o lectura. Por tanto, en lugar de trabajar con estas seis variables, se podría usar solo el primer componente y de este modo se daría cuenta del 86% de variación del conjunto con una solo factor.

Tabla 3.- Autovalores del análisis factorial con medidas de rendimiento.

Variable	Autovalor	%	% acum.
Factor 1	5.18	86.4%	86.4%
Factor 2	0.33	5.5%	91.8%
Factor 3	0.21	3.4%	95.3%
Factor 4	0.20	3.3%	98.5%
Factor 5	0.06	1.0%	99.5%
Factor 6	0.03	0.5%	100.0%

Fuente: http://pisa2012.acer.edu.au/downloads.php

Como puede apreciarse en la tabla de saturaciones, todas las variables contribuyen de modo similar (más de .90) en la configuración de este factor.

Tabla 4.- Matriz de saturaciones del análisis factorial.

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
1ª puntuación en Matemáticas	0.940	-0.173	-0.240	0.130	-0.045	-0.105
2ª puntuación en Matemáticas	0.940	-0.173	-0.200	-0.184	0.044	0.105
1ª puntuación en Ciencias	0.934	-0.152	0.206	0.219	-0.095	0.071
2ªpuntuación en Ciencias	0.934	-0.152	0.254	-0.157	0.099	-0.070
1ª puntuación en Lectura	0.912	0.338	-0.038	0.187	0.130	0.017
2ª puntuación en Lectura	0.915	0.328	0.020	-0.193	-0.133	-0.017

Fuente: http://pisa2012.acer.edu.au/downloads.php

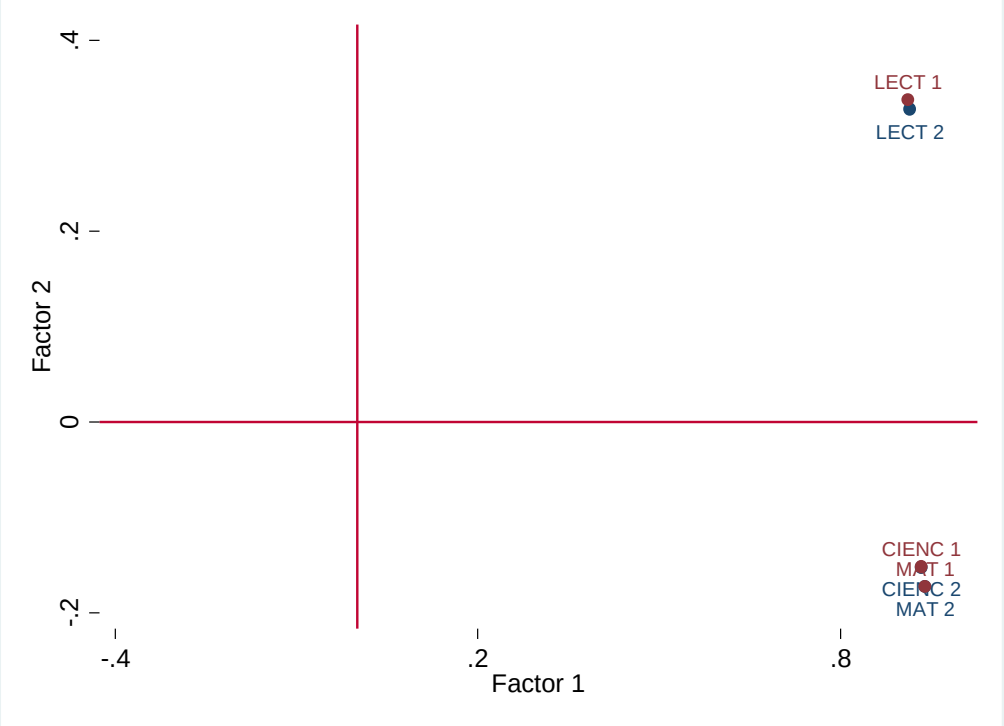

Gráfico 1.- Saturaciones de la solución factorial inicial

Ahora bien, si lo que se desea es más bien conseguir grupos de variables claramente diferenciados, se debería optar por rotar la solución inicial. Como lo que se desea es obtener tres grupos, se solicita la rotación de tres factores, en cuyo caso el resultado es un primer factor con saturaciones altas en las puntuaciones en matemáticas, un segundo factor donde se ubican las puntuaciones en ciencias y un tercer factor en el que se agrupan los rendimientos lectores. Como puede observarse, tras la rotación la varianza de los tres primeros componentes, reflejada en los autovalores, queda más repartida que en la solución sin rotar.

Tabla 5.- Autovalores del análisis factorial rotado (varimax) con medidas de rendimiento

Variable	Autovalor	%	% acum.
Factor 1	2.02	33.6%	33.6%
Factor 2	1.90	31.7%	65.3%
Factor 3	1.79	29.9%	95.3%

Fuente: http://pisa2012.acer.edu.au/downloads.php

En la matriz de saturaciones puede comprobarse como desaparece el factor general y se convierte en una solución con tres factores correspondientes a las materias de matemáticas, ciencias y lectura.

