El análisis multivariado y la investigación en las ciencias de la salud

Juan Luis Londoño Fernández
Facultad Nacional de Salud Pública
Universidad de Antioquia. Apartado 51922

Antecedentes

A medida que avanza la búsqueda del conocimiento de los factores que afectan la salud de los individuos y de los pueblos, se vuelve necesario plantear relaciones multifactoriales en las cuales, la salud, como variable dependiente, pueda explicarse o predecirse a partir de múltiples factores o variables independientes.

Los fundamentos teóricos del análisis multivariado se desarrollaron principalmente durante los tres primeros decenios del presente siglo gracias al trabajo de Ronald Fisher y Karl Pearson; al principio, los modelos multivariados se aplicaron predominantemente a las investigaciones experimentales sobre agricultura, pero a partir de la Segunda Guerra Mundial se inició su aplicación a los estudios epidemiológicos analíticos. Como pioneros en este campo cabe citar los trabajos sobre los factores de riesgo para la enfermedad coronaria realizados por Cornfield en 1962, por Truett y colaboradores en 1967, y por Kleinbaum y colaboradores en 1971.[1-3]

La mayoría de los estudios de este tipo realizados hasta principios de los años setenta utilizaron la estrategia conocida con el nombre de análisis discriminante; en 1970 Cox[4] llama la atención sobre las posibilidades de uso, la denominada función logística, la cual presenta ventajas sobre el análisis discri-

minante que son importantes para el estudio de muchas situaciones propias de la investigación epidemiológica analítica.[5,6]

Desde 1975 se publican con frecuencia trabajos de investigación basados en el modelo logístico multivariado, y, en la actualidad, la utilización de este modelo en los trabajos relacionados con el estudio de la relación de múltiples factores de riesgo con la incidencia de una determinada patología es algo casi rutinario.

El desarrollo de la tecnología de los computadores, por su parte, el cual ha sido paralelo al desarrollo metodológico, ha hecho posible aplicar los planteamientos teóricos a los problemas de naturaleza multivariada, lo cual constituye un poderoso impulso, tanto para el desarrollo teórico como para la investigación aplicada.

A este respecto vale la pena mencionar, entre otros, algunos programas de computador que han constituido un significativo aporte en este sentido, como el SAS —Statistical Analysis System—,[7] SPSS —Statistical Package for the Social Sciences—,[8] y BMDP —Biomedical Package—.[9] El programa True Epistat[10] ofrece algunas posibilidades para el análisis multivariado, y el programa Egret —Epidemiological Graphics Estimation Testing— permite realizar análisis multivariados usados con frecuencia en epidemiología.[11]

El análisis multivariado es un conjunto de estrategias constituido por muchos métodos, entre los cuales se destacan aquellos que ofrecen posibilidades más interesantes para el estudio de situaciones propias del campo de la salud, como el análisis de varianza, el análisis de regresión multivariada, el análisis factorial, el análisis discriminante y el análisis de regresión logística multivariada. Cada uno de tales procedimientos tiene usos y características que le son propios, pero los más importantes son el análisis de la varianza y la regresión multivariada.

Aunque no es posible profundizar aquí en todos los métodos arriba enunciados, la consulta de las siguientes obras puede ayudar al lector interesado en profundizar en el tema: Kerlinger,[12] Daniel,[13] Armitage,[14] Kleinbaum y Kupper,[15] Draper y Smith,[16] Dixon y Massey,[17] Snedecor y Cochran,[18] Morrison,[19] Overall[20] y Sokal.[21]

Este documento analiza una serie de situaciones que permiten apreciar la utilidad de los métodos multivariados en la investigación aplicada, al mismo tiempo que se describen algunos de sus fundamentos teóricos, en particular los de la regresión logística multivariada.

Análisis de varianza

En este tipo de análisis se compara la variación observada en la variable dependiente o continua dentro de las categorías de cada una de las variables independientes, las cuales suelen ser nominales o discretas, con la variación observada en la misma variable dependiente entre las distintas categorías de las variables independientes. Como resultado de esta comparación, se obtiene un valor de la variable probabilística F de Snedecor, el cual, referido a tal distribución, permite probar la hipótesis nula de la igualdad del efecto bajo los distintos tratamientos por medio de una prueba de significancia estadística.

