MÉTODOS CUANTITATIVOS DE ANÁLISIS MULTIVARIANTE APLICADOS A LA EDUCACIÓN. REVISIÓN DE LOS MÉTODOS DE DEPENDENCIA

Alfredo Jiménez Palmero*; Carmen Palmero Cámara**; Juan Alfredo Jiménez Eguizábal***. Universidad de Burgos

RESUMEN

Reconocer la complejidad de la educación no significa tener que admitir la incapacidad para diseñar investigaciones que permitan reducir la incertidumbre de los procesos de toma de decisión en el ámbito educativo. El presente artículo aborda la solvencia epistemológica y el potencial innovador de los métodos cuantitativos de análisis multivariante en el estado actual y cuestiones más disputadas de la investigación educativa, revisando las áreas de aplicación pedagógica de los métodos de dependencia. Emergen, así, nuevas posibilidades metodológicas al servicio de una educación de calidad para todos en función de los valores que guían la convivencia en una sociedad plural.

Palabras clave: investigación educativa, métodos cuantitativos, análisis multivariante, métodos de dependencia, regresión.

QUALITATIVE APPROACHES OF MULTIVARIATE ANALYSIS APPLIED TO EDUCATION. REVIEW OF THE DEPENDENCY METHODS

ABSTRACT

Recognizing the complexity of education does not mean having to admit the inability to design research to reduce uncertainty in decision making processes in the educational field. This article addresses the epistemological soundness and innovative potential of quantitative methods of multivariate analysis at the present moment and the most contested issues of educational research, reviewing the areas of pedagogical application of the methods of dependency. In this way, new methodological opportunities emerge to serve a quality education for all based on the values that guide life in a pluralistic society.

Keywords: educational research, quantitative methods, multivariate methods, dependency methods, regression.

* Departamento de Administración y Dirección de Empresas. Universidad de Burgos. E-mail: ajimenez@ubu.es
** Departamento de Ciencias de la Educación. Universidad de Burgos. E-mail: cpalmero@ubu.es
*** Departamento de Ciencias de la Educación. Universidad de Burgos. E-mail: ajea@ubu.es
Recibido: 15-01-10 Aceptado: 19-02-10

1. Introducción

En un mundo estructurado por la complejidad, los sistemas educativos necesitan, tal vez más que en ninguna otra época anterior, la adopción de decisiones muy controvertidas en un contexto generalizado de incertidumbre. En este sentido y con toda pertinencia, la investigación educativa ha ido avanzando sensiblemente en la configuración y aplicación de diferentes métodos cuantitativos, también es justo subrayarlo cualitativos, cuya revitalización propicia las mejores condiciones para explorar y pilotar, frente a las amenazas del caos, la *poiesis* que exigen las actuales políticas educativas en torno a la acción intencional y proactiva de la educación, erigida en factor decisivo por el valor y la influencia que tiene en la renovación de los hombres y las sociedades.

Al igual que en otros muchos campos científicos, la investigación cuantitativa en educación debe perseguir la obtención de la mayor y más precisa información derivada de las variables disponibles, en aras a reducir esa incertidumbre que rodea la toma de decisiones, tanto por parte de los individuos como de las instituciones.

En la práctica totalidad de diseños de investigación, los fenómenos educativos analizados son complejos al influir en ellos gran número de variables que es necesario estudiar simultáneamente, por lo que resulta imprescindible recurrir a técnicas más depuradas que los métodos clásicos de estadística descriptiva –superando la presentación de resultados en términos de frecuencias y correlaciones– como son las técnicas multivariantes. El análisis de cada variable por separado, propio de los métodos clásicos, y muy usual hasta ahora en el ámbito educativo ha comportado estudios parciales y en muchos casos incluso conclusiones erróneas, al no considerar los efectos conjuntos.

Estos métodos multivariantes, que analizan dos o más variables de manera simultánea, permiten obtener una visión de conjunto, al tener en consideración la interacción de factores que pueda existir en el fenómeno estudiado, eliminando la información redundante y ofreciendo una variable no observable directamente capaz de representar conceptos abstractos.

Se pueden señalar entre las principales aportaciones de estos métodos: el gran volumen de información que se puede obtener al permitir analizar un número elevado de encuestas; la posibilidad de hacernos asimilables y comprensibles gran cantidad de datos con una mínima pérdida de información, en ocasiones incluso proporcionando una imagen gráfica; y la posibilidad de analizar simultáneamente la información obtenida sobre el fenómeno, considerando todos los factores que intervienen en él y permitiendo utilizar tanto variables continuas como nominales, ordinales o textuales (Abascal y Grande, 1989).

Tal y como se puede apreciar en la Figura 1, dentro de los métodos multivariantes se distinguen dos grandes categorías: por un lado los métodos de

interdependencia o descriptivos, en los que se recogen los métodos factoriales, que ofrecen información en forma de plano, las escalas multidimensionales y los métodos de clasificación que la suministran en forma de árboles. Por otro, los métodos de dependencia o explicativos abordan una o varias variables en función de las demás.

Nuestro objetivo consiste en realizar una contribución en el ámbito del diseño de investigaciones educativas mediante un rastreo de las áreas de aplicación educativa de los métodos multivariantes de dependencia, especialmente aquellos que estudian una única variable dependiente y que son ampliamente utilizados en ciencias sociales como la economía, la psicología o la administración de empresas así como en las ciencias naturales, con objeto de motivar su empleo por parte de los investigadores dedicados al ámbito de la educación.

Con el propósito de hacer visibles nuestras motivaciones básicas, nos parece necesario incluir, aunque sea de forma sumaria, una referencia mínima para destacar la importancia de aportar un marco de referencia a las investigaciones en marcha que se han generado a partir de los programas de Doctorado impartidos en Venezuela –sedes de Cabimas, Maracaibo, Puerto Ayacucho, Puerto Ordaz y San Cristóbal– en convenio entre las Universidades de Burgos y de Córdoba (España) y las Universidades del Zulia y Abierta de Venezuela.