Tabla 6.- Matriz de saturaciones del análisis factorial rotado

Variable	Factor 1	Factor 2	Factor 3	Comun.
1ª puntuación en Matemáticas	0.421	0.785	0.420	0.970
2ª puntuación en Matemáticas	0.420	0.759	0.448	0.953
1ª puntuación en Ciencias	0.423	0.469	0.734	0.938
2ªpuntuación en Ciencias	0.422	0.437	0.770	0.961
1ª puntuación en Lectura	0.812	0.404	0.354	0.948
2ª puntuación en Lectura	0.804	0.370	0.402	0.945

Fuente: http://pisa2012.acer.edu.au/downloads.php

Los análisis presentados se han realizado empleando el método de componentes principales. Podría haberse optado por la solución de ejes principales, en cuyo caso, en lugar de tratar de dar cuenta de toda la varianza del conjunto de variables, solo se pretende informar de la varianza común entre las variables, valor que inicialmente puede ser estimado con el coeficiente de correlación múltiple de cada variable observada con el resto de variables.

5.2 La explicación en el análisis multivariable: la regresión múltiple

La regresión múltiple es uno de los procedimientos más empleados para explicar en el caso de disponer de datos transversales. A diferencia del análisis factorial, en el que se disponía de un conjunto múltiple de variables, en el análisis de regresión hay que distinguir entre una variable resultado, respuesta o dependiente y una serie de variables predictoras o independientes.

El modelo de la regresión múltiple se adecua a la siguiente expresión matricial:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{e}$$

Por tanto, en toda regresión múltiple puede distinguirse:

a) Una variable dependiente cuantitativa (**Y**) que es el explanandum del análisis, es decir, aquello que debe ser explicado.
b) Una serie de variables independientes (**X**), de las que se estima el valor de sus coeficientes (**B**) mediante el criterio de mínimo cuadrados, de modo similar al que se realiza en la regresión simple vista en el capítulo anterior, aunque en este caso no hay una sola variable sino varias y se hace una estimación suponiendo que el resto de ellas se mantienen constantes. De cada uno de estos coeficientes puede calcularse el error típico con el fin de decidir si son significativos.
c) Un error residual (**e**), que consiste en la diferencia entre el valor empírico de la variable dependiente y su valor pronosticado aplicando los coeficientes obtenidos a las variables independientes.
d) Un coeficiente de determinación (R^2), que representa la varianza no residual de la variable dependiente, o dicho con otras palabras, el porcentaje de la variación de la variable predicha que es explicado por el conjunto de variables independientes.

En la regresión se obtiene para cada variable independiente el parámetro que mejor ajusta la predicción de la variable respuesta. Por ejemplo, si se desea poner la puntuación en matemáticas en función de la edad y el status, este análisis aportaría un coeficiente para cada una de estas variables, que indicaría cuánto varía el valor medio de la respuesta por cada unidad que varíen los predictores.

Se sabe que Matemáticas es una variable que está medida de tal forma que la media de estudiantes de la OCDE investigados es de 500 puntos y su desviación típica 100, que la edad recoge solo los individuos nacidos en un determinado año y, en consecuencia, solo recoge el mes en el que han nacido, que el estatus socio-cultural de la familia está tipificado de modo que la media es 0 y la desviación típica es igual a la unidad.

La regresión se conforma al modelo lineal. Por ello, además de obtener los parámetros para cada variable incluida en el modelo, se obtiene otro que concierne a la constante y es expresión del valor medio de la variable predicha en el caso de que todas las variables independientes tengan el valor de cero.

Siguiendo con el ejemplo de la puntuación en matemáticas y transformando el mes de nacimiento de modo tal que el mes de diciembre (correspondiente a los estudiantes más jóvenes) sea representado con un 0 y el mes de enero (propio de los alumnos más veteranos) con un 11, se obtiene una ecuación con tres parámetros: la contante (488) indica la puntuación en matemáticas de un estudiante español nacido en diciembre en el seno de una familia con estatus medio en la OCDE. El coeficiente relativo a la edad (0,7) indica que, por cada mes de más que tengan los estudiantes, su media en matemáticas se verá incrementada en algo más de medio punto. Por su lado, el coeficiente del estatus, que presenta un valor aproximado de 33, informa de que, en promedio, por cada punto de estatus más que tenga un estudiante su puntuación en matemáticas se incrementará en más de treinta puntos.

La bondad de la regresión se evalúa con el coeficiente de determinación (R^2), que representa el porcentaje de varianza de la variable dependiente que está informado por las variables independientes y procede del cociente entre la suma cuadrática de la regresión (diferencia entre los valores predichos por la regresión y la media de la variable dependiente) y la suma cuadrática total (diferencia entre los valores reales de la variable dependiente y su correspondiente media). En esta regresión el valor de 0.16 indica que el estatus y la edad de los estudiantes dan cuenta del 16% de la variación de la prueba de matemáticas.

