Cross Classifications, III: Aproximate Sampling Theory", *Journal of the American Statistical Association*, vol. 58, pp. 310-364, 1963.
11. Haggard, E. A.: *Intraclass Correlation and the Analysis of Variance*, The Dryden Press, Inc., Nueva York, 1958, pp. 22-26.
12. Hagood, M. J., y D. O. Price: *Statistics for Sociologists*, Henry Holt and Company, Inc., Nueva York, 1952, cap. 23.
13. Hays, W. L.: *Statistics*, Holt, Rinehart and Winston, Inc., Nueva York, 1963, cap. 16.
14. Heise, D. R.: "Separating Reliability and Stability in Test-Retest Correlation", *American Sociological Review*, vol. 34, pp. 93-101, 1969.
15. Johnston, J.: *Econometric Methods*, McGraw-Hill Book Company, Nueva York, 1963, Parte II.
16. Kendall, M. G.: *Rank Correlation Methods*, Hafner Publishing Company, Inc., Nueva York, 1955, caps. 1, 3 y 4.
17. Kruskal, W. H.: "Ordinal Measures of Association", *Journal of the American Statistical Association*, vol. 53, pp. 814-861, 1958.
18. Mueller, J. H., K. Schuessler, y H. L. Costner: *Statistical Reasoning in Sociology*, 2ª ed., Houghton Mifflin Company, Boston, 1970, cap. 10.
19. Siegel, Sidney: *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill Book Company, Nueva York, 1956, cap. 9.
20. Somers, R. H.: "A New Asymmetric Measure of Association for Ordinal Variables", *American Sociological Review*, vol. 27, pp. 799-811, 1962.
21. Wallis, W. A., y H. V. Roberts: *Statistics: A New Approach*, The Free Press of Glencoe, Ill., Chicago, 1956, cap. 17.
22. Wilson, T. P.: "A Critique of Ordinal Variables", *Social Forces*, vol. 49, pp. 432-444, 1971.
23. Wilson, T. P.: "A Proportional-Reduction-in-Error Interpretation for Kendall's tau-b", *Social Forces*, vol. 47, pp. 340-342, 1969.

XIX. CORRELACIÓN MÚLTIPLE Y PARCIAL

En los dos últimos capítulos nos hemos ocupado de la relación entre dos escalas de intervalo, entre una variable dependiente y una sola variable independiente. Los análisis de correlación y regresión pueden extenderse fácilmente para comprender cualquier número de escalas de intervalo, una de las cuales puede tomarse como dependiente, y las demás como independientes. El problema se puede concebir como un problema de predicción en el que tratamos de predecir una variable dependiente Y a partir de las variables $X_1, X_2, \ldots, X_k$. Habremos de servirnos de nuevo de un modelo muy sencillo, que será directamente análogo a la regresión lineal, excepto en cuanto al hecho de que habrá más de dos dimensiones.

El concepto de correlación se generalizará en dos formas. Emplearemos el término de *correlación parcial* para designar la correlación entre dos variables cualesquiera cuando los efectos de otras variables se han controlado. El de *correlación múltiple*, en cambio, servirá para indicar qué tanto de la variación total de la variable dependiente puede explicarse por todas las variables independientes actuando conjuntamente. Veremos que los materiales examinados en el presente capítulo comportan en su mayor parte extensiones directas de razonamientos presentados anteriormente. Una vez que hayamos ampliado las nociones de correlación y regresión, estaremos en condiciones, en el capítulo siguiente, de emprender el análisis de covariancia, que comporta una combinación de las técnicas de regresión con el análisis de la variancia.

XIX.1. *Regresión múltiple y mínimos cuadrados*

En la regresión múltiple tratamos de predecir una sola variable dependiente a partir de cualquier número de variables independientes. Si se da un gran número de variables de escala de intervalo que deban relacionarse entre sí, será posible, por supuesto, predecir cualquier variable particular a partir de cualquier combinación de las demás. Por lo regular resultará claro del contexto cuáles variables han de considerarse como independientes y cuáles como dependientes.[1] Así, por ejemplo, puede querer predecirse el éxito en la universidad a partir de una serie de marcas de aptitud y del éxito en la escuela secundaria. O puede resultar posible predecir la tasa de crecimiento de una ciudad

[1] Cuando se crea que existe una causación recíproca, o retroalimentación, de la variable "dependiente" hacia alguna de las demás, deberán emplearse ecuaciones simultáneas en lugar de mínimos cuadrados. Véanse [4] y [12].

conociendo factores como la magnitud actual, los porcentajes de mano de obra en las diversas ocupaciones, o la magnitud y la distancia del gran centro urbano más próximo.

En el análisis de regresión múltiple definimos la ecuación de regresión como el curso de la media de la variable dependiente Y para todas las combinaciones de $X_1, X_2, \ldots, X_k$. En otros términos: para cada combinación de X fijas habrá una distribución de las Y. Cada distribución tendrá una media $\mu_{Y|X_1, X_2, \ldots, X_k}$ y una desviación estándar $\sigma_{Y|X_1, X_2, \ldots, X_k}$, y habremos de suponer una vez más que todas estas distribuciones son normales y que las desviaciones estándar son iguales (homoscedasticidad). El recorrido de las medias ya no seguirá siendo una curva en el espacio bidimensional, sino que será, antes bien, una especie de hipersuperficie en un espacio de $(k+1)$ dimensiones. Es obvio que ya no estaremos en condiciones de representar un curso semejante, excepto en el caso en que sólo tengamos dos variables independientes X_1 y X_2.

En el capítulo anterior supusimos una ecuación de regresión lineal de la forma $Y = \alpha + \beta X$. Y habremos de volver a suponer una forma sencilla de la ecuación de regresión. Supongamos, pues, que el curso de las medias de Y adopta la forma:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \qquad \text{(XIX.1)}$$

en donde $\alpha, \beta_1, \beta_2, \ldots, \beta_k$ son constantes. Ésta es la ecuación más sencilla posible de regresión múltiple, y es directamente análoga a la regresión lineal en el caso de dos variables. En efecto, si todas las β, excepto una, son cero, el problema se reduce al caso bidimensional.

