ALGUNAS OBSERVACIONES ACERCA DE LOS PROBLEMAS ESTADISTICOS QUE PUEDEN PRESENTARSE EN EL ANALISIS DE LA ENCUESTA MUNDIAL DE FECUNDIDAD*

Sir Maurice Kendall
Director del Proyecto WFS

SOME NOTES ON STATISTICAL PROBLEMS LIKELY TO ARISE IN THE ANALYSIS OF *WFS* SURVEYS

SUMMARY

This paper examines some of the more important problems that arise in the analysis of the data from the World Fertility Survey (WFS). At the same time it warns of applying indescriminately certain current statistical techniques.

Since the statistical and the demographic approaches are different, althougt both may pursue the same ends, such as the construction of an explanatory model, the author suggests that the best way of approaching the analysis of the WFS data is to maintain a continuing dialogue between the specialists in both disciplines.

---

* El INSTITUTO INTERNACIONAL DE ESTADISTICA ha decidido publicar una serie de *Boletines Técnicos* destinados a analizar problemas metodológicos específicos relacionados con la Encuesta Mundial de Fecundidad, actualmente en curso. Como contribución a dicha Encuesta y a la difusión de sus resultados, NOTAS DE POBLACION reproducirá regularmente esos boletines, a cuyo primer número corresponde el presente trabajo.

INTRODUCCION

Un análisis detenido de una encuesta mundial de fecundidad (EMF) requerirá probablemente una especialización estadística de muy diversa índole, desde la simple tabulación y el manejo de los números hasta las técnicas matemáticas más refinadas. No es posible comentar en estas notas toda la metodología que podría necesitarse, para lo cual se requeriría por lo menos un volumen. El análisis que sigue se limita por lo tanto a algunos de los principales problemas que se preven en esta etapa y a algunas advertencias acerca de los peligros que entrañaría el empleo indiscriminado de ciertas técnicas estadísticas rutinarias.

El proceso analítico consiste en buena parte en ajustar modelos a los datos, o en establecer si los datos concuerdan con las hipótesis que el demógrafo formula para su prueba. El estadístico tiende a buscar modelos mediante el examen de los datos, en muchos casos sin tener una idea previa de la causalidad del sistema, aunque tiene que examinar la coherencia lógica de sus supuestos; el demógrafo tiende a abordar el análisis con una base de hipótesis posibles surgidas de su conocimiento y experiencia previos. Pero ambos persiguen en realidad el mismo objetivo: la construcción de un modelo explicativo. Y aunque el término "explicación" es un término relativo y "causalidad" es un concepto esquivo, parece evidente que el resultado más provechoso de los estudios EMF provendrá de un diálogo continuo entre el estadístico y el demógrafo. En este documento se tratan algunos de los tópicos estadísticos que será necesario que ellos examinen.

CATEGORIAS DE VARIABLES EN EL ANALISIS DE REGRESION

1. En muchos contextos demográficos se necesita proceder a la regresión de una variable y ("regresante") sobre una serie de variables $x_1, x_2 \ldots . x_p$ (regresoras), algunas de cuyas x no son variables continuas pero sí están categorizadas. Por ejemplo, los individuos en estudio pueden estar separados por sexo, clasificados en alguna agrupación ordenada, como el nivel educativo o la clase social, o clasificados en una agrupación no ordenada, como la religión o la raza. Suele proponerse que en tales casos las clases discontinuas se representen mediante una pseudo variable; por ejemplo, los varones y las mujeres por una variable (1,0); las actitudes favorables, neutras o desfavorables por una variable tripartita (+1, 0, -1); tres grupos religiosos A, B, C, por tres variables, una que represente el valor 1 si el sujeto es una A (y cero en caso contrario), otra que represente el valor 1 si el sujeto es B (y cero en caso contrario), y la otra el valor 1 si el sujeto es C (y cero en caso contrario). Son posibles otras variaciones.

2. Estas pseudo variables, especialmente las dicotómicas, son denominadas frecuentemente "mudas". Esto no es muy exacto por-

que, estrictamente hablando, una variable muda permanecería constante, pero el término ha penetrado profundamente en la literatura y tiene el mérito de ser breve.