La cuestión medular entonces es cómo diseñar una adecuada investigación educativa que más allá de los reduccionismos de muy diversa índole que afectan a los sistemas educativos, proporcione certidumbre a los procesos de toma de decisiones orientados a lograr una educación de calidad para todos. En este sentido, intentaremos mostrar cómo para superar las posiciones metodológicas reduccionistas que impiden la discusión, la confrontación y evaluación racional de diferentes procesos educativos, debemos recurrir al conocimiento de los referentes de significado y de acción que nos aportan los métodos cuantitativos de análisis multivariante aplicados de la educación.

Los argumentos aducidos, y otros muchos que podrían añadirse sin dificultad, justifican la singular relevancia de la reflexividad epistemológica y el intento de plantear y recualificar en profundidad los criterios y aplicaciones de los métodos de dependencia. Por ello, además de la presente introducción, el trabajo analiza en la sección segunda las distintas clases de variables que determinan el tipo de método de dependencia a emplear en la investigación; el punto tercero, tras la necesaria referencia al modelo de regresión lineal simple, se centra en el modelo de regresión lineal general; en el cuarto apartado se analizan modelos de regresión diferentes en función de las características de las variables dependientes e independientes a introducir en la investigación, mostrando como ejemplos la regresión de Poisson o la regresión binomial negativa; la sección quinta se adentra en los modelos que emplean variables dependientes no métricas; la sección sexta profundiza en las ecuaciones simultáneas a emplear en caso de que la relación de causalidad entre la variable

dependiente y una o más de las independientes tenga una naturaleza circular y se finaliza con las conclusiones y las referencias bibliográficas.

2. Tipos de variables

El primer paso al diseñar una investigación de carácter cuantitativo multivariante, consiste en la planificación de las variables independientes –también llamadas predictoras, explicativas o regresores– que pensamos que afectan a la variable dependiente –a explicar, de respuesta, predicha o regresando– que queremos estudiar y cuya relación se verificará comprobando la hipótesis que se hayan formulado. Además de las variables independientes, hay que tener en cuenta que puede ser conveniente incluir en el modelo algunas variables de control, que son aquellas que también afectan a la variable dependiente pero de las que no se van a derivar ninguna hipótesis en la investigación que se está elaborando. La inclusión de dichas variables de control debe venir justificada por otros trabajos relevantes que hayan demostrado la influencia de dichas variables en la variable dependiente. Esto también suele ser aconsejable para las variables independientes, si es que existen trabajos anteriores que las hayan empleado en situaciones similares, explicando las diferencias entre dichos estudios y el que se está elaborando, para resaltar la contribución que se pretende lograr.

Tanto las variables dependiente, independientes como las de control, pueden clasificarse según su naturaleza en variables cuantitativas o escalas métricas y variables cualitativas o escalas no métricas o de orden. Las primeras se subdividen a su vez en numéricas, cuando tienen un origen, el cero, con sentido real e invariable, como sucede en el caso de los ingresos de las familias, los costes de la educación, el número de profesores y de alumnos, entre otros y de intervalos cuando no existe un cero natural como en el caso paradigmático de la temperatura, donde el cero no implica ausencia de ésta. En ambos casos, la distancia entre dos valores mantiene el significado para cualquier par, es decir, la distancia de 3 a 5 tiene el mismo valor que entre 7 y 9. Además, permiten emplear casi todas las operaciones estadísticas (media, desviación, correlaciones, test paramétricos, etc.) pero en las de intervalos, la relación entre A y B no es independiente de la unidad de medida, si B= 2A no implica que B posea necesariamente el doble que A de la característica estudiada (Abascal y Grande, 1989).

Por lo que respecta a las segundas, las variables cualitativas o escalas de orden, se subdividen en escalas ordinales y nominales. En las primeras se jerarquizan los objetos colocándolos en un orden relacionado con el grado que poseen de la variable medida, por ejemplo al codificar las preferencias de los padres sobre tipos de educación o preferencias de centro docente, elección de asignaturas optativas, ciclos o programas educativos. En este caso la distancia entre valores no tiene por qué significar lo mismo en cualquier par, sino que

simplemente señala un orden de preferencia. También permiten operaciones estadísticas, aunque más limitadas, como por ejemplo la mediana o el coeficiente de correlación de rangos.

Las nominales, por su parte, se asocian a categorías o conjuntos mutuamente excluyentes, en los que no existe una relación de orden sino únicamente se expresa la pertenencia. Cuando sólo existen dos categorías, como por ejemplo en el sexo, en la aplicación o no de un determinado recurso de enseñanza, puede ser el caso de enseñanza a distancia o la utilización de e-learning, la variable se denomina dicotómica o binaria y se suele codificar con un 0 y un 1 para su operativización en programas informáticos. Cuando existen diversas categorías, como sucede en los tipos de dirección escolar, de gestión educativa o la selección de una determinada Universidad donde estudiar o de un determinado país donde realizar el postgrado o la especialización, la variable se considera multinómica, asignando números en función de la cantidad de categorías existentes.

En función de qué tipo de variables se dispongan, el investigador deberá seleccionar la técnica concreta conveniente para su investigación. En ocasiones, más de una técnica puede ser aplicada, por lo que puede resultar interesante ofrecer los resultados de todas ellas y realizar una comparación. Cada una de las técnicas tiene una serie de requerimientos específicos que se deben cumplir, pero en cualquier caso siempre hay que tener en cuenta que no debe existir multicolinealidad entre las variables independientes y de control, por lo que se debe verificar la tabla de correlaciones bivariadas entre todas ellas. También es aconsejable comprobar que los Variance Inflation Factor (VIFs) no superen el límite de 10 aconsejado por Neter et al. (1985), Kennedy (1992) y Studenmund (1992) o mejor incluso el de 5,3 recomendado por Hair et al. (1999).