Si podemos suponer una población "normal multivariable" en la que cada variable esté distribuida normalmente alrededor de todas las demás, entonces podemos satisfacer los tres supuestos requeridos. En otros términos, una distribución normal multivariable nos asegura que las ecuaciones de regresión serán de la forma anterior, que las distribuciones de las Y para X determinadas serán todas normales, y que las variancias serán también iguales. Esto constituye una generalización obvia de las propiedades de la distribución normal bivariable. Sobra decir que la distribución normal multivariable no puede representarse geométricamente (pese a que tiene una ecuación algebraica perfectamente definida), toda vez que tuvimos ya necesidad de tres dimensiones para representar el caso bivariable.

Con objeto de proporcionar una mejor comprensión intuitiva de la naturaleza de las extensiones implicadas, será conveniente examinar el caso en que no hay más que dos variables independientes (véase la figura XIX.1). La ecuación de regresión $Y = \alpha + \beta_1 X_1 + \beta_2 X_2$ puede representarse en este caso por medio de un

plano en un espacio tridimensional. Si dejamos que $X_1 = X_2 = 0$, obtenemos $Y = \alpha$, lo que indica que el plano de regresión corta el eje de las Y a una altura α. Con objeto de obtener una interpretación de las β, tomamos las intersecciones del plano de regresión con planos perpendiculares a los ejes de X_1 y X_2. Así, por ejemplo, si tomamos un plano perpendicular al eje de X_2,

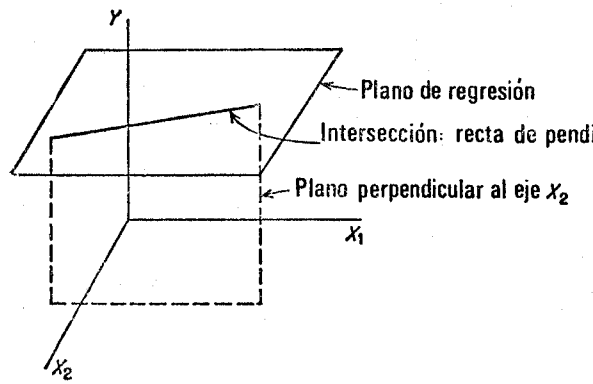

FIG. XIX.1. *Interpretación geométrica de la regresión múltiple de* Y *sobre* X_1 *y* X_2.

mantenemos de hecho a X_2 constante, ya que todos los puntos situados en este plano tendrán el mismo valor de X_2. Este plano corta el plano de regresión en una recta, y la pendiente de esta recta será β_1. En otros términos, si mantenemos X_2 en un valor fijo, β_1 representa la pendiente de la *línea* de regresión de Y a X_1. Y en forma análoga, el hecho de mantener constante a X_1 nos da un plano que intersecta el plano de regresión en una línea de pendiente β_2.

Conviene observar que las β empleadas en la regresión múltiple no serán por lo regular las mismas que las que se obtuvieron en el caso de dos variables. Designando el caso de dos variables como regresión *total*, vemos que la β empleada en la regresión total se obtiene *prescindiendo* de las demás variables independientes, y no manteniéndolas constantes. Las β obtenidas en las ecuaciones de regresión múltiple se designan como coeficientes *parciales*, porque comportan pendientes que se obtendrían eliminando o manteniendo constantes cada una de las demás variables independientes consideradas en la ecuación de regresión.

El concepto de los mínimos cuadrados puede ampliarse en una forma semejante. Como quiera que es casi siempre necesario apreciar una ecuación de regresión adaptando una a los datos empíricos, habremos de requerir una vez más que la ecuación de

estimación revista una forma particular y se sirva del criterio de los mínimos cuadrados para conseguir el "mejor" ajuste. Nos serviremos de una ecuación de mínimos cuadrados de la forma:

$$Y_p = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k \qquad \text{(XIX.2)}$$

y volverá a resultar que, a condición que la ecuación de regresión sea efectivamente de la misma forma, la ecuación de los mínimos cuadrados representa la mejor estimación de la ecuación de regresión. En otros términos, si nos servimos de a para estimar α, y de b_i para estimar β_i, estas estimaciones serán insesgadas y, al propio tiempo, de eficiencia máxima. Por consiguiente, nuestra atención puede fijarse en el análisis de los mínimos cuadrados como método práctico de estimar una ecuación teórica que se aplica a la población. Si sólo hay dos variables independientes, ajustaremos una serie de puntos en el espacio tridimensional con un plano de mejor ajuste. En un espacio de $(k+1)$ dimensiones, por su parte, ajustaremos puntos con un hiperplano de k dimensiones, si es que semejante figura se puede concebir.

Tomando el caso tridimensional, reduciremos al mínimo la cantidad $\Sigma(Y - Y_p)^2$, que representa la suma de las desviaciones al cuadrado respecto del plano de mínimos cuadrados en la dimensión vertical de Y (véase la figura XIX.2). El resultado será un plano único de mejor ajuste, determinado por valores específicos de a, b_1 y b_2. Según veremos, puede utilizarse luego un coeficiente de correlación múltiple para medir la bondad de ajuste de los puntos al plano de mínimos cuadrados. Sería también posible, por supuesto, medir el grado de ajuste mediante una desviación estándar referida al plano, y podríamos comparar esta desviación con la desviación estándar en relación con la $\overline{Y}$ fija (representada ahora como *plano* perpendicular al eje de las Y). Algebraicamente, el caso más general es una ampliación directa del caso de tres variables. La cantidad $\Sigma(Y - Y_p)^2$ se minimiza, y habrá que calcular $(k+1)$ coeficientes, esto es, a, b_1, b_2, $\ldots$ b_k. El cálculo efectivo de estos coeficientes será posible examinarlo más adelante, cuando hayamos efectuado el estudio de la correlación parcial.

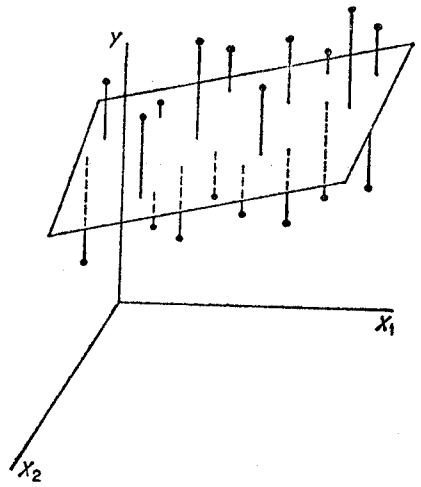

FIG. XIX.2. *Plano de mínimos cuadrados, que reduce al mínimo las sumas de las desviaciones al cuadrado en la dimensión vertical* Y.