La siguiente regresión analizada incluye tres variables más de características diferentes a las que se han considerado hasta ahora. Se trata de incluir variables cualitativas (nominales) en la regresión. En este caso, debe distinguirse entre variables dicotómicas, como es el caso del género (mujer y varón) y la titularidad del colegio (público y privado), de otras variables con más de dos valores posibles, como es la primera lengua hablada por los sujetos, que puede adoptar 6 valores (castellano, catalán, vasco, gallego, valenciano y otros). La cuestión reside en considerar una de las categorías de cada

variable nominal (lo más recomendable es asignar este papel a la categoría de la variable en cuestión con una mayor frecuencia) como categoría de base, a la que se le atribuirá el valor 0. De este modo, si las variables son dicotómicas, bastará con introducir en la regresión la categoría no considerada como base. En el caso del género, si se toma como valor de base a las alumnas, en la regresión aparecerá la característica ser varón como contraste; en el caso de la titularidad del centro, si se toma como valor de base los colegios públicos, en la regresión solo aparecerían los colegios privados. Ahora bien, si se considera la primera lengua aprendida del estudiante como variable, también se tomará solo un valor como base, preferiblemente el "castellano", por ser el más frecuente. De este modo, habrá que introducir en la ecuación cinco variables: catalán, vasco, gallego, valenciano y otros.

Tabla 7.- Regresión múltiple de la puntuación en Matemáticas.

Variables	Simple	Con dicot.	Con inter.
Género=varón		15.50 ***	15.50 ***
Mes	0.73 ***	0.74 ***	0.74 ***
Estatus	33.32 ***	29.65 ***	30.52 ***
Colegio=privado		19.29 ***	19.34 ***
Estatus#privado			-2.67 ***
Lengua=catalán		17.07 ***	17.09 ***
Lengua=vasco		22.12 ***	21.87 ***
Lengua=gallego		5.92	6.08
Lengua=valenciano		7.09 ***	7.12 ***
Lengua=otra		-32.25 ***	-32.11 ***
Constante	488.03	473.16 ***	473.50 ***
Error residual	79.59	78.15	78.14
R^2	0.16	0.19	0.19

*** p<.001

Fuente: http://pisa2012.acer.edu.au/downloads.php

Una vez aprendido cómo se introducen las variables cualitativas en la regresión, habrá también que considerar cómo se interpretan los coeficientes. A fin de comprenderlo mejor, se parte de la constante de la regresión con las nueve variables del ejemplo propuesto: las dos cuantitativas (estatus y mes), las dos dicotómicas (género y titularidad del colegio) y las cuatro concernientes a la lengua materna. Esta segunda ecuación de regresión presenta una constante diferente de la primera, porque se han introducido siete nuevas variables. En consecuencia, el valor de la constante (473 puntos) se referirá a una estudiante nacida en el mes de diciembre, que vive en una familia de estatus medio (en la OCDE), de género femenino, estudiando en un colegio público y cuya lengua materna es el castellano. El parámetro o coeficiente relativo a género informa de la diferencia entre varones y mujeres. En este caso es positivo, porque por término medio y, manteniendo constante el resto de variables consideradas, los estudiantes varones puntúan 15 puntos más en matemáticas que las estudiantes. De igual modo, la variable privado (uno para este valor y 0 para los colegios públicos) refleja la diferencia de medias entre ambos tipos

de colegios. En concreto, los estudiantes de colegios de titularidad privada han puntuado 20 puntos más que los públicos en Matemáticas. Finalmente, habría que referirse a la lengua materna. En este sentido, los de lengua vasca tienen 22 puntos más que los de lengua castellana, los catalanes 17 puntos más, y los valencianos solo 7. En cambio, los que tiene otra lengua nativa, la mayoría inmigrantes de países distintos de Latinoamérica, puntuaron en matemáticas con 32 puntos menos.

Intencionalmente, se ha dejado de comentar el coeficiente de los que tienen como lengua materna el gallego. Ello ha sido debido a que es el único que no ha presentado resultados significativos, como así lo indica la ausencia de estrellas en la tabla. En realidad, no es significativo porque el error típico es casi tan alto como su correspondiente coeficiente. Como en muestras grandes es asumible que el coeficiente adopte una distribución normal, si el cociente entre el coeficiente y se error típico no supera el valor de 1.96, no se está en condiciones de rechazar la hipótesis nula, o dicho de otro modo, no puede afirmarse que haya diferencias significativas en la puntuación en Matemáticas entre quienes hablan en casa en castellano y quienes hablan en gallego.

Falta por introducir una complejidad en el análisis de regresión. Se trata de la posibilidad de trabajar con interacciones entre las variables independientes. Se da interacción cuando la relación entre dos variables es distinta según los valores de una tercera variable. En el ejemplo actual, se diría que la titularidad del colegio y el estatus del estudiante interactúan en el caso de que la influencia del estatus fuera distinta según el alumno estudiara en uno u otro tipo de colegio. Para generar un modelo de interacción en la regresión basta con generar una tercera variable que sea el producto de aquellas que se suponen que interactúan e introducirla en la regresión. El coeficiente obtenido medirá las diferencias en la influencia de una variable a medida que se produce un cambio de unidad en la tercera variable. En el ejemplo de la tabla, se ve que la influencia del estatus en la puntuación de matemáticas es de 30,5 puntos por unidad. Pero esta cantidad ahora solo es válida para los colegios públicos (el valor 0 de la titularidad). En el caso de los colegios públicos (el valor 1 de la titularidad) la influencia es menor en 2 puntos y medio, lo que quiere decir que por un cambio de una unidad en el estatus, las puntuaciones en matemáticas cambian solo 28,5 puntos. Esto podría explicarse por el hecho de que en las familias con más estatus los padres solo llevan a sus hijos al colegio en el caso de que les garantice un buen nivel. Por el contrario, la estrategia de las familias con bajo estatus es la de llevar solo a sus hijos más inteligentes a colegios privados.