3. La regresión lineal

La primera técnica de dependencia que se suele afrontar, a pesar de que sus limitaciones y diversos supuestos de partida previos impiden a menudo su empleo, es la regresión lineal. Al igual que el resto de técnicas que se mostrarán a lo largo del trabajo, la regresión lineal pretende analizar la dependencia de una variable a explicar con respecto a una o más variables explicativas, con el objetivo de determinar la estructura o forma de la relación entre las variables independientes con la dependiente, verificar las hipótesis derivadas de la teoría analizada y predecir los valores de la variable dependiente.

El modelo de regresión lineal simple se expresa de la siguiente forma:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Donde Y_i es la variable dependiente, X_i representa la variable independiente, $_0$ y $_1$ los parámetros a estimar y $_1$ la perturbación aleatoria. Cuando la

regresión lineal simple se generaliza al contexto multivariante, la expresión se formula así:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Donde en lugar de existir una variable independiente única existen p variables independientes que se relacionan con la variable dependiente.

Normalmente se emplea una técnica de estimación denominada Mínimos Cuadrados Ordinarios (MCO), u Ordinary Least Square (OLS) empleando el término en inglés, ya que ofrece estimadores eficientes e insesgados. Consiste en elegir la recta de regresión o plano de regresión de forma que se minimicen las distancias (medidas en el eje de la variable dependiente) entre el plano y los puntos observados (Fernández, 2005). Sin embargo, existen otros métodos de estimación que se emplean cuando alguno de los supuestos de partida de los MCO no se cumplen, como por ejemplo el de máxima verosimilitud, el de Mínimos Cuadrados Generalizados o el de Mínimos Cuadrados Ponderados cuando se quiere asignar un mayor peso a un dato en particular.

En concreto, los MCO necesitan que se cumplan las siguientes hipótesis básicas de partida:

1. Se supone que la forma funcional que liga la variable explicada con las variables explicativas es de tipo lineal al menos en los parámetros. En caso de que la forma sea circular, se deben emplear las ecuaciones simultáneas.

2. Las variables explicativas son fijas en el muestreo o al menos serán independientes de las perturbaciones. Éste es un requisito que se puede extender a todos los demás métodos cubiertos por este trabajo y consiste en evitar que exista algún tipo de relación lineal entre las variables explicativas del modelo. En caso de concurrir, algunos coeficientes tenderán a perder significatividad.

Para detectar, se deben revisar las correlaciones entre parejas de regresores, los Variance Inflation Factor (VIFs) o regresiones auxiliares para observar el Factor de Agrandamiento de la Varianza y el Índice de Condición. Como posibles soluciones a este problema, se puede intentar ampliar el tamaño de la muestra, eliminar aquellas variables más colineales con el resto o tomar como variable explicativa el resultado de componentes principales entre las variables explicativas colineales.

En todo caso, la multicolinealidad no es necesariamente perjudicial si el objetivo de la investigación es únicamente la predicción. Sin embargo, sí lo si se pretenden realizar tests de hipótesis.

3. Las perturbaciones aleatorias son heterocedásticas, es decir tienen una

varianza constante. Este problema suele aparecer si en la muestra se incluyen centros educativos muy grandes y muy pequeños, alumnos muy jóvenes y alumnos adultos como en el acceso de personas mayores a la Universidad o en los casos de estudios comparados entre países con una gran diferencia en la financiación de los estudios. En caso de homocedasticidad, las pruebas de significancia t y F dejan de ser válidas y de aplicarse pueden conducir a errores.

Para detectarla se puede recurrir a métodos gráficos de residuos, la prueba Goldfeld-Quandt o la prueba de Park y para solventarla se debe utilizar la estimación por Mínimos Cuadrados Generalizados o por Mínimos Cuadrados Ponderados.

4. La covarianza de las perturbaciones es distinta de cero, es decir el término de perturbación relacionad con una observación cualquiera no está influido por el término perturbación relacionado con cualquier otra observación. En caso contrario se presenta un problema de autocorrelación que provocará que las estimaciones continúen siendo lineales e insesgadas pero ineficientes, de nuevo las pruebas de significancia t y F dejan de ser válidas y de aplicarse pueden conducir a errores y cabe la posibilidad de que un coeficiente aparezca no significativo cuando en realidad lo sea.

Para detectarla se suelen utilizar los gráficos de residuos, el test de Durbin-Watson y el contraste de Beusch-Godfrey, resultando aconsejable emplear la estimación por Mínimos Cuadrados Generalizados si se observa este problema.

Además, la variable dependiente ha de ser métrica y debe poder tomar todo tipo de valores numéricos, tanto negativos como positivos así como con decimales. Aunque diversas variables los cumplen –beneficios empresariales, temperatura, saldos de la balanza de pagos de un país– estos requerimientos impiden que esta técnica sirva para estudiar muchos fenómenos relacionados con el ámbito educativo (con las excepciones de incrementos y disminuciones de las calificaciones de un año a otro, inversiones y desinversiones en el sistema educativo o los beneficios de entidades educativas privadas). En estos casos es necesario acudir a otro tipo de técnicas multivariantes de dependencia como las que se describen en los apartados subsiguientes.

4. Otros tipos de regresiones: regresión de poisson y regresión binomial negativa

En determinadas ocasiones, la variable dependiente que se desea estudiar será métrica, pero con algunas especificidades. De hecho, resulta muy habitual que sea imposible que tome valores negativos y que siempre tome valores enteros, como sucede al querer analizar el número de escuelas en una localidad, o el número de profesores en un instituto, región o Estado. En este

caso la regresión de Poisson resultaría más apropiada que la tradicional regresión por mínimos cuadrados (García-Canal y Guillén, 2008). Sin embargo, la variable dependiente podría sujeta a sobredispersión, en cuyo caso se necesitaría emplear el modelo binomial negativo, que consiste en una generalización del modelo de Poisson en el que se relaja el supuesto de igual media y varianza (Hausman et al. 1984; Cameron y Trivedi, 1998).