XIX.2. *Correlación parcial*

Podemos servirnos de este modelo de regresión múltiple para obtener medidas del grado de relación entre una variable dependiente Y y cualquiera de las variables independientes, controlando una o más de ellas. El término de *correlación parcial* se emplea para designar este tipo de procedimiento de control, el cual, según veremos, es básicamente muy similar al referente al análisis de la variancia por dos métodos. En la correlación parcial controlamos ajustando valores de las variables dependientes e independientes con objeto de tomar en cuenta las puntuaciones de las variables de control. Para comprender la naturaleza de la correlación parcial y el procedimiento de ajuste, limitaremos por ahora nuestra atención a los problemas más sencillos, en los que figuran sólo tres variables, y supondremos modelos de regresión lineal entre las tres combinaciones de variables tomadas de dos en dos.

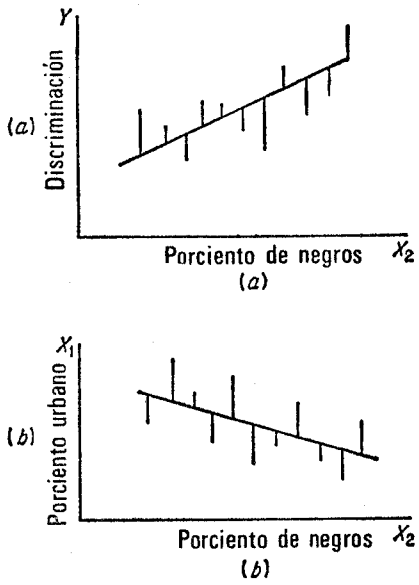

FIG. XIX.3. *Rectas de mínimos cuadrados indicando los residuos entre: a)* Y *y* X_2, *y b) entre* X_1 *y* X_2 ...

Supongamos que queremos medir el grado de relación entre una variable dependiente Y y una variable independiente X_1, controlando en relación con otra variable independiente X_2. Para servirnos de un ejemplo concreto, podemos tener interés en predecir la tasa de discriminación económica contra los negros, medida por las diferencias de ingreso entre los blancos y los negros, y el grado de urbanización, según resulta del porcentaje de un distrito designado como urbano. Se espera con seguridad que el porcentaje de negros en el distrito afectará asimismo la tasa de discrimi-

nación, y se decide en consecuencia utilizar el porcentaje de negros como variable de control.

Supóngase que las líneas de mínimos cuadrados entre la discriminación Y y el porcentaje de negros X_2 y entre el porcentaje urbano X_1 y el porcentaje de negros son como las que se indican en la figura XIX.3. La relación entre la discriminación y el porcentaje de negros es positiva, lo que indica que tasas elevadas de discriminación se hallan asociadas a porcentajes elevados de la minoría en cuestión. Por otra parte, la relación entre el índice de urbanización y el porcentaje de negros es negativa. Sobre la base de la sola información, anticiparíamos una relación negativa entre las tasas de discriminación y la urbanización. En otros términos: las áreas urbanas podrían tener tasas bajas, debido simplemente al hecho de que en promedio cuentan con relativamente menos negros. Supóngase, sin embargo, que estuviéramos en condiciones, en alguna forma, de "forzar" todos los distritos a tener los mismos porcentajes de la minoría. Podríamos eliminar en esta forma el efecto perturbador de esta tercera variable. En realidad, por supuesto, no podemos hacer efectivamente todos los porcentajes de la minoría iguales, pero podemos por lo menos efectuar ajustes en relación con el hecho de que difieren. Como quiera que sabemos (o podemos apreciarla) la relación entre la variable de control y cada una de las otras dos variables, podemos predecir cómo se comportarían éstas respecto de cambios en la variable de control.[2] De hecho, las ecuaciones de mínimos cuadrados representadas en la figura XIX.3 constituyen nuestras ecuaciones de predicción y pueden emplearse en el proceso de ajuste.

Al relacionar la discriminación Y al porcentaje de negros X_2, podemos concebir la variación de las tasas de discriminación como resultado de dos componentes, siendo la primera de ellas el porcentaje de negros y, la segunda, debiéndose a otros factores, uno de los cuales puede ser la urbanización. Como ya vimos, esta segunda componente puede representarse como *desviaciones* respecto de la ecuación de mínimos cuadrados que comporta Y y X_2. En términos de X_2, estas desviaciones o *residuos* representan error. Aun si X_2 se mantuviera constante, subsistirían. En estos residuos, por consiguiente, son en los que estamos en verdad interesados, ya que representan la cantidad de variación en la discriminación que subsiste una vez que el porcentaje de negros ha explicado todo lo que podía de la variación.

Y en forma análoga, nos interesaremos también en los residuos o desviaciones respecto de la ecuación empleada para predecir el porcentaje urbano a partir del porcentaje de negros. En otros

[2] Debe insistirse de nuevo en que la justificación para interpretar resultados de tal ajuste, hecho con lápiz y papel, implica el supuesto causal de que la variable de control puede afectar a las otras dos variables.

términos: dejamos que el porcentaje de negros explique la variación en las otras dos variables todo lo que pueda. Y si ahora ponemos los residuos en relación de unos con otros, obtenemos una medida de la relación entre Y y X_1 que es independiente de los efectos de X_2. *La correlación parcial entre* Y *y* X_1 *controlando* X_2, *puede definirse como la correlación entre los residuos de las regresiones de* Y *sobre* X_2 *y de* X_1 *sobre* X_2. En cierto sentido, pues, la correlación parcial representa la correlación entre "errores" respecto de la variable de control.

El que tenga algún sentido controlar en relación con X_2 correlacionando residuos podrá parecer oscuro aún. Quizá la explicación sea más aceptable si examinamos más de cerca una relación hipotética entre dichos residuos. Supongamos, por ejemplo, que para el distrito A encontramos un gran residuo negativo al correlacionar Y con X_2. Esto significa que el distrito A presenta considerablemente menos discriminación de lo que se esperaría conociendo solamente su porcentaje de minoría. El punto que representara dicho distrito particular se situaría en algún lugar por debajo de la línea de mínimos cuadrados. Supóngase, asimismo, que el residuo para este mismo distrito fuera positivo al correlacionar X_1 con X_2. En tal caso sabemos que el distrito en cuestión está más urbanizado de lo que se esperaría conociendo solamente su porcentaje de minoría. Tenemos, por lo tanto, un distrito relativamente urbanizado con tasas bajas de discriminación, y sabemos, además, que dichos valores son altos y bajos respectivamente en comparación con otros distritos del mismo porcentaje de minoría. No podemos, por consiguiente, atribuir la relación negativa entre los residuos al hecho de que la cifra del porcentaje de negros acontezca ser alta o baja. Y en forma análoga, el distrito B puede tener grandes residuos positivos para Y, pero negativos para X_1. Por consiguiente, este distrito tendría mayores tasas de discriminación de lo que se esperaba, pero estaría al propio tiempo menos urbanizado que otros distritos del mismo porcentaje de minoría. Es obvio que si muchos distritos son similares a A o B, obtendremos una correlación negativa, entre los residuos, indicando una correlación negativa entre la discriminación y la urbanización, ajustando en relación con el porcentaje de negros.