3. Lo que hay que examinar es si estas pseudo variables pueden utilizarse en un análisis ordinario de regresión de mínimos cuadrados y arrojar resultados significativos. La situación en general no es de ninguna manera clara. Considérese en primer término el caso sencillo en que una variable "regresante" Y se "regresa" simplemente sobre una variable "regresora" continua x y los individuos están clasificados por sexo. Una manera simplista aunque frecuente consistiría en analizar el modelo:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon \tag{1}$$

donde Z es la variable que representa el sexo, que supondremos igual a 1 para los hombres (n_1 en la notación) y cero para las mujeres (n_2 en la notación). Un análisis de mínimos cuadrados correcto llevaría a los estimadores:

$$b_0 = \bar{y} - b_1\bar{x} - b_2\bar{z} \tag{2}$$

$$b_1 = \frac{n_1 \ \mathrm{cov}_1 \ (y,x) + n_2 \ \mathrm{cov}_2 \ (y,x)}{n_1 \ \mathrm{var}_1 \ x + n_2 \ \mathrm{var}_2 \ x} \tag{3}$$

$$b_2 = \bar{y}_2 - b_1\bar{x}_2 - b_0 \tag{4}$$

donde las barras indican las medias de las observaciones y los subíndices 1 y 2 de las x se refieren a las categorías masculina y femenina respectivamente.

Ahora, si hubiéramos analizado los dos grupos separadamente siguiendo el mismo procedimiento, habríamos obtenido.

para el grupo masculino:

$$b_1 \ (\text{masculino}) = \frac{\mathrm{cov}_1 \ (y,x)}{\mathrm{var}_1 \ x} \tag{5}$$

y para el grupo femenino:

$$b_1 \ (\text{femenino}) = \frac{\mathrm{cov}_2 \ (y,x)}{\mathrm{var}_2 \ x} \tag{6}$$

Comparando en seguida con la ecuación (3) se ve que el coeficiente de regresión b_1 de los dos grupos juntos es un promedio ponderado del coeficiente que se obtiene tratando los dos grupos por separado.

4. El efecto de la variable muda ha consistido por consiguiente en promediar dos relaciones que pueden ser totalmente diferentes. Evidentemente, sería más conveniente mantener separadas estas relaciones, a menos que se pudiese demostrar que son lo bastante semejantes como para justificar la amalgamación.

5. Un modelo más elaborado requiere agregar a la ecuación (1) un término de "interacción" XZ, de modo que el modelo se convierte en:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \epsilon \qquad (7)$$

Una solución de mínimos cuadrados ahora da:

$$Y = \bar{y}_2 - \frac{\text{cov}_2\,(y,x)}{\text{var}_2\ x} + \frac{\text{cov}_2\,(y,x)}{\text{var}_2\ x}\,X$$

$$+ \left\{\bar{y}_1 - \bar{Y}_2 + \frac{\text{cov}_2\,(y,x)}{\text{var}_2\ x}\,\bar{x}_2 - \frac{\text{cov}_1\,(y,x)}{\text{var}_1\ x}\,\bar{x}_1\right\} Z \qquad (8)$$

$$+ XZ\left\{\frac{\text{cov}_1\,(y,x)}{\text{var}_1\ x} - \frac{\text{cov}_2\,(y,x)}{\text{var}_2\ x}\right\}$$

Si $x_2 = 1$, esta ecuación se transforma en:

$$Y = \bar{y}_1 - \frac{\text{cov}_1\,(y,x)}{\text{var}_1\ x}\,\bar{x}_1 + \frac{\text{cov}_1\,(y,x)}{\text{var}_1\ x}\,X \qquad (9)$$

que es la regresión ordinaria de y sobre x_1 en el grupo masculino. Del mismo modo, si $x_2 = 0$, obtenemos la regresión ordinaria en el grupo femenino. La pseudo-variable x_2 ha amalgamado, aparentemente, los dos grupos en la ecuación (8), pero en realidad sólo constituye una expresión sintética de las dos relaciones (posiblemente diferentes), una de las cuales es la ecuación (9).

6. Sin embargo, la ecuación (8) efectivamente prueba si las dos relaciones tienen la misma β_1. En realidad, el coeficiente del último

término de esta ecuación es la diferencia de las β_1 estimadas para los dos grupos. Si es cero o demasiado pequeña, las dos regresiones tienen la misma pendiente y pueden amalgamarse en el caso de b_1. Pueden aún tener valores diferentes de las β_0 estimadas, es decir, pueden representarse por líneas paralelas.