Un análisis de este tipo puede usarse para evaluar la hipótesis estudiada por Patrick y colaboradores en la isla de Ponape,[22] según la cual, la discrepancia entre el grado de participación de sus habitantes en la vida moderna, su actitud hacia la misma (X_1), y su preparación para este tipo de vida (X_2), tenían que ver con los valores de presión arterial observados en dicha población (Y).

Un análisis de varianza puede ser utilizado, por ejemplo, para estudiar la relación entre el consumo de alcohol —variable dependiente continua— el nivel socioeconómico —variable independiente discreta— y el sexo —variable independiente discreta—.

Análisis de regresión multivariada

Por medio de este método se evalúa el poder predictivo de un conjunto de variables usualmente continuas como X_1, X_2, ... , X_p sobre una variable dependiente Y continua, de acuerdo con un modelo lineal de la forma:

$$Y = \alpha + \beta X_1 + \beta X_2 + ... + \beta X_p \pm \varepsilon$$

En donde α representa el efecto no explicado por las variables incluidas en el modelo y los distintos coeficientes β representan los coeficientes de regresión, y ε, el error aleatorio.

La estrategia central de este método consiste en descomponer la variación observada en la variable dependiente (Y) en 1), la variación debida a la regresión, que es la variación explicada por cada uno de los factores considerados como variables independientes (X_1, X_2, ..., X_p), y 2) la variación debida a otros factores, la cual no puede ser explicada por los ya vistos (Véase figura 1).

De la razón entre medidas específicas o medias de cuadrados de estas dos fuentes de variabilidad observada en la variable dependiente resulta un valor en la variable probabilística F de Snedecor, la cual permite evaluar la importancia relativa que tiene cada una de las variables independientes incluidas en el análisis, una vez descontado el efecto que tienen sobre Y los factores

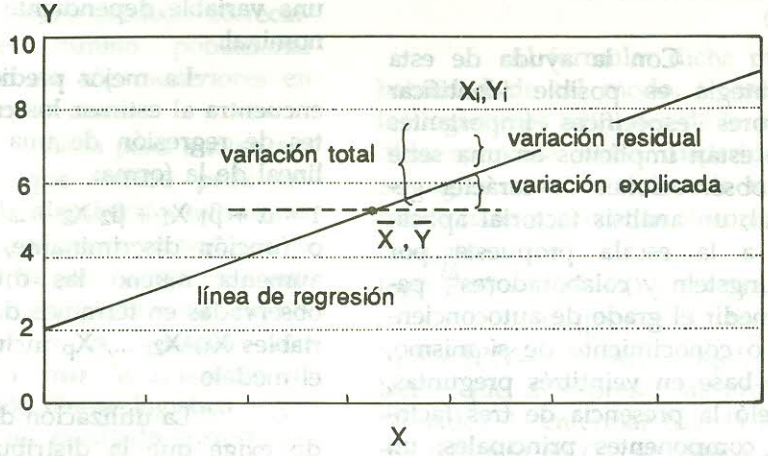

Figura 1
Variación explicada, variación residual y variación total

que se han introducido previamente en el modelo; éste es el proceso secuencial de adición o de exclusión de variables también llamado stepwise.

Incluyendo unos factores y excluyendo otros se obtiene un determinado modelo cuya capacidad de predicción puede evaluarse en términos del porcentaje de la variación observada en Y, el cual es atribuible a los factores incluidos en el modelo por medio de coeficiente de regresión elevado al cuadrado R^2.

Kleinbaum y Kupper[15] proponen una situación que ilustra las posibilidades de aplicación del análisis de regresión múltiple, en la cual se estudia la asociación que existe entre el tamaño poblacional (X_1), el índice de desempleo (X_2) y el índice en el aumento del costo de la vida (X_3), con el índice de homicidios observado en una región durante varios años.

El estudio del comportamiento del índice de homicidios por medio de varios modelos alternativos permite descartar el tamaño poblacional (X_1) como variable predictora del efecto; recordemos que su contribución a la variabilidad de Y es insignificante, y que es necesario seleccionar como mejor modelo aquel en el cual el índice de desempleo (X_3) y el índice en el aumento del costo de la vida (X_3), en su orden, se revelen como las variables que mejor predigan el índice de homicidios o Y.