Para verificar qué modelo debe emplearse, se puede recurrir al test de bondad del ajuste, para analizar la posible sobredispersión de la variable dependiente. También se puede analizar el ratio *"likelihood ratio test"*, que analiza la sobredispersión del parámetro alpha. Cuando la sobredispersión del parámetro es cero, la distribución binomial negativa es equivalente a la de Poisson. Sin embargo si alpha es significativamente diferente de cero, la regresión de Poisson no resultaría apropiada.

Estas técnicas, empleadas con resultados brillantes en la investigación relacionada con la internacionalización de empresas, en aquellos casos en que la unidad de medida son empresas o filiales (García-Canal y Guillén, 2008; Jiménez, 2008) han sido analizadas con gran profundidad en sus fundamentos estadísticos por Maddala (1993) y son transferibles a la investigación en materia educativa, especialmente fecundas en la evaluación de la efectividad de estrategias educativas o en las relacionadas con la planificación educativa de recursos físicos –centros docentes–, humanos –necesidad de profesores– y económicos –inversiones educativas–.

5. Métodos con variables dependientes no métricas

Cuando la variable dependiente es cualitativa, no es posible emplear los modelos de regresión hasta ahora descritos. Si todas las variables independientes son cuantitativas se puede emplear el análisis del discriminante, aunque resulta mucho más frecuente que la regresión logística ya que permite que las variables independientes sean cualquier combinación entre cualitativas y/o cuantitativas, y permite analizar fenómenos en los interesa distinguir entre dos categorías mutuamente excluyentes, como la decisión de invertir o no en un país (De la Fuente et al. 2008) o si optar por la creación de una empresa en propiedad o una alianza con un socio local al introducirse en un determinado mercado (Slangen y van Tulder, 2009). En el caso educativo, sirven de ejemplos la decisión de ir o no a la Universidad, implantar o no una determinada titulación, o la posible apertura de un aula de informática en un centro educativo.

Esta técnica que ya comenzó a ser empleada a mediados del siglo XX se convierte en los años 60 en un método ampliamente utilizado en el análisis de regresión de datos dicotómicos, principalmente en las ciencias de la salud (Hosmer y Lemeshow, 1989 así como Cramer, 1991 ofrecen una amplia bibliografía al respecto), para posteriormente expandirse en campos como las ciencias empresariales, la sociología o las ciencias de la educación, donde

puede ser empleada en estudios sobre el comportamiento de profesores, el análisis del éxito de nuevos métodos educativos o de un determinado material de apoyo docente, el fracaso escolar, la predicción de situaciones de demanda educativa sobre la base de ratios sociodemográficos, la planificación de las universidades de una región/Estado o la predicción de actividades de I+D en universidades innovadoras.

El fundamento de la regresión logística se basa en la estimación de los parámetros por el procedimiento de máxima verosimilitud, ya que el procedimiento típico de las regresiones lineales de mínimos cuadrados (en el que se minimiza las diferencias al cuadrado entre los valores reales y los predichos de la variable dependiente) presenta diversos inconvenientes (Luque Martínez 2000):

En primer lugar, se viola la asunción de normalidad en los errores, de manera que los estimadores minimocuadráticos no serán eficientes. Este problema sin embargo minimiza su importancia en muestras grandes.

En segundo lugar, el término error tampoco cumple el supuesto de homocedasticidad en su varianza, sino que por el contrario presenta heterocedasticidad. También este problema se podría solucionar mediante transformaciones o estimación en dos etapas.

En tercer lugar la hipótesis de normalidad de la variable dependiente tampoco se cumple cuando es dicotómica.

Finalmente, y como inconveniente más grave, no se cumple el requisito de que las probabilidades estimadas para cualquier valor de las variables independientes se encuentren entre 0 y 1, lo que de manera provoca que no se pueda utilizar el método de mínimos cuadrados en este caso.

Incluso esta limitación podría resolverse mediante mínimos cuadrados restringidos o programación cuadrática pero aun entonces continúa presente el problema fundamental de asociar que la probabilidad de la variable dependiente aumenta linealmente con la independiente, es decir que el efecto marginal permanece constante (Gujarati 2003). En realidad, cuando en este tipo de variables dependientes se producen pequeños incrementos en los extremos de la distribución de la variable independiente la probabilidad de la variable dependiente no se ve prácticamente afectada. Aldrich y Nelson (1984) sostienen que la relación entre p_i y X_i debe ser no lineal, es decir, "uno se acerca a cero a tasas cada vez más lentas a medida que X_i se hace más pequeño y se acerca a uno a tasas cada vez más lentas a medida que X_i se hace muy grande" e indican que emplear mínimos cuadrados cuando la variable dependiente es cualitativa ofrece escasas ventajas y numerosos inconvenientes, ya que de antemano se sabe que las asunciones del modelo no se van a poder mantener, o lo que es igual, se especifica incorrectamente la relación entre las variables indepen-

dientes con la variable dependiente.

También prestan atención a un pequeño inconveniente que el modelo logit comparte con el modelo probit, consistente en la imposibilidad de resolver de manera algebraica las ecuaciones de verosimilitud debido a que son no lineales para los parámetros que se deben estimar, lo que fuerza a emplear algoritmos iterativos para lograr aproximaciones. Sin embargo subrayan que los estimadores obtenidos con esta técnica cuando la muestra es grande mantienen todas las propiedades que tienen los obtenidos mediante la regresión lineal, ya que son insesgados (los estimadores se encuentran centrados alrededor de los verdaderos valores), eficientes (ningún otro estimador insesgado tiene una varianza menor) y normales (por lo que se pueden realizar tests de hipótesis y obtener otras inferencias). Otro inconveniente demostrado por Robinson (1982) consiste en que cuando los residuos están serialmente correlacionados, los estimadores por máxima verosimilitud se mantienen insesgados pero ya no son eficientes, al igual que sucede en los modelos de regresión lineal. Desafortunadamente, no se han logrado correcciones para la correlación serial en los modelos logit ni probit (Aldrich y Nelson, 1984).