7. Aparecerán efectos similares cuando una pseudo variable está formada por más de dos clases o cuando existen varias pseudo variables. Dos actitudes son posibles: mantener las líneas de regresión claramente diferenciadas dentro de cada categoría antes de intentar cualquier tipo de amalgamación: o elaborar un modelo totalmente interactivo, examinar si puede eliminarse algún término y usar el resultado para derivar las regresiones individuales. Sin embargo,

i) Si existen muchas categorías de variables, las frecuencias dentro de las subcategorías pueden resultar pequeñas, tan pequeñas que las regresiones dentro de ellas adolezcan de tal variabilidad de muestreo que carezcan de confiabilidad. Por ejemplo, una muestra de 5 000 dividida por sexo, tres grupos étnicos, cuatro categorías de educación y cinco regiones geográficas, que da un total de 2 x 3 x 4 x 5 = 120 sub-categorías, sólo tendría un número promedio de muestra en las categorías de 42, y algunas serían inferiores. Valdría la pena combinar algunas de las categorías para lograr una mayor confiabilidad de la muestra y prevenir la posibilidad de que surjan relaciones no idénticas.

ii) A veces puede que no exista interés en mantener categorías diferenciadas. Por ejemplo, si tenemos una muestra de mujeres y efectuamos la regresión de su fecundidad (es decir, los números de hijos tenidos) sobre el ingreso en cuatro zonas geográficas diferentes, y si la muestra es representativa en cuanto a los números correspondientes a esas regiones y si sólo nos interesa la relación entre la fecundidad y el ingreso *para toda la zona*, puede bastar una relación.

iii) A veces las categorizaciones *ordenadas* pueden representarse, con un grado de aproximación satisfactoria, mediante números de orden, que luego se tratan como variables ordinarias. Supóngase, por ejemplo, que tenemos la siguiente clasificación por nivel social de 1 000 individuos, siendo *A* el más alto:

A	B	C	D	E
50	150	500	200	100

Si ésta fuera una clasificación de 1 000 personas, podríamos considerar que las primeras 50 se ordenan por rango de 1 a 50 y atribuir a cada uno el promedio de esos órdenes $1/50(1+2.....+50) = 25.5$, tratándolos como rangos ligados. En el siguiente

grupo, cada uno tendría un promedio de 1/150(51 + ...200) = 125,5, etc. En tales casos podrían darse cifras más complicadas, pero ellas dependen de algún supuesto acerca de la distribución que dio origen a la categorización observada.

iv) Sin embargo, puede ser de interés trabajar con la ecuación del tipo (7), en la que interviene cierto número de variables mudas, y aplicarle alguna de las formas de regresión rutinarias que eliminan las variables no contributivas, con el objeto de lograr la representación más sintética. Este procedimiento debe aplicarse con cautela y teniendo siempre presente las realidades de la situación.

8. Como ejemplo del efecto "promediador" que puede resultar de una variable muda, he aquí algunos datos referentes a la encuesta EMF de Fiji. Se efectuó la regresión de la edad al matrimonio y sobre la edad de la mujer (x_1), sus años de educación (x_2) y la raza (x_3)(fijianos = 0, indios = 1). El resultado para todo el grupo fue:

$$y = 0.09x_1 + 0.30x_2 - 0.18x_3 \qquad (10)$$

habiéndose medido las variables con su media.

El resultado para el grupo fijiano fue:

$$y = 0.20x_1 + 0.16x_2 \qquad (11)$$

y para el grupo indio:

$$y = -0.04x_1 + 0.33x_2 \qquad (12)$$

Aparece con claridad que las relaciones expresadas por las ecuaciones (11) y (12) son totalmente diferentes y se confunden cuando se las reune en la ecuación (10).

Este ejemplo, aunque se basa en datos reales acerca de 5 000 casos aproximadamente en la ecuación (10), más de 2 500 en la ecuación (12) y más de 2 000 en la ecuación (11), se da sólo con fines ilustrativos. Para un estudio más profundo se necesitaría examinar la relación de la edad al matrimonio con otras variables.

RELACIONES ENTRE LAS VARIABLES "REGRESORAS"

9. En una ecuación como:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon \qquad (13)$$

ha sido costumbre referirse a Y como a la variable "dependiente" y a las X como a las variables "independientes". Es relativamente raro que las X sean independientes en el sentido estadístico y muy a menudo están estrechamente correlacionadas. Esto crea problemas especiales en la interpretación de tal ecuación y, particularmente, respecto de las contribuciones relativas de las X individuales a Y.

Para evitar que el término ambiguo de "independiente" se aplique a variables mutuamente dependientes, también se acostumbra a denominar a Y con el término de variable "explicada" y a las X con el término de "explicativas". Aunque esto constituye un progreso terminológico, no está totalmente exento de objeciones, pues las variables "explicativas" pueden no influir en la variación de Y estando ligadas a ella únicamente por algún mecanismo causal indirecto. Una terminología absolutamente neutral es denominar a la Y variable "regresante" y a las X variables "regresoras".