La estimación del porcentaje de la variabilidad de Y, explicada por dicho modelo, permite afirmar que las dos variables independientes X_2 y X_3 logran explicar el 79,63% de la variabilidad observada en Y.

Análisis factorial

Con la ayuda de esta estrategia es posible identificar factores específicos importantes que están implícitos en una serie de observaciones de carácter general; un análisis factorial aplicado a la escala propuesta por Feningstein y colaboradores[23] para medir el grado de autoconciencia o conocimiento de sí mismo, con base en veintitrés preguntas, reveló la presencia de tres factores componentes principales; un subgrupo de diez preguntas que medían la autoconciencia privada, es decir, la percepción del yo frente a sí mismo, otro subgrupo de siete preguntas que medían la autoconciencia pública, o sea, la percepción del yo frente a los otros y un tercer subgrupo de seis preguntas que medían la ansiedad social, es decir, la percepción del yo frente a la sociedad. La identificación de estos tres factores hizo posible el estudio de su relación con el tabaquismo observado en una muestra de escolares adolescentes a quienes se les aplicó la escala de autoconciencia.[24]

Análisis discriminante

En esencia, el método conocido como análisis discriminante evalúa la capacidad que tiene un conjunto de variables independientes para predecir un resultado, el cual casi siempre es una variable dependiente de tipo nominal.

La mejor predicción se encuentra al estimar los coeficientes de regresión de una función lineal de la forma:

$1 = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$,

o función discriminante, la cual aumenta mucho las diferencias observadas en términos de las variables X_1, X_2, ..., X_p incluidas en el modelo.

La utilización del método exige que la distribución de las variables independientes incluidas en el modelo, sea normal multivariada, una condición que no se satisface en aquellos problemas que incluyen variables independientes de tipo nominal, los cuales son comunes en la investigación epidemiológica.

Press y Wilson[6] proponen una aplicación interesante del análisis discriminante en una situación en la cual se trata de predecir el cambio poblacional, bien sea alto o bajo, de diversas regiones de los Estados Unidos con base en el ingreso per cápita (X_1), la tasa de mortalidad (X_2), la tasa de natalidad (X_3), el grado de urbanización (X_4) y la localización con respecto al mar (X_5), propios de tales regiones.

Se obtuvo entonces la función discriminante:

$$Y = -11,84 + 2,11 X_1 - 2,55 X_2 + 3,22 X_3 - 0,36 X_4 + 1,16 X_5$$

Cuando se aplicó dicho modelo a una muestra de diez re-

giones con el propósito de validarlo, se logró predecir correctamente el cambio poblacional observado en años anteriores en el 68% de los casos.

Vale la pena anotar, sin embargo, que debido principalmente a la distribución no normal de las variables continuas y a la inclusión de dos variables nominales dicotómicas, el grado de urbanización y la situación con respecto al mar, la aplicación de un modelo lineal logístico logró predecir un resultado correcto en la muestra de validación en un 72% de los casos, lo cual indica una moderada superioridad para esta situación del análisis de regresión logística sobre el análisis discriminante.

Análisis de regresión lineal logística multivariada

Al usar esta estrategia, la probabilidad de que se presente la enfermedad Y bajo un conjunto X de factores $X_1, X_2, ..., X_p$, se expresa como:

$$P(Y = y / X_1, X_2, ..., X_p) = \frac{1}{1 + e^{-(\alpha + \Sigma \beta i\ Xi)}}$$

$$= 1 / \{1 + \exp -(\alpha + \Sigma \beta_i X_i)\} \quad (1)$$

Al formular dicha probabilidad de tal modo, y hallar los logaritmos naturales de la medida conocida como disparidad, en inglés odds, se obtiene la relación que se conoce con el nombre de logit:

$$\text{logit } P(X) = \ln [P(X) / (1- P(X)] = \ln \{[P(Y=1/X_1, X_2, ..., X_p)] / [P(Y=0/X_1, X_2, ..., X_p)]\},$$

relación que, al expresar las probabilidades de enfermar bajo la forma indicada en (1), se reduce a:

$$\alpha + \sum_{i=1}^{p} \beta_i X_i \quad (2)$$

De esta manera, el logit que corresponde a un conjunto de factores de exposición, es una función lineal de éstos.