La regresión logística nos permite emplear variables independientes cualitativas, lo que configura, junto a su mayor robustez cuando no se cumplen los supuestos sobre normalidad y homocedasticidad, las ventajas de esta técnica frente al análisis discriminante. Incluso algunos investigadores la prefieren por su similitud con la regresión en el sentido de que permite contrastes estadísticos directos, capacidad para incorporar efectos no lineales así como una amplia variedad de diagnósticos (Hair y otros 1999).

El modelo por lo tanto se basa en la utilización de la función logística, ya que es no lineal, monótona y creciente, y acotada entre 0 y 1. La regresión logística en definitiva consiste en aplicar esta función para modelizar la relación de probabilidad de Y=1 condicionada a un determinado valor de las variables independientes, empleando la estimación por máxima verosimilitud, que proporciona unos valores para los parámetros desconocidos que maximiza la probabilidad de que con ellos se obtengan los valores observados (Hair y otros 1999).

Así el modelo se puede expresar:

$$p_i \quad E(Y \quad 1/X_i) \quad \frac{e^{(\quad 1 \quad 2X_i)}}{1 \quad e^{(\quad 1 \quad 2X_i)}}$$

Y su generalización a un contexto multivariante:

$$p_i \quad E(Y \quad 1/X_i) \quad \frac{e^{(\quad 0 \quad 1X1i \quad 2X2i \quad \ldots \quad nXni)}}{1 \quad e^{(\quad 0 \quad 1X1i \quad 2X2i \quad \ldots \quad nXni)}}$$

Donde X_i un vector que recoge las características individuales del decisor *í*, es decir, las variables independientes, y β es un vector de parámetros.

En este tipo de investigaciones suele ser aconsejables mostrar, como medida de la significación global del ajuste, los coeficientes de regresión de la prueba ómnibus junto con medidas de la bondad del ajuste, tales como la devianza (-2 log de verosimilitud) y la tabla de clasificación de eficacia predictiva. En esta técnica no se emplea el coeficiente de determinación R^2 tan habitual en la regresión lineal, ya que no existe un equivalente directo. Se han propuesto diversas alternativas, conocidas como pseudos R^2, pero todas ellas sufren de falta de aceptación general entre los investigadores y ninguna tiene una interpretación simple como sí sucede con el R^2 (Aldrich y Nelson, 1984; Luque Martínez, 2000).

Existen modelos incluso más avanzados, que se emplean cuando la variable dependiente resulta de una elección discreta entre diversas alternativas. En este caso se debe emplear el Conditional Logit Model (CLM) de McFadden (1984). En este modelo los coeficientes se estiman mediante el procedimiento de máxima verosimilitud y se mantiene la propiedad de independencia de alternativas irrelevantes (conocida como IIA en sus términos en inglés). Esto quiere decir que la probabilidad de elegir la región j sobre la alternativa i, dada su probabilidad condicional, depende únicamente de las características de las dos alternativas y no de una tercera posible elección. Debido a esta última propiedad, las diferentes alternativas deben ser comparables en términos de sustitución.

Otro modelo más avanzado es el Nested Logit Modelo (NLM), en el que una primera decisión, como por ejemplo elegir en qué continente va a invertir una empresa, condiciona una segunda selección, como puede ser el país concreto en que se va a localizar (Durán et al. 2008). Esta estructura de decisión en forma de árbol con dos niveles puede ser estudiado gracias al NLM. Para distinguir cuándo se debe emplear el CLM y cuando el NLM, se recurrirá a la prueba de Hausman (1978) y al valor del *inclusive value* en el NLM. En el primer caso si el p-valor resulta bajo querrá decir que la asunción de independencia de alternativas irrelevantes no se sostiene por lo que será necesario recurrir al NLM. En el segundo, el valor deberá encontrarse entre 0 y 1. En caso de que sea 1 o una cifra cercana, el modelo CLM podría resultar adecuado ya que las alternativas de la primera elección se pueden considerar como sustitutos equivalente, mientras que cuanto más cercano a 0 únicamente la primera elección resulta relevante, al considerarse sustitutos equivalentes todas las opciones de la segunda elección. Cuando el valor del *inclusive value* se encuentra entre medias, el NLM resulta el modelo correcto.

Una descripción más detallada de ambos modelos en lo relativo a sus desarrollos matemáticos, fórmulas y asunciones básicas puede ser encontrado

en McFadden (1984), Cramer (1991), Maddala (1993), Mayer y Mucchielli (1999) y Disdier y Mayer (2004).

6. Ecuaciones simultáneas

En ocasiones, algunos fenómenos no se caracterizan por una circularidad en la causalidad. Así, por ejemplo, se puede argumentar que los beneficios empresariales son un factor determinante de la imagen de una empresa, pero a su vez también es cierto que la imagen de la empresa juega un papel relevante en los beneficios obtenidos por ésta (De Quevedo y De la Fuente, 2003). De igual manera, el mayor empleo de servicios privados para el cuidado de hijo viene determinado por la participación femenina en el mercado laboral, pero también el florecimiento de este tipo de servicios en una determinada región favorece la incorporación de la mujer a un puesto de trabajo (Herbst y Barnow, 2008). Otro ejemplo podría consistir en que cuanto mayor sea el nivel de comercio exterior de un país (medido por ejemplo a través de sus exportaciones e importaciones), mayor será la cantidad de inversión directa del exterior que recibirá. A su vez, cuanto mayor sea el nivel de inversión directa en el exterior recibida en un país, mayor su crecimiento económico. Finalmente, el mayor nivel de crecimiento económico de un Estado influye positivamente en su nivel de comercio exterior (Metwally, 2004).

Esta tipología de relaciones se puede extrapolar al ámbito educativo. Así, por ejemplo, podemos sostener que la demanda educativa que recibe una escuela o una universidad es un factor determinante de la imagen de ese centro docente, pero a su vez también es cierto que la imagen de una universidad o de una escuela, pensemos por ejemplo en Harvard, juega un papel relevante en la inducción de la demanda de acceso a ese centro. Otro ejemplo podría consistir en que cuanto mayor sea el nivel de estancias formativas en el extranjero de los universitarios de un país (medido a través de los universitarios que realizan postdoctorado en el extranjero y los que vienen a estudiar al país), mayor será la cantidad de gestión de conocimiento e innovación del exterior que recibirá, lo que resulta muy perceptible en los campos tecnológicos, biosanitarios y educativos. A su vez, cuanto mayor sea el nivel de conocimiento y de técnicas exteriores recibida en un país, mayor su progreso y crecimiento educativo. Finalmente, el mayor nivel de progreso educativo de un Estado influye positivamente en su nivel de internacionalización y conexiones con el extranjero.