Resumiendo, la regresión múltiple es un procedimiento que sirve para explicar la variación de una variable en función de otras siguiendo un modelo lineal. El procedimiento averigua si los predictores que se empleen influyen positivamente, negativamente o no influyen de modo claro. Esto se puede detectar a través de la significación de los coeficientes o parámetros de la regresión. Además de ello, el procedimiento incluye un estadístico que refleja la bondad de ajuste de la regresión o, dicho de otro modo, un indicador de cuánto pueden ajustarse los valores de la variable dependiente en función de los de las variables independientes.

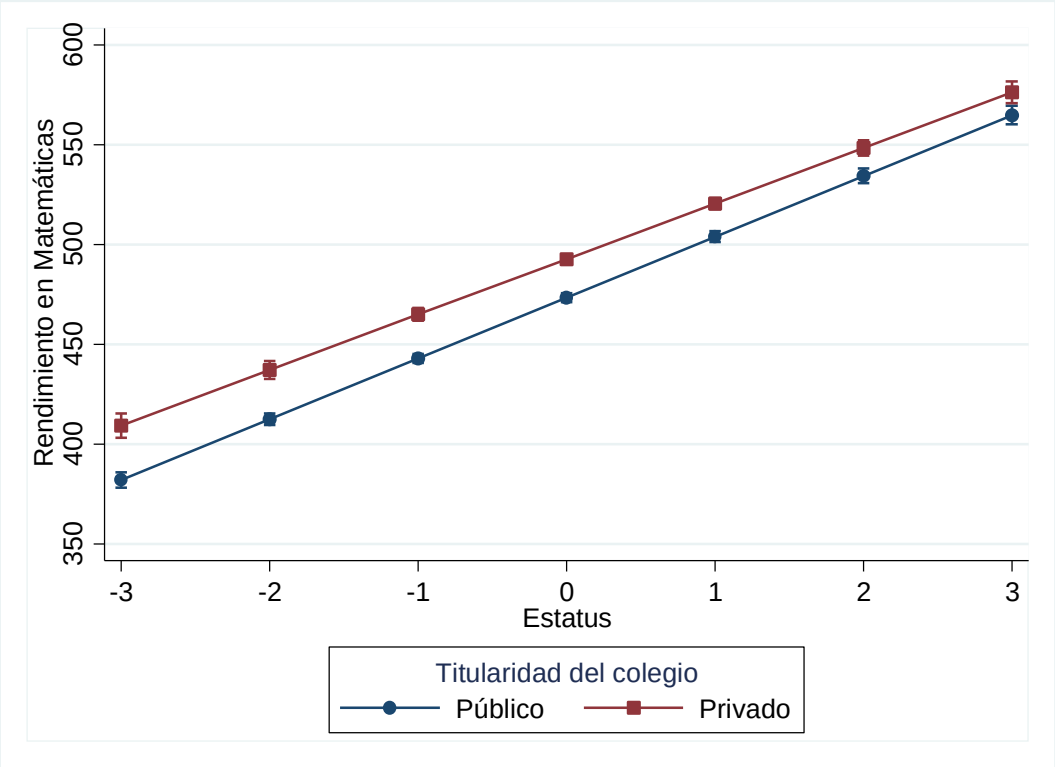

Gráfico 2.- Modelo de regresión de Matemáticas sobre estatus y tipo de colegio con interacción.

5.3 Modelos logit para las variables dependientes cualitativas

La regresión está diseñada fundamentalmente para variables de tipo cuantitativo. A pesar de ello, con ciertas precauciones también pueden introducirse como predictores variables de naturaleza cualitativa. Lo que no es procedente es que la variable dependiente sea cualitativa en una regresión múltiple. Sin embargo, hay un modo de realizar predicciones sobre variables nominales que actúa con una lógica similar a la de la regresión. Se trata de los modelos logit, los probit y los multi-logit, de los que aquí solo se verán los primeros.