La correlación parcial da una sola medida que resume el grado de relación entre dos variables al controlar en relación con otra. Según veremos al examinar los procedimientos de cálculo, el razonamiento puede extenderse fácilmente a variables de control adicionales. Podemos concebir varias ecuaciones de regresión múltiple, una de las cuales comporte Y y todas las variables de control, y la otra relacionando X_1 con estas mismas variables. Pueden obtenerse los residuos de cada una de esas ecuaciones de regresión múltiple y relacionarlos luego. Ajustaremos en esta

forma en relación con todas las variables de control al mismo tiempo. El punto importante, aquí, es que sólo obtenemos una correlación parcial, en tanto que al controlar con las tablas de contingencia (con concesiones para la interacción) obteníamos una medida separada para cada una de las categorías de las variables de control.

En el capítulo XV vimos que el grado de relación entre dos variables podía variar de una categoría de la variable de control a otra. Así, por ejemplo, si el porcentaje de negros se hubiera categorizado, es perfectamente posible que hubiéramos obtenido una elevada correlación negativa entre la discriminación y la urbanización para distritos de porcentajes de minoría muy bajos, pero con una correlación positiva, de todos modos, en el extremo opuesto del continuo porcentaje de negros. Por lo tanto, el hecho de que en la correlación parcial hayamos obtenido una sola medida de resumen puede acaso oscurecer cierta información acerca de la interacción.

Resulta que el coeficiente de correlación parcial puede ser también interpretado como un *promedio ponderado* de los coeficientes de correlación que se hubieran obtenido si la variable de control hubiera sido dividida en muy pequeños intervalos y calculando correlaciones separadas dentro de cada una de estas categorías. La naturaleza exacta de este procedimiento de ponderación carece de importancia, ya que nunca se hace uso de él en la práctica. No tendría, por tanto, sentido pensar que las correlaciones parciales relacionan dos variables que "mantienen constante" a una tercera, ya que la fuerza de su relación puede variar de acuerdo con el valor particular en que se mantiene constante la variable de control.

En el caso de la distribución normal multivariable, sabemos que todas las ecuaciones de regresión tendrán la forma especial descrita por la ecuación (XIX.1). Pero la distribución normal multivariable posee además otra propiedad notable. Y es que la fuerza de la relación entre dos variables será la misma independientemente de los valores de las variables de control. En otros términos: si se seleccionara un gran número de categorías de una variable de control y se obtuvieran correlaciones dentro de cada una de dichas categorías, todas las correlaciones tendrían el mismo valor. Por lo tanto, la correlación parcial tendría el mismo valor que cada una de esas correlaciones dentro de las categorías. En este caso especial, tendría así cierto objeto pensar en términos del mantenimiento constante de la tercera variable de control. Sin embargo, como quiera que en el mejor de los casos sólo podemos aproximarnos a la distribución normal multivariable con datos reales, es más seguro pensar en la correlación parcial como promedio ponderado, o como si comportara un ajuste en relación con la variable de control.

Cálculo de los coeficientes de correlación parcial. El cálculo de las correlaciones parciales es sumamente sencillo, a menos que se desee controlar en relación con tres o más variables a la vez. Antes de presentar la fórmula de la correlación parcial, hemos de introducir un cambio de notación. Por desgracia, lo que constituye una notación conveniente para un objeto no lo es para otro, ni es el uso convencional totalmente concordante. Hemos venido representando la variable dependiente por Y y las variables independientes por $X_1, X_2, \ldots, X_k$. En reconocimiento del hecho de que la elección de la variable dependiente es a menudo más o menos arbitraria y que, por consiguiente, podemos querer calcular correlaciones parciales entre varias combinaciones de variables, convendrá renumerar simplemente las variables de 1 a $k+1$ y representar la correlación entre las variables 1 y 2, controlando en relación con 3 mediante $r_{12\cdot3}$. Y en forma análoga, la correlación entre las variables 2 y 3, controlando en relación con 1, por medio de $r_{23\cdot1}$.

Esta notación puede extenderse fácilmente a cualquier número de variables de control añadiendo más números a la derecha del punto central del subíndice. Así, por ejemplo, la relación entre las variables 5 y 7, con control de las variables 1, 2, 3, 4 y 6, nos vendría dada por $r_{57\cdot12346}$. El orden de las dos variables a la izquierda del punto no juega papel alguno, lo mismo que el de la derecha. Para distinguir entre parciales con números diferentes de control, designamos el número de controles como el *orden* de la correlación. Así, pues, un primer orden parcial tendrá un control; un segundo orden, dos controles, y así sucesivamente. En concordancia con esta terminología, la correlación sin controles se designa a menudo como correlación de orden cero. Según se ha indicado más arriba, el término *correlación total* se emplea también para designar una correlación entre dos variables sin controles.

Podemos dar ahora la fórmula del primer orden parcial $r_{ij\cdot k}$:

$$r_{ij\cdot k} = \frac{r_{ij} - (r_{ik})(r_{jk})}{\sqrt{1 - r_{ik}^2}\sqrt{1 - r_{jk}^2}} \qquad \text{(XIX.3)}$$

Obsérvese que la primera correlación del numerador es la correlación total entre las dos variables a relacionar (i y j). La variable de control figura en la segunda expresión del numerador, en donde se la relaciona con cada una de las otras variables, así como en ambos términos del denominador. Cualquier correlación parcial particular puede obtenerse a partir de esta fórmula general, sustituyendo i, j y k por los números apropiados. Así, por ejemplo:

$$r_{13\cdot 2} = \frac{r_{13} - (r_{12})(r_{23})}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{23}^2}}$$

En un estudio de 150 distritos del Sur [3], la correlación entre la discriminación en los ingresos y el porcentaje de negros fue de .536; aquélla entre la discriminación en los ingresos y el porcentaje urbano fue de .139, y la correlación entre los porcentajes de negros y urbano fue de −.248. Si llamamos al índice de discriminación variable 1, al porcentaje de negros variable 2 y al porcentaje urbano variable 3, podemos obtener la correlación parcial entre la discriminación y el porcentaje urbano, controlado en relación con el porcentaje de negros. Tenemos así:

$$r_{13\cdot 2} = \frac{.139 - (.536)(-.248)}{\sqrt{1 - (.536)^2}\sqrt{1 - (-.248)^2}} = \frac{.2719}{.8178} = .332$$

Este resultado puede interpretarse como correlación entre la discriminación y el porcentaje urbano una vez que se ha dejado que el porcentaje de negros explique todo lo que puede de ambas variables.