En aquellos casos en que se sospeche la posibilidad de simultaneidad en la causalidad entre dos o más variables, se debe emplear el modelo de ecuaciones simultáneas. Si no existiera simultaneidad, la estimación por método de mínimos cuadrados ordinarios obtendría estimadores consistentes y eficientes, mientras que en su presencia no resultan ni tan siquiera consistentes, debiendo recurrir a métodos que permitan abordar la endogeneidad. Sin embargo, en

caso de aplicarlos en ausencia de endogeneidad, los estimadores así obtenidos resultarían consistentes pero no eficientes, por lo que se preferiría el método de mínimos cuadrados (Gujarati, 1997).

Por tanto, para verificar la endogeneidad y justificar así el empleo de ecuaciones simultáneas, debe aplicarse la prueba de especificación de Hausman. Dicha prueba consta de dos pasos (Maddala, 1996; Gujarati, 1997).

En el primero se obtienen las ecuaciones de forma reducida para los regresores que se cuestionan como endógenos, es decir, obteniendo estas variables sólo en términos de las variables predeterminadas y las perturbaciones estocásticas únicamente, con objeto de obtener sus valores previstos.

En el segundo se estima una segunda ecuación que se corresponde con la original del modelo pero a la que, además, se le añaden como variables explicativas las previsiones obtenidas en la fase uno de las variables que se cuestiona su endogeneidad. Mediante la significatividad de la prueba F, se puede verificar si efectivamente los regresores presentan endogeneidad, lo que conllevaría recurrir a técnicas de ecuaciones simultáneas, o por el contrario exogeneidad, obligando al empleo de estimadores a través de mínimos cuadrados ordinarios.

Es necesario también verificar que los parámetros del modelo son estimables, para cual se deben cumplir tanto la condición de orden como la de rango. Con respecto a la primera, el número de variables predeterminadas excluidas en una determinada ecuación debe ser, al menos, tan alto como el número de variables endógenas incluidas en dicha ecuación menos uno (Gujarati, 1997).

En cuanto a la segunda, debe ser posible construir un determinante diferente de cero, de orden (M-1)x(M-1), donde M es el número de ecuaciones y variables endógenas, a partir de los coeficientes de las variables (endógenas y predeterminadas) excluidas de esa ecuación articular, pero incluidas en las otras ecuaciones del modelo (Gujarati, 1997).

Una vez superadas ambas condiciones, debe seleccionarse el método concreto de estimación, aunque normalmente resulta recomendable el método de información completa “3 stage least square” (3SLS) ya que, aunque resulta más sensibles a errores en los datos o en la especificación de las ecuaciones, realiza la estimación de manera conjunta para todos los parámetros del modelo, en lugar de ecuación a ecuación, por lo que preserva mejor el objetivo que se persigue con las ecuaciones simultáneas que la simple estimación aislada de cada ecuación (De Quevedo y De la Fuente, 2003). Además, se trata de una alternativa mejor que otras como los modelos de información limitada [por ejemplo el “2 stage least square” (2SLS)] ya que no pierde eficiencia cuando

existe correlación entre los errores de las diferentes ecuaciones del modelo (Cho, 1998; Kim, 2007). Aún así, puede resultar interesante, y de hecho algunos trabajos así proceden, ofrecer los resultados de dos o más métodos como prueba de robustez.

7. Conclusiones

El presente trabajo ha pretendido contribuir a explorar las aplicaciones educativas de diversas técnicas cuantitativas de análisis multivariante, con especial atención a los métodos de dependencia. El investigador deberá seleccionar aquella técnica que resulte conveniente, en función del tipo de variables a su disposición en el análisis de un fenómeno concreto, respetando los requerimientos y supuestos que el método exija, para evitar obtener conclusiones sesgadas, incompletas o incluso incorrectas.

Aún existen otras técnicas adicionales, como el análisis de la varianza (tanto univariante o ANOVA como multivariante o MANOVA) y convarianza (ANCOVA Y MANCOVA respectivamente), el modelo probit, las ecuaciones estructurales o los datos de panel. Especialmente relevante es esta última, al permitir introducir la variable tiempo en el análisis, que bien podría aplicarse, por ejemplo, a la evolución de la matrícula en las Universidad o centros docentes. Además, permite superar algunas limitaciones de las regresiones por mínimos cuadrados.

Así, Temple (1999) afirma que este tipo de regresiones pueden adolecer de errores de medida, heterogeneidad de los parámetros y pérdida de información dinámica relevante. Gyimah-Brempong y Traynor (1999), Tsangarides (2001) y Jiménez (2009), en estudios relacionados con inversiones y África, señalan que los resultados obtenidos pueden ser inconsistentes y sesgados al no tomar en consideración la posible endogeneidad de los regresores, además de sufrir un posible sesgo por las variables omitidas, limitación que puede ser superada, incluso si la endogeneidad se da en la propia variable dependiente, empleando el estimador dinámico de datos de panel de Arellano y Bond (1991), también conocido como Método Generalizado de los Momentos (Generalized Method of Moments o GMM), aunque a cambio no permite incorporar variables explicativas que no varíen de un año a otro, ya que se trata de un modelo que se estima en primeras diferencias por lo que en realidad se toman los incrementos. En caso de querer incluir este tipo de variables, y siempre y cuando no existan problemas de endogeneidad, se puede recurrir a otros tipos de datos de panel tales como los de efectos fijos o los de efectos dinámicos en función de los resultados de la prueba de Hausman (1978), tanto para variables métricas como binarias y cuya operativización en los programas informáticos se pueden encontrar, por ejemplo, en los manuales de STATA (Statacorp, 2001).