Los modelos logit consisten en transformar una variable dependiente dicotómica (ocurre o no ocurre un determinado suceso) en una escala sin límites, mediante la expresión: ln(p/(1-p)). Esta expresión tiene el valor 0 en el caso de que p sea igual al 50% (0.5), es negativa si p es menor del 50% y positiva si supera esta cantidad. El procedimiento para hallar los coeficientes de los predictores es similar al de la regresión múltiple, si bien emplea el método de máxima verosimilitud, en lugar del de mínimos cuadrados. Una vez obtenidos, se pueden hacer predicciones de la variable dependiente siguiendo el modelo lineal, aunque en este caso, en lugar de predecir directamente el valor de la variable respuesta se obtiene su logit, que es bastante más complejo de interpretar que una proporción. Para hacerse una idea de lo que se está hablando, no está mal recordar que un logit de -3 equivale a un 5%, un logit de -2 equivale al 12% y uno de -1 al 37%. Obviamente, estos tienen sus equivalentes positivos en los porcentajes complementarios, es decir, +1 significa un 63%; +2 un 88% y +3 un 95%. En general, la fórmula para

obtenerlos es mediante la expresión: exp(x)/(1+exp(x)). La expresión de los modelos logísticos es, por tanto,

$$\ln \frac{P(\mathbf{Y}=1)}{P(\mathbf{Y}\neq 1)} = \mathbf{XB} + \mathbf{e}$$

Para entender mejor el procedimiento, se va a emplear la misma variable dependiente que se usó para la regresión múltiple convirtiéndola en una variable dicotómica que asume el valor 1 en el caso de que el estudiante haya sacado una puntuación en Matemáticas por encima de la media de la OCDE, es decir por encima de 500, que representa solo al 43.8% de los estudiantes que realizaron esta prueba en el España.

El resultado de la regresión logística con las mismas variables independientes, interacción incluida (tercera columna de la tabla 8) se presenta con una pauta similar a la regresión múltiple, esto es, mediante unos coeficientes con sus errores típicos y significaciones, así como una medida equivalente al R2, llamada pseudo-R2, basada en la verosimilitud del modelo comparada con la del modelo base. Como puede apreciarse, el pseudo-coeficiente de determinación (0,10) es menor que en la regresión múltiple. Aunque ambos coeficientes no sean del todo comparables, los modelos con variable dependiente continua suelen ser más precisos que aquellos que trabajan con variable discreta.

Tabla 8.- Regresión logística de buena puntuación en Matemáticas (>500)

Variables	Simple	Con dicot.	Con inter.
Género=varón		0.34 ***	0.34 ***
Mes	0.02 ***	0.02 ***	0.02 ***
Estatus	0.71 ***	0.65 ***	0.65 ***
Colegio=privado		0.40 ***	0.40 ***
Estatus#privado			-0.02
Lengua=catalán		0.50 ***	0.50 ***
Lengua=vasco		0.58 ***	0.58 ***
Lengua=gallego		0.23 *	0.23 *
Lengua=valenciano		0.06	0.06
Lengua=Otra		-0.47 ***	-0.47 ***
Constante	-0.22	-0.56 ***	-0.55 ***
Pseudo R^2 (McFadden)	0.08	0.10	0.10
Clasificados correctamente	63.1%	64.0%	64.0%

* p<.05; ** p<.01; *** p<.001
Fuente: http://pisa2012.acer.edu.au/downloads.php

Centrándose en los coeficientes, la constante se refiere al valor logit estimado para una estudiante nacida en diciembre, de familia de estatus medio, que asiste a clases en un colegio privado y habla castellano en casa. Una persona con estas características tiene una probabilidad del 36,6% (logit=-0,55) de tener una puntuación en matemáticas por encima de la media de los países de la OCDE. Todos los coeficientes son congruentes con los obtenidos en la anterior regresión: los varones, los nacidos en meses anteriores,

los de estatus alto y en colegio privado rinden mejor en Matemáticas. La influencia de la lengua es similar en los casos de estudiantes, siendo solo significativamente mayor la probabilidad de tener una puntuación en Matemáticas mayor de la media la de quienes hablan catalán o vasco en su familia y significativamente menor en el caso de los que hablan otra lengua no oficial en España. Sin embargo, a diferencia de la regresión anterior, la interacción entre estatus y titularidad del colegio no es significativa, en la medida en que el valor de p correspondiente al cociente entre el valor del coeficiente y su error típico no es menor que 5%.