El resultado anterior facilita entonces la estimación del riesgo relativo,[a] correspondiente a diferentes valores de exposición en los múltiples factores X_1 (conjunto X^*) con respecto a otros valores de exposición en los mismos factores o conjunto X°), puesto que se tiene entonces que:

$$\ln RR \approx \ln RD = \ln \{[P(X^*)/(1-P(X^*))] / [P(X^\circ)/(1-P(X^\circ))]\}$$

$$= \ln \{P(X^*)/[1-P(X^*)]\} - \{\ln [P(X^\circ)/[1-P(X^\circ)]\},$$

a Medida que se utiliza frecuentemente en la investigación epidemiológica para expresar el grado de asociación entre la exposición a un factor de riesgo y la incidencia de una determinada enfermedad; su valor se estima, de modo aproximado, por medio de la razón de disparidades u Odds Ratio.

expresión que, según el resultado obtenido en (2), se convierte finalmente en una suma de la forma:

$$\sum_{i=1}^{p} \beta_1 (X_i^* - X_i^\circ) \qquad (3)$$

Según este resultado, ln RD = $\Sigma \beta_i (X_i^* - X_i^\circ)$, y la estimación del riesgo relativo bajo unos ciertos valores de exposición X_1^*, X_2^*, ..., X_p^* con respecto a otros valores X_1°, X_2°, ... , X_p°, se reduce a estimar los coeficientes β_i propios de una regresión lineal que expresa el logit de P(X) según (2).[b]

Una vez que se han estimado tales coeficientes de regresión se estima fácilmente el valor del riesgo relativo, puesto que

$$RR \approx RD = \exp \left[\sum_{i=1}^{p} \beta_i (X_i^* - X_i^\circ) \right]$$

Otra ventaja adicional que ofrece la formulación arriba descrita, consiste en que se vuelve posible controlar de manera inmediata el efecto de las variables de confusión incluidas en el modelo igualando los valores de tales variables X_i en los conjuntos X_i^* y X_i° de la expresión (3). Además, la detección de la interacción que pueda existir entre una exposición estudiada y otros factores se posibilita en el modelo logístico al definir algunos de los X_i en términos de la interacción estudiada; así, si la exposición en estudio se denota con la letra E y un cierto factor se denota con la letra W, la interacción EW se podría estudiar si en el modelo (2) uno de los factores X_i representa el producto EW.

El modelo arriba descrito, el cual se aplica estrictamente al análisis de observaciones obtenidas por un diseño no equiparado como la regresión logística incondicional, se puede aplicar, con algunas modificaciones, a la situación propia de las observaciones obtenidas por equiparamiento, como la regresión logística condicional.

Al comparar el análisis de regresión logística multivariada con el análisis estratificado se advierte que, si bien el primero es un método más complejo, que puede presentar mayor dificultad para interpretar los resultados obtenidos, ofrece ventajas importantes, tales como la posibilidad de controlar simultáneamente el efecto de múltiples variables, una tarea que se complica en el análisis estratificado, cuando el número de observaciones es relativamente escaso, hay una mayor facilidad para el estudio de la interacción, y, en el caso de información equiparada, la posibilidad de estimar el efecto que, sobre la variable dependiente, tiene el factor por el cual se ha equiparado la información.

b La estimación de los coeficientes β_i en la regresión lineal logística se hace usualmente por el método de máxima verosimilitud.

El siguiente ejemplo ilustra las características más sobresalientes del método.

En un estudio sobre factores de riesgo para el cáncer de esófago en China realizado por De Jong y colaboradores y citado por Breslow,[25] se estudió la relación que con dicha enfermedad presentaban los factores dialecto (X_1), el consumo de samsu (X_2), que es una bebida común en la región, el consumo de cigarrillos (X_3), y la frecuencia de consumo de bebidas calientes (X_4).

Con el fin de estimar la probabilidad de enfermar P(Y), se utilizó la función lineal:

$$P(Y / X_1, X_2, X_3, X_4) = 1 / [1 + \exp -(\alpha + \sum_{i=1}^{4} \beta_i X_i)]$$

Haciendo uso del método de máxima verosimilitud, se estimaron los coeficientes de regresión β_1 con sus respectivos errores estándar para finalmente obtener las estimaciones de los riesgos relativos ajustados por las demás variables incluidas en el modelo utilizando la forma:

$RRx_i = \exp(\beta_i)$, (23) (24) (25) (Véase tabla 1)