Para utilizar estas técnicas, los investigadores contamos con la inestimable ayuda de los programas estadísticos diseñados para ordenador. Las

regresiones más sencillas así como toda la estadística descriptiva y el estudio de la multicolinealidad pueden llevarse a cabo con el programa SPSS. Sin embargo para técnicas más complejas como la regresión de Poisson, la regresión binomial negativa, el Conditional Logit Model, el Nested Logit Model o los datos de panel resulta necesario acudir a otros programas entre los que se pueden destacar STATA o E-VIEWS, para cuyos manejos existen manuales de usuario muy ilustrativos acerca del empleo y requisitos de cada una de las técnicas.

A través del análisis crítico de los diversos métodos que constituyen el cuerpo de este trabajo hemos pretendido justificar la singular relevancia de la reflexividad epistemológica del análisis multivariante en educación, precisando, con las aportaciones pertinentes en cada caso, cuándo, cómo y en qué condiciones puede aplicarse racionalmente como diseño de investigación.

En el reto de construir la estructura epistémica de la investigación educativa, nos encontramos con el corpus de conocimiento científico que nos proporcionan los métodos de dependencia, que son capaces de satisfacer en el conocimiento pedagógico, los recursos, las estrategias y las condiciones que ha de satisfacer la investigación educativa.. En consecuencia, las limitaciones reales en la tarea investigadora no son motivo de desesperación, sino que, por el contrario, nos impulsan a modificar su horizonte conceptual para aplicar nuevos criterios y ensayar soluciones críticas innovadoras y abiertas para responder a la pregunta inicial sobre si la investigación educativa puede hacer frente a las posibilidades y necesidades de la educación y de los sistemas educativos para conseguir el pleno desarrollo de la personalidad humana en el respeto a los principios democráticos de convivencia y a los derechos y libertades fundamentales, que no pierden su justificación, aunque no se pueda encontrar una fundamentación definitiva.

Éste bien puede ser el momento para avanzar en las dimensiones innovadoras del campo de conocimiento pedagógico, razonablemente creativo, en un mundo estructurado por la complejidad. El empleo de enfoques multimetodológicos, que combinen diferentes técnicas tanto cuantitativas como cualitativas, puede enriquecer en gran medida la discusión de resultados y las conclusiones que se pueden alcanzar en el ámbito educativo. En todo caso, somos más conscientes de que los métodos de dependencia, al margen de los cauces de la discusión convencional, contribuyen al alza de los niveles de reflexividad epistemológica exigibles en la investigación educativa.

Agradecimientos

Los autores agradecen la colaboración y sugerencias de la profesora María Isabel Landaluce Calvo. Asimismo, Alfredo Jiménez Palmero agradece el apoyo financiero del Ministerio Español de Ciencia e Innovación a través del programa FPU.

Referencias bibliográficas

Abascal, E. y Grande, I. (1989). Métodos multivariantes para la investigación comercial: teoría, aplicaciones y programación BASIC. 1ª Ed. Ariel Economía.

Aldrich, J.H. y Nelson, F. (1984). Analysis with a limited dependent variable: Linear probability, logit and probit models. Sage series on Quantitative Analysis.

Arellano, M. y Bond, S. (1991). "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations". Review of Economic Studies. Vol. 58, pp. 277-297.

Cameron, A.C. y Trivedi, P.K. (1998). "Regression analysis of count data". Cambridge University Press: Cambridge, UK.

Cho, M. (1998). "Ownership structure, investment, and the corporate value: an empirical análisis". Journal of Financial Economics. Vol. 47, pp. 103-121.

Cramer, J.S. (1991). The logit model: an introduction for economists. Edward Arnold, Londres.

De la Fuente, J.M. Durán, J.J. y Jiménez, A. (2008). "La importancia del riesgo país y político en la presencia exterior de las empresas multinacionales españolas". Comunicación presentada en el congreso nacional XVIII Congreso Nacional de ACEDE los días 14, 15 y 16 de Septiembre de 2008 en León (España).

De Quevedo, E. y De la Fuente, J.M. (2003). "El estudio de endogeneidad y circularidad en la relación reputación-creación de valor aplicando modelos de ecuaciones simultáneas". En Camisón, C. Oltra, M.J. y Flor, M.L. (Eds) (2003) "Enfoques, problemas y métodos de investigación en Economía y Dirección de Empresa. Actas del VIII Taller de Metodología de ACEDE". Acede y Fundació Universitat Empresa. Pp. 209-224.

Disdier, A.T. y Mayer, T. (2004). "How different in Eastern Europe? Structure and determinants of location choice by French firms in Eastern and Western Europe". Journal of Comparative Economics. Vol. 32, pp. 280-296.

Durán, J.J. De la Fuente, J.M. y Jiménez, A. (2008). "Political risk as a determinant of investment by Spanish multinational firms in Europe. Is there an East-West structure?" Comunicación presentada en el congreso internacional 34th EIBA Annual Conference "International Business and Catching-up economies: Challenges and opportunities", los días 11, 12 y 13 de Diciembre de 2008 en Tallin (Estonia).

Fernández, A. (2005). Econometría. Ed. Pearson Prentice Hall.

García -Canal, E. y Guillén, M.F. (2008). "Risk and the strategy of foreign location choice in regulated industries". Strategic Management Journal. Vol. 29, Iss. 10, pp. 1097-1115.

Gujarati, D.N. (1997). Econometría. McGraw-Hill Interamericana, Santa Fe de Bogotá.

Gujarati, D. (2003). "Econometria / Damodar N. Gujarati". 4ª Ed. McGraw Hill. México.

Gyimah-Brempong, K. y Traynor, T.L. (1999). "Political instability, investment and economic growth in sub-Saharan Africa". Journal of African Economies. Vol. 8, Iss. 1, pp. 52-86.