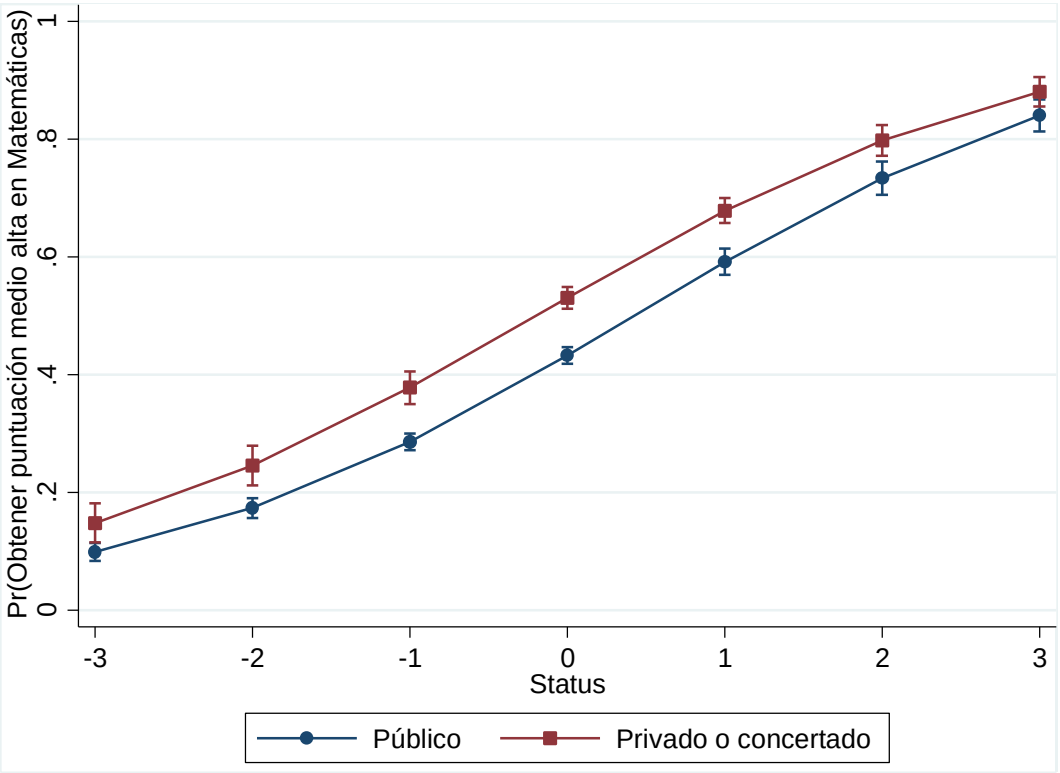

Gráfico 3.- Probabilidades de puntuaciones medias y altas en matemáticas según modelo logit.

5.4 Clasificación con variables cualitativas: el análisis de segmentación

Una alternativa a las regresión logística son los árboles de clasificación y regresión, también denominados análisis de segmentación, que constituyen un conjunto de técnicas que dividen jerárquicamente la muestra en un número indeterminado de segmentos homogéneos bajo un determinado criterio mediante la selección progresiva de sus mejores predictores entre un conjunto de variables candidatas. En consecuencia, se trata de una técnica de dependencia, al tenerse que determinar el criterio sobre el que se van a conformar las distintas subdivisiones de la muestra.

Como se deduce de la definición son diversos los procedimientos para realizar las segmentaciones. En este capítulo, solo se abordará el procedimiento CHAID, muy apropiado para el caso de variables dependientes nominales al emplear la métrica del chi2 explicada en el capítulo anterior. Las variables independientes de este procedimiento también convienen que sean nominales u ordinales, aunque obviamente también puede

emplearse un predictor cuantitativo, siempre y cuando se transforme en un conjunto finito de categorías.

El procedimiento consiste en el siguiente proceso: en primer lugar se comprueba si las distintas categorías de las variables predictoras son similares o no entre sí. Siguiendo el mismo ejemplo de la regresión logística en el que la variable dependiente era obtener una alta puntuación en Matemáticas (por encima de 500 puntos), si se dispone de un predictor con más de dos categorías, como es el caso de la lengua que se habla en la casa del estudiante, se comprueba si sus distintas categorías tienen un comportamiento similar en relación con la variable dependiente, en este caso, sería saber si hay un porcentaje parecido de buenas puntuaciones entre los que hablan catalán y los que hablan vasco por ejemplo, en cuyo caso se agruparían estas categorías. Como se aprecia en la tabla 9, no hay diferencias significativas entre quienes hablan en casa vasco y catalán, como tampoco las hay entre quienes lo hacen en valenciano y en gallego. Una vez fusionados estos dos pares de categorías, no se pueden seguir combinando, porque hay diferencias significativas entre cualquier par de valores (2º paso).

Tras haber agrupado categoría similares dentro de los posibles predictores, tiene lugar el proceso más importante de la segmentación: la división de la muestra, que se realiza mediante la selección del mejor predictor entre el conjunto de variables candidatas. En este caso, entre género, mes de nacimiento, estatus de la familia, titularidad del centro, o lengua hablada en casa, se trata de averiguar cuál de ellas ayuda más a discriminar a los estudiantes con buenas puntuaciones.

Tabla 9.- Fusión de categorías de la variable lengua hablada en casa según semejanza en puntuaciones en Matemáticas.

Matemáticas	Vasco	Catalán	Castellano	Valenciano	Gallego	Otros
<500	37.5	41.2	49.9	55.3	58.6	73.4
>=500	62.6	58.8	50.1	44.7	41.4	26.6
n	1084	1188	20794	150	565	1532
Prueba	χ^2 =3.26; p=0.07			χ^2 =0.51; p=0.47		

2º paso

Matemáticas	Vasco-Cat.	Castellano	Valen-Gall.	Otros
<500	39.9	49.9	57.9	73.4
>=500	60.61	50.1	42.1	26.6
n	2272	20794	715	1532
Prueba	χ^2 =90.90; p=0.00		χ^2 =17.60; p=0.00	χ^2 =53.99; p=0.00

Como se refleja en el gráfico 4, la variable por la que se segmenta en primer lugar la puntuación medio-alta en Matemáticas es el estatus. Esta variable, al ser cuantitativa, ha sido dividida en tres grupos de igual tamaño: el primero (con estatus por debajo del -0,6) solo presenta un 32,8 de estudiantes con más de 500 puntos en la prueba de Matemáticas en PISA; el segundo (con un estatus entre las puntuaciones de -0,6 y +0,4) se caracteriza por tener un 48,7% de estudiantes por encima de la media de la OCDE, y el tercero (con

estatus mayor de 0,4) por ofrecer un porcentaje de alumnos de este tenor del 67,7%. Es, por tanto, clara la influencia del estatus familiar en el rendimiento en matemáticas.