Referencias

1. CORNFIELD, J. "Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis". *Fed Proc.*, 21: 58-61, 1962.
2. TRUETT, J.; CORNFIELD, J. and KANNEL, W. "A multivariate analysis of the risk of coronary heart disease in Framingham". *Journal of Chronic Diseases* 20: 511-524, 1967.
3. KLEINBAUM, D. et al. "Multivariate analysis of risk of coronary heart disease in Evans County, Georgia". *Archives of Internal Medicine* 128: 943-948, 1971.
4. COX, D. R. *The analysis of binary data.* London, Methuen, 1970.
5. HALPERIN, M.; BLACKWELDER, W. and VERTER, J. "Estimation of the multivariate logistic risk function

Tabla 1 Coeficientes de las variables más o menos error estándar en la función del riesgo relativo múltiple, estimados al utilizar un análisis de regresión lineal logística para el diseño no equiparado (Estudio del *IARC* sobre cáncer de esófago en Singapur).

Variables en el modelo	*Coeficientes β ≠ E.S*	*RR*
Constante	- 3,2062 ± 0,3650	
X_1 (dialecto)	1,4145 ± 0,3301	4,11
X_2 (samsu)	0,5352 ± 0,2766	1,71
X_3 (cigarrillos/día)	0,0121 ± 0,0095	1,01
X_4 (bebidas muy calientes)	0,7556 ± 0,1493	2,13

and maximum likelihood approaches". *Journal of Chronic Diseases* 24: 125-158, 1971.

6. PRESS, J. and WILSON, S. "Choosing between logistic regression and discriminant analysis". *Journal of the American Statistical Association* 73 (364): 699-705, 1978.

7 SAS Institute Inc., *"Sas Circle"*, Po. Box 8000, Cary, North Carolina, USA.

8. SPPS Inc., 444 N. Michigan Avenue, Chicago, Illinois, USA.

9. University of California, *Health Sciences Computing Facilities*, Los Angeles, California, 90024, USA.

10. EPISTAT SERVICES, 2011 Cap Rock Circle, Richardson, Texas, USA.

11. Statistics and Epidemiology Research Corporation, 909 Northeast 43 rd. Street, Suite 310, Seattle, WA 98105, USA.

12. KERLINGER, F. N. *Foundations of behavioral research*. New York, Holt, Rinehart and Winston, 1973.

13. DANIEL, W. *Biostatistics: a foundation for analysis in the health sciences*. New York, John Wiley and Sons, 1983.

14. ARMITAGE, P. *Statistical methods in medical research*. Oxford, Blackwell Scientific Publications, 1971.

15. KLEINABUM, D. G. and KUPPER, L. L. *Applied regression analysis and other multivariable methods*. Belmont, Calif., Wadsworth Publishing Company, 1978.

16. DRAPER, N. R. and SMITH, H. *Applied regression analysis*. New York, John Wiley And Sons, 1966.

17. DIXON, W. J. and MASSEY, F. J. *Introduction to statistical analysis*. Tokyo, McGraw-Hill, 1963.

18. SNEDECOR, G. W. and COCHRAN, W. G. *Statistical methods*. Iowa, The Iowa State University Press, 1967.

19. MORRISON, D. F. *Multivariate statistical methods*. New York, Mc Graw Hill, 1967.

20. OVERALL, J.E. and KLETT, C. J. *Applied multivariate analysis*. New York, McGraw-Hill, 1972.

21. SOKAL, R. R. and ROHLF, F. J. *Biometry*. San Francisco. W.H. Freeman, 1969.

22. PATRICK, R. "The ponape study of the health effects of cultural change". Trabajo presentado en el Simposio Anual de la Sociedad para la Investigación Epidemiológica, Berkeley, California, 1974.

23. FENINGSTEIN, A; SCHEIER, M. F. and BUSH, A. H. "Public and private selfconsciousness: assesment and theory". *Journal of Consulting and Clinical Psichology* 43: 522-527, 1975.

24. LONDOÑO, J. L. "Factores relacionados con el consumo de cigarrillo en escolares adolescentes de la ciudad de Medellín". En prensa, Medellín, 1989.

25. BRESLOW, N. E. and DAY, N. E. *Statistical methods in cancer research*. Vol. 1: The Analysis of Case-Control Studies. Lyon, IARC Scientific Publications No. 32, 1980, p.273.