Si bien no resultará inmediatamente evidente que la fórmula anterior pueda derivarse de la definición de la correlación parcial en términos de una correlación de residuos, la fórmula de cálculo, por lo menos, tiene un sentido. En efecto, en el numerador sustraemos esencialmente un factor de corrección de la correlación total. En cuanto al denominador, éste consta de dos factores de corrección, ninguno de los cuales puede ser mayor que la unidad, que toman en cuenta el hecho de que la variable de control explica cierta proporción de la variación de las otras variables. Si elevamos al cuadrado el coeficiente de correlación parcial, el número resultante representará la proporción de variación de la variable 1 (discriminación), dejada inexplicada por 2 (porcentaje de negros), pero que puede explicarse por los valores ajustados de X_3 (porcentaje urbano).

Examinemos la ecuación (XIX.3) con mayor atención, para ver cómo la correlación parcial se comporta en relación con las tres correlaciones totales. Con fines de simplificación, supongamos primero que r_{ij} es positiva. Si r_{ik} y r_{jk} tienen ambas el mismo signo (ya sea positivo o negativo), su producto será positivo, y el numerador será o bien un número positivo menor que r_{ij}, o será incluso cero o negativo. Por otra parte, el denominador será siempre menor que la unidad, a menos que $r_{ik} = r_{jk} = 0$. Por consiguiente, la fracción resultante puede ser casi cualquier número entre −1.0 y +1.0, según sea la magnitud de las tres correlaciones totales. Veremos más adelante exactamente lo que

podemos y lo que no podemos decir acerca del comportamiento de la parcial en estas circunstancias.

Supongamos ahora que las correlaciones con la variable de control son de signos opuestos. Obtenemos en tal caso un producto negativo a sustraer de un número positivo, y el resultado será un número positivo mayor. Esto significa que si empezamos con dos variables relacionadas positivamente y si podemos encontrar una variable de control relacionada negativamente con una de ellas pero positivamente con la otra, la parcial resultante será mayor que la correlación de orden cero. Si la correlación de la variable de control con una u otra de las otras variables acontece ser cero, el factor de corrección del numerador será cero. Pero si la variable de control se halla correlacionada ya sea positiva o negativamente con la variable restante, el denominador será menor que la unidad, y la correlación parcial volverá a ser mayor que la correlación total.

Si hubiéramos empezado con una correlación total negativa, una variable de control relacionada con cada una de las otras dos en la misma dirección (ya sea positiva o negativa) produciría una correlación negativa mayor. Sin embargo, si se relacionara con ellas en sentido opuesto, el resultado sería análogo al que se ha descrito en primer término (en donde la correlación total era positiva y el factor de corrección positivo asimismo). ¿Por qué? En cambio, si la variable de control no se relacionara con una de las otras variables, el resultado sería una correlación parcial con un valor absoluto mayor que la total. Y si la variable de control no se relacionara con *ninguna* de las otras variables, la correlación parcial sería exactamente igual, por supuesto, a la correlación total. Una vez que hayamos examinado la relación entre la correlación parcial y las interpretaciones causales, estaremos en condiciones de dar una justificación intuitiva del comportamiento de las correlaciones parciales en estas diversas condiciones.

Las fórmulas de las parciales de segundo orden o superior son directamente análogas a las de la parcial de primer orden. En efecto, vamos añadiendo simplemente variables de control sucesivas, empezando cada vez con la parcial de orden uno menos que el deseado. Así, por ejemplo, las fórmulas de $r_{ij\cdot kl}$ y $r_{ij\cdot klm}$ serán:

$$r_{ij\cdot kl} = \frac{r_{ij\cdot k} - (r_{il\cdot k})(r_{jl\cdot k})}{\sqrt{1 - r_{il\cdot k}^2}\sqrt{1 - r_{jl\cdot k}^2}} \qquad \text{(XIX.4)}$$

y

$$r_{ij\cdot klm} = \frac{r_{ij\cdot kl} - (r_{im\cdot kl})(r_{jm\cdot kl})}{\sqrt{1 - r_{im\cdot kl}^2}\sqrt{1 - r_{jm\cdot kl}^2}} \qquad \text{(XIX.5)}$$

Obsérvese que en la ecuación (XIX.4) suponemos que ya hemos controlado en relación con la variable X_k. Por lo tanto, la k aparece a la derecha del punto en las tres parciales de primer orden. Y en forma análoga, en la ecuación (XIX.5) hemos controlado previamente en relación con X_k y X_l, y de aquí que estas cantidades figuren en cada una de las parciales de segundo orden.

Las parciales de cuarto y quinto orden podrían obtenerse en forma análoga, y resultará instructivo tratar de escribir las fórmulas de estas parciales de orden superior. De modo que la manera de calcular estas últimas es idéntica a la que empleamos en el caso del primer orden. Pero el trabajo que supone se hace prontamente aburrido. Así, por ejemplo, con objeto de obtener una parcial de tercer orden con este método, han de haberse obtenido previamente tres parciales de segundo orden, cada una de las cuales ha de haberse obtenido a su vez calculando parciales de primer orden a partir de correlaciones de orden cero. Si el lector tratara de expresar la fórmula de las parciales de tercer orden directamente en términos de las correlaciones de orden cero, se daría cuenta del trabajo que esto representa.

Afortunadamente, en la investigación sociológica rara vez resulta necesario ir mucho más allá de las parciales de segundo o tercer orden. Por lo regular, la adición de controles más allá del segundo o tercer control proporciona muy pocos conocimientos nuevos. Si se hace necesario servirse de parciales de orden superior, o de ecuaciones de regresión múltiple de cuatro o más variables, existen ciertas rutinas de cálculo que facilitan considerablemente la labor. Para tratar tales problemas el lector podrá referirse ya sea al método abreviado de Doolittle o al de Dwyer, de la raíz cuadrada (véanse [9] y [11]). De estos dos métodos, el primero tal vez sea más satisfactorio, por cuanto permite obtener directamente las parciales sucesivas $r_{12\cdot3}$, $r_{12\cdot34}$, $r_{12\cdot345}$, etcétera.