Una vez que se han conformado estos tres nodos o grupos, el análisis de segmentación prosigue realizando divisiones sobre estos grupos con el resto de variables. En este concreto ejemplo, ocurre que en cada uno de los grupos de estatus formados aparece una variable diferente de segmentación:

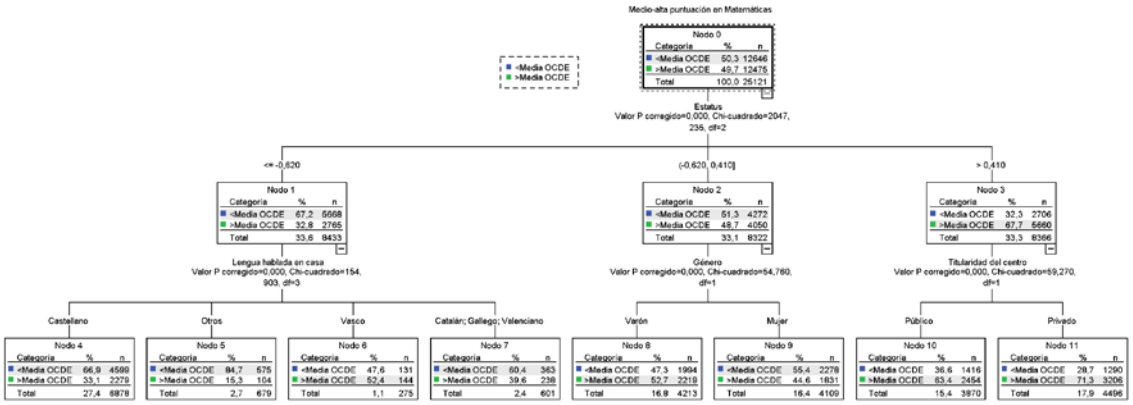

Gráfico 4.- Árbol de la nota en Matemáticas según estatus, lengua, género y titularidad.

Entre los de estatus bajo (nodo1), la variable que segmenta es la lengua que se habla en casa: en este sentido, los vascos de bajo estatus tienen un 52% de probabilidad de estar por encima de la media de la OCDE; en cambio, solo el 15% de aquellos estudiantes en los que no se habla ninguna de las lenguas oficiales españolas puntuaron por encima del valor medio del 500. Es de notar en el nodo 7 del gráfico 3 cómo los estudiantes que en su casa hablan catalán, gallego y valenciano aparecen en el mismo grupo, debido a que no presentan pautas muy distintas entre ellos en las puntuaciones matemáticas.

En los estudiantes de estatus medio (nodo 2), la variable más diferenciadora es el sexo, pues con ella se distingue entre varones, con una probabilidad del 52,7% de encontrarse por encima de la media, y mujeres con una del 44,6%.

Finalmente, la variable que segmenta a los estudiantes de estatus alto (nodo 3) es la titularidad del colegio: si cursan en uno público, tendrán una probabilidad de sacar buena nota en Matemáticas del 63,4%; si, en cambio, están matriculados en uno privado o concertado, la posibilidad de sacarla es del 71%.

En conjunto, puede distinguirse un alto contraste entre los estudiantes de buena familia que estudian en un colegio privado (nodo 11), entre quienes un 71% han obtenido calificaciones por encima de la media, y los estudiantes que viven en entornos de bajo estatus y hablan una lengua distinta de las oficiales en el país (nodo 5), entre quienes solo un 15% han superado el valor promedio del estudio.

El análisis de segmentación implica una doble clasificación: la una proporcionada por las variables por las que se segmenta. Con el análisis realizado, se han generado 8 grupos o nodos finales distintos (del 4 al 11): Grupo de bajo estatus/habla castellano en casa, grupo

de bajo estatus/habla otros idiomas, grupo de bajo estatus/habla vasco, grupo de bajo estatus/habla catalán, gallego o valenciano, grupo de alumnos varones de estatus medio, grupo de alumnas de estatus medio, grupo de estatus alto asistentes a colegio privado y grupo de estatus alto asistentes a colegio público. Cada uno de ellos ofrece un porcentaje de buenas puntuaciones en matemáticas. En función de este porcentaje se puede hacer una segunda clasificación entre aquellos que son buenos y aquellos que no son buenos en Matemáticas. Para generarla, aquellos grupos con más del 50% de buenos en esta materia serán clasificados como tales; mientras que los que no alcancen esta proporción serán clasificados como no buenos. Dicho de otro modo, podríamos hacer una predicción clasificadora de la variable dependiente (ser bueno o no en Matemáticas) en función de las características reflejadas en las variables predictoras.