10. Es conveniente por lo tanto examinar las relaciones entre las X antes de entrar en un análisis de regresión. Para esto se requiere un enfoque un tanto complicado que es difícil resumir en términos no técnicos. Un examen de las correlaciones individuales entre pares de variables no es suficiente. Lo que se necesita es un análisis de toda la serie de correlaciones o covariancias. Uno de los mejores métodos consiste en calcular la matriz de covariancia o correlación de las variables "regresoras" y determinar las constantes, conocidas como raíces latentes o valores característicos. Un valor característico cero implicará una relación lineal entre algunas de las X, y por consiguiente, una redundancia entre ellas. Un valor característico pequeño indica una casi colinealidad entre las X y advierte que los coeficientes b, los estimadores de β, carecerán individualmente de confiabilidad.

En realidad, los estimadores b, en términos de matriz, son:

$$b = yx^1 \ (xx^1)^{-1} \qquad (14)$$

donde x es la matriz pxn de las observaciones x (p variables, n observaciones); x^1 es su transposición, y es el vector $1\,xn$ de las observaciones de y y por consiguiente, la matriz de covariancia (con las x medidas cer-

ca de sus medias) es $(xx)^{-1}$. El hecho de que esta matriz aparezca como inversa implica que si tiene un determinante pequeño (correspondiente a uno o más valores característicos pequeños), los estimadores b estarán inflados y carecerán individualmente de confiabilidad. De presentarse esta situación, es preferible suprimir algunas variables.

11. Un ejemplo, tomado también de los datos de Fiji, servirá para ilustrar este punto. La variable "regresante" y es la paridez (número de hijos). Las variables "regresoras" son la edad de la madre en años, x_1; los años de educación de la madre, x_2; el tamaño deseado de la familia, x_3; y la duración de la vida marital, x_4.

La matriz de la correlación de las variables "regresoras" es la siguiente:

	x_1	x_2	x_3	x_4
1	1.000			
2	.914	1.000		
3	.500	.548	1.000	
4	- .321	- .430	- .280	1.000
Corr. con y	.635	.700	.774	- .342

Las variables "regresoras" están altamente correlacionadas y un análisis del componente principal da los siguientes valores característicos:

Componente	Valor característico (porcentaje del total)	Porcentaje acumulativo
1	2.56 (64 o/o)	64.0
2	0.77 (19 o/o)	83.3
3	0.59 (15 o/o)	98.0
4	0.08 (2 o/o)	100.0

La pequeñez del valor característico más bajo indica que las cuatro variables "regresoras" son casi colineales y que los coeficientes de una regresión de y carecen totalmente de confiabilidad.

En realidad, la regresión de y sobre las cuatro variables (no medidas cerca de sus medias) es:

$$y = -.006x_1 - .015x_2 + .842x_3 + .130x_4 - 1.123$$
$$R^2 = 0.77 \qquad (15)$$

La regresión sobre x_1, x_2 y x_3 es:

$$y = .145x_1 - .077x_2 + .809x_3 - 3.606$$
$$R^2 = .74 \qquad (16)$$

La regresión sobre x_2, x_3, x_4 es:

$$y = -.021x_2 + .744x_3 + .159x_4$$
$$R^2 = .77 \qquad (17)$$

En cuanto a la bondad del ajuste, medido por el cuadrado del coeficiente de correlación múltiple R^2, las ecuaciones (15), (16) y (17) son casi tan buenas unas como otras. Es evidente que no puede atribuirse un significado preciso a los coeficientes individuales.

12. Es importante comprender la fuerza de este argumento. La inferencia clásica de la ecuación (13) sería, por ejemplo, que si X_2..... X_3 permanecen fijas, una variación α en X_1, ocasionaría una variación $\alpha\beta_1$ en Y. Esto es verdad, pero en muchos casos no tiene importancia. Puesto que las X están intercorrelacionadas, una variación de X_1 entrañaría en general variaciones con las otras X de modo que de hecho no permanecerán constantes.

13. De lo anterior se desprende que a menos que las variables "regresoras" no estén correlacionadas o lo estén muy débilmente, no cabe atribuir un significado especial a los coeficientes individuales en una ecuación de regresión. Es la ecuación en su totalidad la que importa, es decir, la excelencia del ajuste según el tamaño del coeficiente de

correlación múltiple R^2. También se desprende que en general no podemos usar esos coeficientes para medir la contribución relativa de las variables "regresoras" individuales a Y. La incapacidad para apreciar este punto ha perjudicado muchos de los análisis de regresión publicados.

14. Alguien se preguntará naturalmente si, en caso de existir dependencia entre las variables "regresoras", es posible establecer su contribución relativa a la variable "regresante". La respuesta general es negativa, en cuanto se refiere únicamente a la técnica de regresión. Para avanzar más hacia una explicación causal se requiere elaborar un modelo causal para análisis, como el que aparece en el *Technical Bulletin* número 2. 1/.