Correlación parcial de rangos ordenados. La teoría de las correlaciones parciales de rangos ordenados está menos bien desarrollada. Puede extenderse al caso de las parciales de primer orden la tau de Kendall, aunque la interpretación de la tau parcial no resulta tan aceptable intuitivamente como en el caso de la correlación de producto-momento. Si no hay empates, resulta que la fórmula de la tau parcial es idéntica a la que hemos estado empleando. (Véanse [13] y [19].) Así, por ejemplo:

$$\tau_{ij\cdot k} = \frac{\tau_{ij} - (\tau_{ik})(\tau_{jk})}{\sqrt{1-\tau_{ik}^2}\sqrt{1-\tau_{jk}^2}} \qquad \text{(XIX.6)}$$

En el caso que haya un gran número de empates podrá usarse un procedimiento alternativo, sugerido por Davis [7] para el caso

de gamma, pero su principio puede aplicarse a cualquiera de las medidas de tau o a d_{yx} y d_{xy}. Si controlamos para W, categorizaremos simplemente W, computando gammas (u otras medidas) dentro de las categorías de W, obteniendo un promedio ponderado de dichas gammas. Pero en lugar de ponderar según el número de casos en cada categoría, lo haremos según el número de *pares* afectados. De esta manera, en el caso de una gamma parcial, estamos considerando tan sólo aquellos pares que no están empatados, bien en X o en Y, pero que lo *están* con respecto a la categoría de la variable de control. Davis demuestra que tal promedio ponderado puede recibir una simple interpretación de reducción proporcional en el error. Quade [16], ofrece un procedimiento análogo de promedio ponderado para el caso de tau$_a$, facilitando asimismo una prueba de significancia para dicho parcial.

En la investigación exploratoria puede tener sentido el utilizar múltiples variables de control, bien por ampliación de la fórmula (XIX.6) o dividiendo las variables de control en múltiples subcategorías. Los cimientos teóricos de tales procedimientos no son, sin embargo, muy firmes, particularmente cuando se dan numerosos empates (véase [20]). Somers [19] ha observado que en el caso de las relaciones no monotónicas marcadas, el procedimiento que Davis sugiere puede ser engañoso. Como norma general, en vista de nuestra ignorancia acerca de las propiedades y comportamiento de las medidas ordinales parciales, puede resultar prudente utilizarlas con precaución, complementándolas con medidas momento-producto aun allí donde las escalas legítimas de intervalo no estén plenamente justificadas. En un terreno ideal debe, por supuesto, intentarse mejorar los procedimientos de medición, justificando así el uso de pruebas y medidas paramétricas más poderosas.

Como está implícito en nuestras anteriores consideraciones sobre los datos ordinales, una de las razones fundamentales por las que resulta difícil llegar a conclusiones definitivas en orden a la adecuación de medidas alternativas está en que tales respuestas parecen depender del concepto que uno tenga acerca de la "realidad básica" que los datos reflejan. Ya hemos observado esto en relación con la manipulación de los empates, y, en forma implícita, con el proceso de categorización. Una manera muy prometedora de atacar este difícil problema supone la construcción de una "realidad" cuyas propiedades sean conocidas, mediante el empleo de datos originados en la computadora, o de datos simulados. Pueden, por ejemplo, crearse variables con distribuciones de frecuencia normales, rectangulares o desviadas. Pueden usarse modelos lineales o no lineales, variar las magnitudes relativas de las variancias de error y formar grupos de datos multivariados con estructuras causales conocidas (por ejemplo, X y Y con relación espuria debida a Z o varias Z_i). Los datos podrían

a continuación ser agrupados de distintas maneras, utilizando diversos procedimientos, comparando las diferentes medidas ordinales en vista de su conformidad con el comportamiento deseado. Por ejemplo: ¿se reduce casi a cero la parcial entre X y Y cuando se controla para Z, allí donde los datos han sido creados de conformidad con el modelo $X \leftarrow Z \rightarrow Y$?

Reynolds [17] ha logrado algunos resultados esperanzadores utilizando variedad de modelos, tipos de distribución de frecuencias y modelos no lineales, y mediante la introducción de cierto número de complicaciones adicionales, habiendo encontrado que si se utilizan por lo menos cinco niveles de cada variable (aunque preferentemente deban ser hasta diez), pueden lograrse muy buenas aproximaciones al comportamiento de las parciales momento-producto, utilizando diferentes procedimientos de separación y cualesquiera de las medidas τ_b, τ_c, d_{yx} o r_s, corregida para empates. Si el número de estos últimos es apreciable, los valores numéricos de las asociaciones totales que utilicen τ_a (la que no corrige para empates) tienden a ser tan bajos que resulta difícil distinguir sus valores de los de las parciales. Si el total τ_a es de solamente .20, el error de muestreo puede ser suficiente para que resulte difícil determinar si hubo o no una reducción suficientemente grande en la parcial que permita apoyar la hipótesis de que la relación es espuria.

También ha encontrado Reynolds que la gamma no se comporta tan bien bajo el seccionamiento como las otras medidas, tal vez por causa de su sensibilidad extremada ante marginales desiguales. En los casos en que el modelo correcto implica una relación espuria entre X y Y debida a W, los controles sobre W no reducen la gamma parcial a cero. Los datos de Reynolds parecen también favorecer el empleo de los procedimientos de seccionamiento de promedios ponderados por comparación con el uso de la fórmula de seccionamiento de la ecuación (XIX.6), aun cuando debe tenerse presente la advertencia de Somers relativa a las variables de control no monotónicas. Por último, y esto es importante, Reynolds ha encontrado que el seccionamiento (usando promedios ponderados), con τ_b, τ_c y d_{yx} dio excelentes resultados en el caso de las relaciones monotónicas, pero no lineales, en tanto que los procedimientos momento-producto o paramétricos no los daban. En el último caso, si se conocen las puntuaciones reales, sería preferible trabajar con modelos explícitos no lineales y procedimientos paramétricos. En ausencia de tal conocimiento, el empleo de los procedimientos paramétricos con puntuaciones asignadas arbitrariamente (y conservando el orden) dio resultados engañosos.