Herbst, C.M. y Barnow, B.S. (2008). "Close to home: a simultaneous equations model of the relationship between child care accessibility and female labor force participation". Journal of Family and Economic Issues. Vol. 29, pp. 128-151.

Kim, Y.H. Rhim, J.C. y Friesner, D.L. (2007). "Interrelationships among capital structure, dividends, and ownership: evidence from South Korea". Multinational Business Review. Vol. 15, Iss. 3, pp. 25-42.

Hair, J.F. Anderson, R.E. Tatham, R.L. y Black, W. (1999). Análisis multivariante. Prentice Hall. 5ª Edición. Madrid.

Hausman, J. (1978). "Specification tests in econometrics". Econometrica. Vol. 46, pp. 1251-1271.

Hausman, J.A: Hall, B.H. y Griliches, Z. (1984). "Econometric models for count data with an application to the patents-R&D relationship". Econometrica. Vol. 52. pp. 909-938.

Kennedy, P.A. (1992). Guide to econometrics. MIT Press: Cambridge, MA.

Liao, T.F. (1994). "Interpreting probability models: logit, probit and other generalized linear models". Sage University Papers. Series on Quantitative applications in the social sciences. Sage Publications Series No. 07-101.

Jiménez, A. (2008). "Does political risk affect the scope of expansion abroad? Evidence from Spanish MNEs". Working Paper en proceso de revisión para la revista International Business Review.

Jiménez A. (2009). "Habilidades políticas e inversión directa en el exterior: un análisis de datos panel de los flujos desde el Sur de Europa hacia los nuevos miembros de la Unión Europea y el Norte de África" Working paper presentado en el XIX Congreso Nacional de ACEDE los días 9, 10 y 11 de Septiembre en Toledo (España).

Luque Martínez, T. (2000). Técnicas de análisis de datos en investigación de mercados. Ed. Pirámide. Madrid.

McFadden, D.L. (1984). "Econometric analysis of qualitative response models". In Grilicjes, Zvi, Intriligator, M.D. (Eds), Handbook of Econometrics, vol. 2. Elsevier/North-Holland, Amsterdam, pp. 1396-1457.

Maddala, G.S. (1993). Limited dependent and qualitative variables in econometrics. Cambridge University Press. Cambridge.

Maddala, G.S. (1996). Introducción a la econometría. 2ª Ed. Prentice Hall.

Mayer, T. y Mucchielli, J.L. (1999). "La localisation à l'étranger des entreprises multinationales". Économie et statistique. Nº 326-327. pp. 159-176.

Metwally, M.M. (2004). "Impact of EU FDI on economic growth in Middle Eastern countries". European Business Review. Vol. 16, Iss. 4, pp. 381-389.

Neter, J. Wasserman, W. y Kutner, M.H. (1985). Applied linear statistical models: regression, analysis of variance and experimental designs. 2ª Ed. Irwin, Homewood.

Robinson, P.M. (1982). "On the asymptotic properties of estimators with limited dependent variables". Econometrica. Vol. 50, pp. 27-42.

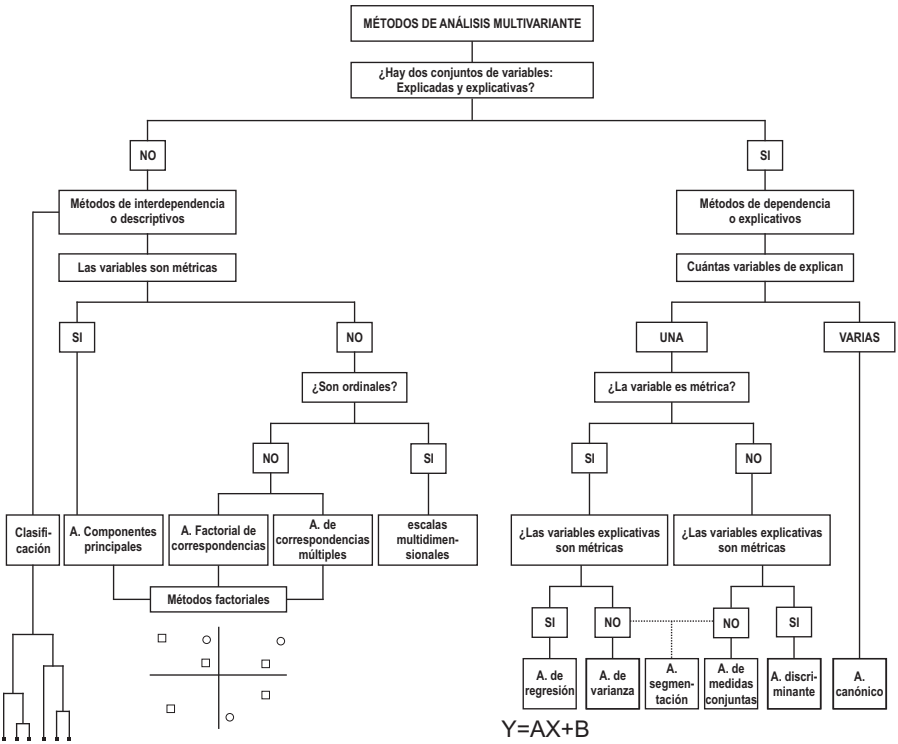
Temple, J. (1999). "The new growth evidence". Journal of Economic Literature. Vol. 37, pp. 112-156.

Slangen, H.L. y Van Tulder, R.J.M. (2009). "Cultural distance, political risk, or governance quality? Towards a more accurate conceptualization and measurement of external uncertainty in foreign entry mode research". International Business Review. Vol. 18, Iss. 3, pp. 276-291.

Statacorp (2001). “Stata statistical software: Release 7.0.”. College Station, Tx: Stata Corporation.

Tsangarides, C.G. (2001). “On cross-country growth and convergence: evidence from African and OECD countries”. Journal of African Economies. Vol. 10, Iss. 4, pp. 355-389.

Stundemund, A. H. (1992). Using econometrics: a practical guide. HarperCollins: New York.



Fuente: Abascal y Grande (1989)