EFECTOS DE LA AGRUPACION

15. En el trabajo demográfico ocurre a menudo que los individuos en estudio se agrupan en clases de frecuencia. Por ejemplo, en un grupo de 5 000 mujeres sería una costumbre corriente agruparlas en categorías de edad, por ejemplo 15-19, 20-24, 25-29, etc. Surge un problema cuando las correlaciones o regresiones basadas en tales datos son marcadamente diferentes de lo que serían si los datos no estuvieran agrupados. El tema ha sido extensamente analizado de un modo bastante sofisticado por Haitovsky (1973).

16. Considérese en primer lugar la correlación entre dos variables x_1 y x_2 que están agrupadas respectivamente en intervalos de h_1 y h_2. El cálculo de la variancia de los datos agrupados exagera la verdadera (no agrupada) variancia en una cantidad que queda bastante bien representada por un término correctivo conocido con el nombre de Sheppard.

$$\text{var (no agrupadas)} = \text{var (agrupadas)} - h^2/12 \qquad (18)$$

En cambio, la covariancia no necesita esta corrección.

Así, la correlación estimada para datos no agrupados es:

$$\left\{ \frac{\text{cov } (x_1, x_2)}{\text{var } x_1 \ \text{var } x_2} \right\}^{1/2} \qquad (19)$$

1/ M.G. Kendall y C.A. O'Muircheartaigh, *Path Analysis and Model Building*.

donde var x_1 y var x_2 se calculan con el material no agrupado.

Si trabajamos con material agrupado sin corregir, el denominador de la ecuación (19) será demasiado amplio y la correlación sería:

$$\frac{\text{cov}\ (x_1,\ x_2)}{\left\{(\text{var}\ x_1 - h_1{}^2/12)\ (\text{var}\ x_2 - h_2{}^2/12)\right\}^{1/2}} \qquad (20)$$

En la ecuación (20) var x se refiere a la variancia estimada a base del material agrupado. Sin embargo, sin correcciones de agrupación, la correlación calculada será demasiado pequeña, en un grado que depende de la tosquedad del sistema de agrupación. El efecto de la agrupación no es uniforme, sino que depende de la distribución de frecuencia de las X. Pueden presentarse casos en que la atenuación debida a la agrupación se invierte.

17. Por lo tanto, parece mejor trabajar con datos no agrupados, cuando sea posible. Pueden formularse consideraciones similares respecto de las regresiones. En general, la agrupación sacrifica información y puede distorsionar las relaciones entre las variables.

18. Otro efecto bastante sutil aparece cuando la agrupación se lleva a cabo. En el modelo clásico de la ecuación (13) se presume que el residuo aleatorio ε es homocedástico, es decir, tiene la misma variancia cualquiera que sea el valor de Y. Cuando las observaciones de Y se agrupan, esta propiedad tiende a perderse debido a que una serie de Y, digamos de número n, aglomeradas en un punto único, tiene una variancia dependiente de n; de este modo, si los números correspondientes a las frecuencias de clase difieren (como casi siempre ocurrirá), el término error tiene diferentes variancias en diferentes puntos de la amplitud de Y. Esta es otra razón para trabajar con datos no agrupados.

19. En la mencionada monografía, Haitovsky examina dos posibilidades: (*a*) restablecer variancias iguales de datos agrupados mediante una transformación lineal, y (*b*) estimar coeficientes de regresión cuando sólo se dispone de frecuencias marginales; por ejemplo, en el caso de dos variables, cuando no existe una clasificación completa, pero sí una clasificación de cada una de ellas. Existen algunos peligros serios en esta parte del asunto, aunque muy a menudo no existen recursos mejores, como ocurre cuando se trabaja con tablas publicadas. En el contexto de la Encuesta Mundial de Fecundidad (EMF), parecería deseable trabajar con datos no agrupados, donde se estudian medias, variancias, correlaciones, regresiones o tipos similares de estadísticas.

VALORES QUE FALTAN

20. En la tabulación ordinaria, los valores que faltan pueden insertarse en columnas encabezadas por las expresiones "no disponibles", "no proporcionados" y otras similares. Para procedimientos más complicados que incluyan el análisis con variables múltiples, la falta de datos constituye un engorro y conviene disponer de algún método para considerarlos. También es éste un asunto que ofrece algunos peligros serios.