Debe observarse, por último, que el problema de crear medidas de correlación múltiple, usando técnicas ordinales, no ha sido estudiado sistemáticamente. Morris [15] ha encontrado incluso

que tanto la gamma como la d_{yx} tienen la indeseable propiedad de que si se forman medidas de correlación múltiple usando procedimientos plenamente razonables, la agregación de más valores explicativos puede traducirse realmente en la *disminución* de los valores de dichas dos medidas. Sugiere una medida alternativa γ_k que es una generalización multivariada de la d_{xy} (no de la d_{yx}) de Somers, como medida asimétrica de asociación múltiple ordinal más apropiada.

XIX.3. *Correlación parcial e interpretaciones causales*

Ya se señaló que el análisis de correlación no se puede emplear directamente para establecer causalidad debido al hecho de que las correlaciones sólo miden la covariación, o sea el grado en que diversas variables cambian juntas. Sin embargo, uno de los objetivos básicos de toda ciencia está en establecer relaciones causales. Independientemente de las reservas filosóficas que se puedan sentir en cuanto a las nociones de causa y efecto, es sumamente difícil pensar teóricamente en cualesquiera otros términos. En el capítulo II se señaló que existe una brecha muy real entre el lenguaje teórico, que empleamos para pensar, y el lenguaje operativo, del que nos servimos para verificar las hipótesis. El problema espinoso de la causalidad no es más que otra indicación de la existencia de dicha brecha. Pensamos a menudo en términos de relaciones causales que comportan secuencias temporales *necesarias*. Así, por ejemplo, si A es causa de B, entonces B ha de seguir necesariamente a A, y si A está ausente, B ha de estarlo asimismo. Por supuesto, este concepto de la causalidad está excesivamente simplificado. Por lo pronto, no se han tenido en cuenta otras variables, y sólo tiene sentido hablar de causa y efecto cuando se pueden establecer ciertos supuestos a propósito de esos otros factores. Por otra parte, A y B pueden variar en grado, y no simplemente estando presentes o ausentes.

Empíricamente, por supuesto, nunca podemos probar que la conexión entre dos variables sea necesaria. Podemos averiguar, en cambio, el grado en que varíen juntas, y resulta asimismo posible, en ocasiones, registrar la secuencia temporal implicada. A partir de estos dos fragmentos de información podemos formular deducciones causales si queremos. Si nuestra teoría puede demostrar una conexión lógica entre dos variables, o predecir que B seguirá a A, no necesitamos atormentarnos demasiado por el hecho de efectuar el salto intelectual a la interpretación causal. Por otra parte, si no logramos hallar razón teórica alguna para enlazar directamente dos acontecimientos, solemos, por lo regular, sentirnos más vacilantes. Tenemos mayor propensión, por ejemplo, a considerar la relación como *espuria*. Por desgracia, nada hay en el análisis de correlación que nos ayude a decidir

al respecto, *a menos* que estemos dispuestos a admitir algunos supuestos a propósito de las variables particulares consideradas y a propósito de otras, que acaso puedan producir también sus efectos. Veamos cómo habrán de ser dichos supuestos.

Supóngase que estamos investigando la relación entre el consumo *per capita* de helados y las tasas de la delincuencia juvenil. Es probable que hallemos una relación negativa. Una de las interpretaciones causales posibles sería la de pensar que los helados son tan buenos para los niños que previenen la delincuencia. Otra podría ser la de que las tasas elevadas de delincuencia hacen que los niños pierdan su gusto por los dulces. Por supuesto, descartaríamos inmediatamente dichas interpretaciones por absurdas, pese a que otras no menos absurdas se hayan tomado en serio en algún momento u otro. Se razonaría probablemente en el sentido de que la relación hallada era *espuria*, por cuanto una tercera variable, el ingreso, por ejemplo, era causa de que las dos variables variaran de tal modo que resultara de ello una correlación negativa.

Una prueba del carácter espurio, válida además si se emplea adecuadamente, consiste en controlar en relación con el nivel del ingreso. Si la correlación parcial entre el consumo de helados y la delincuencia se reduce a cero, o a cerca de cero, podemos deducir que no se da relación causal entre las dos variables. ¿Podemos, efectivamente? Tomemos otro ejemplo muy parecido. Supóngase que encontramos una relación negativa entre el nivel del ingreso y la delincuencia, y decidimos controlar en relación con el porcentaje de hogares deshechos en el área considerada. Podemos hallar de nuevo que la parcial se reduce a cero. ¿Es por ello esta relación espuria? Esta vez ya no estamos tan seguros, pese a que no haya acaso absolutamente nada en la magnitud de las correlaciones o en el comportamiento de las parciales que difiera en modo alguno del primer caso. Con el propósito de atacar el problema básico que aquí se nos plantea, volvamos atrás y considerémoslo en forma un poco más sistemática.

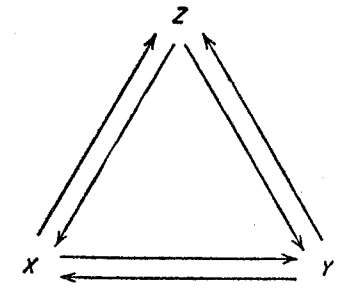

FIG. XIX.4. *Las seis flechas causales posibles entre* X, Y *y* Z.

Limitando de momento nuestra atención al caso de tres variables, observamos que se dan seis conexiones causales posibles entre éstas. Si designamos las variables como *X*, *Y* y *Z* e indicamos la dirección de la causalidad por medio de flechas, podemos trazar un diagrama de las conexiones posibles, como en la figura XIX.4. En todo problema determinado, por supuesto, algunas de esas flechas habrán de borrarse. Descartamos la posibilidad de la causalidad de doble sentido razonando en el sentido de que, si se seleccionan acontecimientos discretos, la secuencia temporal habrá de ser en un sentido o en otro, pero no en ambos a la vez.[3] Así, por ejemplo, en lugar de sostener que el desempleo produce recesión económica y viceversa, digamos que el desempleo de Jones es causa de que éste gaste menos dinero, lo cual deja a su vez sin empleo a Smith, etcétera. Nos quedamos entonces con sólo ciertas relaciones causales posibles, que se han indicado en la figura XIX.5. Con objeto de reducir el número de figuras de la figura XIX.5, se ha decidido arbitrariamente escoger a *Y* como variable dependiente, o sea como aquella que ha de ocurrir como última en cualquier secuencia temporal. De ahí que no se hayan trazado flechas de *Y* a *X* o a *Z*. De todas estas relaciones posibles, las tres primeras no revisten interés ni requieren comentario ulterior. Además, con objeto de simplificar la cosa, limitemos nuestra atención a aquellas figuras en las que sólo se han trazado dos flechas causales (*d*, *e*, *f*, *g* y *h*).