Consecuentemente, a los estudiantes de bajo estatus social, salvo en el caso de los que hablan en su casa vasco, podríamos pronosticarles que no serán buenos en Matemáticas. Lo mismo se podría decir de las estudiantes que se crían en un estatus medio, pues solo el 45% de ellas tienen puntuaciones por encima de los 500 puntos. En contraste, los restantes cuatro grupos (los dos de estatus superior, los varones de estatus medio y los de estatus bajo que hablan vasco) poseen más de la mitad de posibilidades de ser buenos en Matemáticas, por lo que serán cualificados así en una posible predicción.

El final de este proceso se puede presentar en la llamada tabla de clasificación. En ella, se cruza por un lado, los distintos valores empíricos de la variable dependiente; por el otro, los distintos valores pronosticados en función de su pertenencia a uno u otro de los grupos conformados por las variables independientes. En el ejemplo seguido se comprueba que, tras la segmentación, se califican correctamente el 63% de los sujetos. De ahí que la estimación del riesgo sea su complementario, esto es, el 37%. Esta cifra solo puede ser útil comparándola con la estimación del riesgo de partida, es decir, de la que dispondríamos sin contar con ningún predictor. En este caso, sería del 49,7%, que es el porcentaje de la(s) categoría(s) no modal(es), correspondiente a los alumnos y alumnas con buenas notas en matemáticas.

Tabla 10.- Tabla de clasificación del análisis de segmentación

	Pronosticado		
Observado	<500	>500	% correcto
<500	7815	4831	61.8
>500	4452	8023	64.3
Porcentaje	48.8	51.2	63.0

Fuente: http://pisa2012.acer.edu.au/downloads.php

En consecuencia, para una mejor evaluación de la segmentación, conviene realizar, además de la tabla de clasificación, la tabla donde se cruza los grupos segmentados con la variable de criterio (tabla 11). De esta tabla, que con toda seguridad será significativa, se puede obtener la V de Cramer (0,30) o la λ asimétrica (0,26) como coeficientes de asociación que evalúen la bondad de predicción del análisis efectuado.

Tabla 11.- Perfil en el rendimiento en matemáticas de los nodos terminales

	Estatus								
	Estatus bajo				Estatus medio		Estatus alto		
	Lengua				Género		Titularidad		
Matemat.	Otra	Castel.	Cat.Gal.	Vasco	Mujer	Varón	Público	Privado	Total
< OCDE	84.7	66.9	60.4	47.6	55.4	47.3	36.6	28.7	50.3
> OCDE	15.3	33.1	39.6	52.4	44.6	52.7	63.4	71.3	49.7
Total	679	6878	601	275	4109	4213	3870	4496	25121

χ^2 = 2290.39; p=0.00; V de Cramer=0.30; λ =0.26

6 Nuevos avances en análisis multivariable

No han quedado cubiertas en este capítulo todas las posibilidades del análisis multivariable. Solo como bosquejo cabría mencionar algunas de las más novedosas técnicas que se están abriendo paso en la literatura académica de las ciencias sociales. Entre ellas, en el márquetin se ha de mencionar la popularidad que está adquiriendo el análisis conjunto para ver los factores que intervienen en la valoración de los productos de consumo (Hair et al., 1999:407-454); en procedimientos exploratorios, bajo la influencia del análisis de los grandes datos, se están empleando cada vez más los bosques aleatorios (Breiman, 2001), como una extensión del análisis de segmentación o las máquinas de soporte vectorial (Betancourt, 2005) para el reconocimiento de pautas, y en el terreno del análisis de causas se están empleando con profusión la regresión multinivel (Cebolla, 2013), el análisis de la historia de acontecimientos (Bernardi, 2006) o las redes neuronales (Pérez Delgado y Martín Martín, 2003).

7 Bibliografía recomendada

CEA, M. Á. (2002): *Análisis Multivariable. Teoría y Práctica en la Investigación Social*. Madrid, Síntesis.

CUADRAS, C. M. (2014): *Nuevos Métodos de Análisis Multivariante*. Barcelona, CMC Editions.

DÍAZ DE RADA, V. (2002): *Técnicas de Análisis Multivariante para Investigación Social y Comercial. (Ejemplos Prácticos Utilizando SPSS Versión 11)*. Madrid, RA-MA.

GARCÍA FERRANDO, M. (2004): *Socioestadística: Introducción a la Estadística en Sociología*. Madrid, CIS.

HAIR, J. F. *et al.* (1999): *Análisis Multivariante*. Madrid, Prentice Hall.

MARTÍNEZ ARIAS, R. (1999): *El Análisis Multivariante en la Investigación Científica*. Madrid, La Muralla/Hepérides.

PEÑA, D. (2002): *Análisis de Datos Multivariantes*. Madrid, McGraw Hill.

SÁNCHEZ CARRIÓN, J. J. (ed.). (1984): *Introducción a las Técnicas de Análisis Multivariable*. Madrid, CIS.