21. Para los efectos de precisar los conceptos, supóngase que tenemos que proceder a la regresión de Y sobre cuatro variables X_1 a X_4 y que no se dispone de algunas Y y de algunos valores de X. Por supuesto, una manera simple de proceder consistiría en prescindir de todos los registros incompletos. Pero este procedimiento significa sacrificar una buena cantidad de información. Otro camino consiste en revisar los registros completos hasta encontrar uno que contenga la información que falta en el incompleto, y reemplazar los datos que faltan con los de aquel: tal es el método denominado "hot-deck". Otro procedimiento consiste en reemplazar los datos que faltan por números elegidos al azar dentro del margen permisible de la variable que falta: es el método denominado "cold-deck". Ambos son métodos de imputación, que pueden ser objetados desde el punto de vista ético o político y en cualquier caso requieren de una cantidad bastante grande de datos completos para obtener los equivalentes necesarios.

22. Existen métodos más elaborados que tratan de usar toda la información existente, inclusive la contenida en los registros incompletos, estimando los valores que faltan a base de los registros completos. Por ejemplo, si tenemos cierto número de registros completos con los valores conocidos de X_1 a X_4, podemos proceder a la regresión de X_4 sobre X_1 a X_3 y utilizarla para estimar X_4 en los casos en que X_4 falta pero se conocen X_1 a X_3. El tema ha sido estudiado por Beale y Little (1975), quienes, examinando seis enfoques diferentes basándose en parte en la teoría y en parte en estudios de simulación, llegan a la conclusión de que el mejor procedimiento es el que ellos llaman verosimilitud máxima modificada. Efectivamente, es un método de iteración a la convergencia. Los registros completos se utilizan para estimar las medias y covariancias de todas las variables.

Este resultado se utiliza para estimar las cantidades que faltan, las que se sustituyen repitiéndose la estimación de las medidas y covariancias y así hasta obtener la convergencia.

23. Un peligro serio que debe evitarse es el uso en un análisis simple de estimaciones de medias y covariancias de muestras de tamaños diferentes. Por ejemplo, si de 1 000 registros existen 900 casos en que aparecen X_1 y X_2, es factible calcular las medias y la covariancia de X_1, X_2 basándose en esos valores; y si existen 950 casos en que aparecen X_1

y X_3, lo mismo, y así sucesivamente. Las covariancias diferentes pueden sustituirse entonces en una matriz de covariancia y resolverse las ecuaciones de teoría de regresión resultantes. Este procedimiento puede ser desastroso si los valores que faltan no forman una serie al azar (como ocurriría, por ejemplo, si los ingresos altos tendiesen a ser omitidos). Haitovsky (1968) construyó 100 observaciones de acuerdo con la fórmula:

$$Y = 150 + 5.0X_1 - 2.0X_2 + 0.3X_3 + 3.0X_4 + \varepsilon \quad (21)$$

donde las X eran variables normales correlacionadas.

Luego eliminó 6 y, 25 x_1, 15 x_2, cero x_3 y 10 x_4. El procedimiento se repitió siete veces, con el mismo número de eliminaciones, pero con distintos miembros eliminados. Las x_1 se eliminaron en forma parcialmente sistemática, diez de los valores más altos y las otras 15 al azar. Los resultados promedios, basados en la estimación de las covariancias de diferentes números de la muestra, fueron los siguientes:

	Constante	x_1	x_2	x_3	x_4
Valores verdaderos	150.0	5.0	- 2.0	0.3	3.0
Mínimos cuadrados ordinarios sobre los 100 valores	150.732	4.968	- 1.922	0.514	2.922
Estimados como se indica	414.443	4.116	- 0.660	- 6.582	2.699

TABLAS DE CONTINGENCIA DE VARIABLES MULTIPLES

24. En el pasado, el material sobre relaciones se presentaba generalmente en forma de cuadros de doble entrada, sobre todo tratándose de datos que se clasifican en categorías. A veces se daban cuadros de tres y hasta de cuatro entradas, especialmente cuando la clasificación era sencilla (por ejemplo, una dicotomía según el sexo). Pero las dificultades de tabulación, impresión y sobre todo de interpretación han impedido o al menos limitado las tabulaciones por más de dos variables a la vez.

25. En las dos últimas décadas se aprendió bastante acerca de estos cuadros de entrada múltiple y de los métodos mecánicos para su análisis. Ahora se dispone de varios programas para este efecto. En particular, Goodman (Chicago), Nelder (Londres) y Brown (Los Angeles), han diseñado programas especialmente para este objeto. Mayor información al respecto puede solicitarse a EMF.

26. Sin embargo, estos programas no pueden aplicarse a ciegas y para utilizarlos mejor es conveniente conocer algo sus fundamentos teóricos. Existe al respecto una abundante literatura. Pueden señalarse como un buen resumen la monografía de Plackett (1974) sobre *Contingencia de viarables múltiples* ("Multivariate Contingency") y uno de los capítulos del libro de Kendall (1975) sobre *Análisis de variables múltiples* ("Multivariate Analysis"); Bishop *et al* (1975) hacen una exposición más amplia. Estos libros son los más recientes, pero el tema sigue desarrollándose con bastante rapidez. En lo que resta de esta sección se describe muy brevemente la clase de problemas que se presentan.

27. Las tablas de contingencia pueden consistir en una categorización ordenada, en una categorización desordenada o en una combinación de ambas (por ejemplo, clasificación por clase social, que es ordenada; educación, que es ordenada; grupo étnico, desordenada; área geográfica, desordenada). El procedimiento que se describe en las líneas que siguen se aplica por igual a ambos tipos.

28. Otra distinción entre los tipos de categorización es análoga a la que se encuentra en la teoría de las variables continuas. Por un lado pueden elegirse algunas variables para estudiarlas como dependientes de otras (análisis de dependencia); y por otro, el interés puede radicar en la relación de un grupo de variables entre sí (análisis de interdependencia). Ejemplos del primero son la regresión y el análisis de variancia; ejemplos del segundo son el análisis de los componentes, el análisis factorial y el análisis de conglomerados. Afortunadamente para la reducción del número de hipótesis que deben considerarse, el primer caso es más frecuente que el segundo.

29. En el análisis de contingencia múltiple existen dos problemas fundamentales. El primero consiste en elaborar una medida de la relación entre dos o más variables. Esto se hace habitualmente usando la X^2 estadística o alguna función de ella. El otro consiste en encontrar el camino a través de un cúmulo de hipótesis posibles de manera sistemática.

Por ejemplo, en un cuadro de doble entrada es costumbre comparar la frecuencia observada en cada casilla (supongamos F) con la frecuencia que se habría observado si las variables fueran independientes (supongamos T), la última de las cuales se calcula considerando fijos los totales marginales de una entrada. Existen entonces dos medidas (que

son asintóticamente equivalentes) de uso habitual para probar la hipótesis de independencia:

$$X^2 \text{ (Pearson)} = \sum \frac{(F-T)^2}{T} \qquad (22)$$

$$X^2 \text{ (razón de probabilidad)} = 2 \sum F \log \frac{F}{T} \qquad (23)$$

donde la suma se efectúa en las casillas del cuadro.

Antes de la computadora de bolsillo, el primero era más fácil de calcular, pero es preferible el segundo.

30. En el caso de un cuadro de triple entrada (supongamos las variables *A, B, C*), ya no se trata de probar una sola hipótesis sino 17, algunas de las cuales son triviales. Puede mostrarse en la siguiente forma:

A	*A, B*	*A, B, C*	*AB*	*AB, C*	*AB, C*	*AB, AC*	*ABC*
B	*A, C*		*BC*	*BC, A*	*BC, BA*		
C	*B, C*		*CA*	*CA, B*	*CB, CA*		

Aquí, por ejemplo, *A, B* se refiere a una hipótesis basada en la "fijación" de los márgenes de una variable de *A* y *B*. *AB* representa la hipótesis de que todo el cuadro de triple entrada está determinado por la distribución conjunta de *A* y *B*. *AB, AC* es una prueba de "fijación" de los márgenes de doble entrada de *AB* y *AC*. En realidad, la prueba de que una simple variable *A* "explica" todo el cuadro es trivial: ella prueba simplemente si las frecuencias en la misma categoría de *A* son todas iguales dentro de los límites del muestreo. Lo mismo ocurre con *B* y con *C*. Asímismo, el modelo *ABC* (que se da para tener el cuadro completo) tampoco necesita ser probado por que fija todas las casillas del cuadro; es el modelo "saturado". Las otras 13 en cambio pueden ser interesantes. Una prueba basada en *AB, C*, por ejemplo, es semejante a la prueba de una correlación parcial - ¿son *A* y *B* dependientes cuando se abstrae el efecto de *C*? .

31. El número de posibilidades por examinar aumenta en forma alarmante con el número de dimensiones. Para los cuadros de cuádruple entrada existen 167 y para los de quíntuple entrada hay miles. Dentro de los límites de esta nota es imposible realizar un análisis sistemático detallado. A veces, la especificación previa de las hipótesis en es-

tudio reducirá el número de posibilidades que se van a examinar. Cuando esto no se hace, parece mejor comenzar por los modelos más simples para luego irse remontando hacia los modelos más complejos y detener el análisis cuando se ha alcanzado un modelo sintético (es decir, que tiene el menor número de parámetros).

Se espera preparar un boletín técnico en que se trate este tema en forma más detallada.

TRANSFORMACIONES DE VARIABLES

32. Puesto que, aún en la era del computador, las matemáticas lineales son relativamente simples, se ha observado de parte de los estadísticos la tendencia a dedicar casi toda su atención a modelos expresados en forma lineal. (La ecuación de regresión (13) es un ejemplo ilustrativo). Tales modelos imponen una fuerte limitación en los datos y es muy conveniente considerar al principio si la linealidad es realista y si no, qué puede hacerse para mejorar el modelo.

33. Existen dos procedimientos tradicionales:

(i) Si se ha concebido un modelo multiplicativo y no aditivo como en el tipo de función de la demanda en economía, de Cobb-Douglas), puede llegarse a la linealidad trabajando con los logaritmos de los datos en lugar de los datos originales.

(ii) Aun si las relaciones no son lineales, el margen de interés puede ser lo bastante estrecho como para que una relación curvilínea pueda ser adecuadamente representada por una línea recta.

34. Aparte de esto, hay muchas circunstancias en que es conveniente introducir modificaciones en las variables antes de someterlas al análisis matemático. Por ejemplo, en el plano de la fecundidad, considérese:

(i) El impacto marginal de las variables de base.

En ciertos contextos, como la relación entre el estímulo y la respuesta en psicofísica (Ley de Weber-Fechner) o la relación entre el ingreso y su utilidad en economía, el efecto de un cambio en la primera variable sobre la segunda depende del nivel y de la cantidad del cambio ocurrido en la primera variable. Puede que existan razones para esperar relaciones similares entre las variables de las EMF. Por ejemplo, un año más de educación o una unidad más de ingreso puede muy bien tener un efecto reducido sobre la fecundidad o las

intenciones de fecundidad si el nivel de educación o de ingreso es ya alto.

(ii) El impacto marginal de las variables intermedias

En el estudio del efecto de algunas variables sobre otras, es difícil especificar las relaciones matemáticas entre las variables intermedias (por ejemplo, la costumbre de una lactancia prolongada y la duración de la amenorrea post-partum) y su relación con la fertilidad o la fecundidad. Sin embargo, el modelo lineal habitual es claramente inapropiado para la mayoría de las relaciones y su adopción indiscriminada produciría a lo sumo una primera aproximación. De modo similar, es posible, por ejemplo, que la probabilidad de que una mujer conciba en un mes sea una función lineal de la frecuencia de las relaciones sexuales, porque la probabilidad está limitada por cero y la unidad.

En algunos casos, el efecto del cambio de una variable sobre la esperanza condicional de otra variable puede suponerse mediante una investigación relacionada anterior. En algunos casos se ha incluído una relación en la definición de las variables. Por ejemplo, una tasa general de fecundidad es el producto de una tasa de fecundidad marital y la proporción de casados, y una relación similar aparece en los índices de fecundidad marital de Coale. A veces se emplea un procedimiento gradual y sistemático, ya aludido, incluyendo progresivamente interacciones de mayor orden y polinomios hasta lograr el "ajuste" mejor. En otros casos se puede disponer de los puntos de los datos obtenidos mediante el computador y, mediante examen, identificar una pauta de las esperanzas condicionales.

35. Existe otro tipo de transformaciones cuya función consiste en alcanzar la homocedasticidad del término "error". Estas son distintas de las transformaciones que produce la linealidad de las esperanzas condicionales. Los dos tipos pueden usarse conjuntamente.

Si las variables pueden ser transformadas para lograr su linealidad y homocedasticidad, se pueden emplear entonces los procedimientos de estimación de los mínimos cuadrados usuales. Sin embargo, ya no existe una limitación para esta clase de forma final. Nelder y Wedderburn han descrito un método para estimar parámetros cuando las observaciones se distribuyen de acuerdo con una familia exponencial. Existen asimismo poderosos programas iterativos de computador que permiten estimar los parámetros prácticamente en cualquier forma de ecuación.

REFERENCIAS

E.M.L. Beale and R.J.A. Little, *Missing Values in Multivariate Analysis.* (J. Roy. Statist. Soc. B, 37, 129, 1975).

Y.M.M. Bishop, S.E. Fienberg and P.W. Holland, *Discrete Multivariate Analysis*. (M.I.T. Press, 1975).

Y. Haitovsky, *Missing Data in Regression Analysis.* (J. Roy. Statist. Soc. B, 30, 67, 1968).

Haitovsky, Y. *Regression Estimation from Grouped Observations.* (Charles Griffin & Co., 1973).

M.G. Kendall, *Multivariate Analysis.* (London: Charles Griffin & Co., 1975).

R.L. Plackett, *The Analysis of Categorical Data.* (London: Charles Griffin & Co., 1974).

J.A. Nelder and H.W.M. Wedderburn, *Generalised Linear Models.* (J. Roy. Statist. Soc. A, 135, 370, 1971).