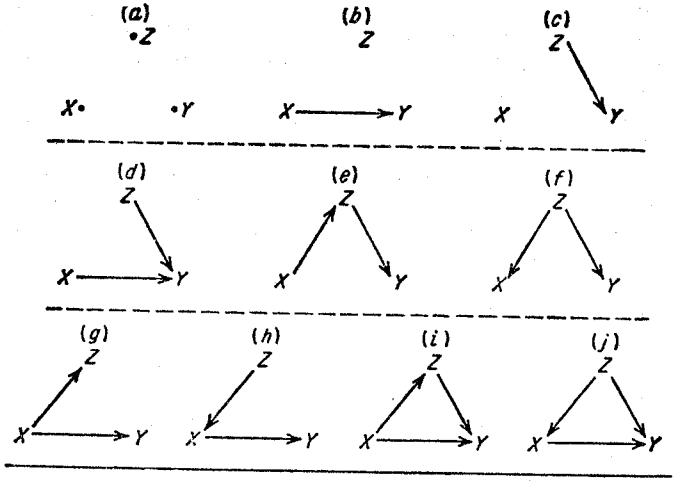

FIG. XIX.5. *Relaciones causales posibles entre* X, Y *y* Z, *tomando a* Y *como variable dependiente y excluyendo la causalidad en dos direcciones.*

[3] La mayoría de las situaciones empíricas son, por supuesto, mucho más complejas de lo que esta sencilla ilustración sugiere, requiriéndose técnicas más avanzadas, procedimiento que se aplica cuando los supuestos adecuados para las pruebas de mínimos cuadrados no se reúnen. Véanse [2], [4] y [12] para un examen más detallado de este problema.

¿Podemos distinguir entre estos varios modelos inspeccionando las magnitudes relativas de los coeficientes de correlación? La respuesta es afirmativa si estamos dispuestos a hacer dos clases de supuestos. Simon [18] ha demostrado matemáticamente lo que dichos supuestos deban ser. *Primero,* hemos de estar en condiciones de eliminar algunos de los modelos, postulando que por lo menos algunas de las relaciones posibles son *inconsistentes.* Esto ya se ha hecho hasta cierto punto, al eliminar todas las dobles flechas, así como al tomar a Y como variable dependiente, o sea suponiendo que no podía ser causa ni de X ni de Z. Habrán de hacerse además otros supuestos, pero éstos los dejamos para más adelante.

El *segundo* tipo de supuesto general que hemos de establecer se refiere a otras variables que podrían eventualmente actuar. Supongamos, siguiendo a Simon, que todas las demás variables que influyen sobre X no están relacionadas con todas las otras que afectan a Y y Z, etcétera. En otros términos: podemos admitir la existencia de otras variables incontroladas, pero hemos de suponer que la influencia que ejercen sobre X, Y y Z es esencialmente fortuita. Obsérvese que esto implica en realidad una combinación de supuestos más débil que la que suele comportar el modelo del experimento ideal, en el que se supone que todas las variables "relevantes" han sido controladas. Reconocemos la influencia perturbadora de otras variables en que no esperamos que las correlaciones sean perfectas. Por otra parte, hemos de suponer que operan de tal modo que no perturben el *patrón* de las relaciones entre X, Y y Z. Esta condición puede realizarse aproximadamente en la práctica si actúa un gran número de otras variables, ninguna de las cuales ejerce gran efecto sobre más de una de las variables consideradas.

Si existe una variable externa de efecto perturbador, deberá introducirse en el modelo como cuarta variable. Simon argumenta en el sentido de que esto es lo que debemos hacer siempre, y que el hecho de que no nos satisfaga la explicación causal en el caso de dos variables es la razón de que introduzcamos la noción de una relación espuria. Así, por ejemplo, si estuviéramos convencidos de que no existía variable tal alguna que perturbara la relación entre el consumo de helados y la delincuencia, y si pudiéramos excluir la posibilidad de que la delincuencia produce una baja de la venta de helados, entonces nada tendríamos que oponer a la explicación de que la flecha causal vaya en sentido opuesto. E introducimos el factor ingreso precisamente porque esperamos que esta última variable afecte a la relación entre las dos primeras. Y en forma análoga, añadiríamos al sistema una cuarta o quinta variable, pero hemos de estar dispuestos a detenernos en algún lugar. En este punto, si hemos de formular en principio alguna deducción causal, cualquiera que sea, hemos

de suponer que el sistema está *cerrado,* en el sentido que acabamos de describir.

Obsérvese que nos encontramos en la posición con la que estamos ya familiarizados de tener que adoptar algunos supuestos que no se dejan verificar empíricamente mediante el análisis estadístico. No será posible, por consiguiente, establecer el carácter correcto de un modelo causal particular cualquiera, pero podemos proceder por eliminación. Así, por ejemplo, uno de los modelos indicados en la figura XIX.5 podrá parecer eficaz; sin embargo, el modelo correcto podría comportar en realidad cuatro o más variables, y el cuadro presentarse en forma totalmente distinta. Con todo, habiendo adoptado los supuestos en cuestión, podemos servirnos del análisis matemático formulado por Simon, para llegar a ciertas relaciones anticipadas que deberían verificarse entre las correlaciones si el modelo particular es efectivamente correcto. Según veremos, exactamente las mismas relaciones empíricas se anticipan por algunos de los modelos anteriores, obligándonos a escoger sobre otras bases. Es aquí donde hemos de servirnos del primer tipo de supuesto examinado más arriba, o sea de que algunas relaciones causales *no* se realizan. Con todo, sin embargo, examinemos primero las predicciones matemáticas relativas a las interrelaciones entre coeficientes de correlación.

Si nos fijamos en la figura XIX.5*g*, vemos que las relaciones entre X y Y y entre X y Z son directas, en tanto que la relación entre Y y Z sólo es indirecta. Lo propio se aplica a la figura *h*. En estos dos casos, el sentido común sugeriría que, si todas las demás variables actuaran de modo esencialmente fortuito, esperaríamos encontrar que la correlación entre Y y Z es menor en magnitud que cualquiera de las otras dos. Y en forma análoga, en las figuras XIX.5*e* y *f* esperaríamos que la relación entre X y Y fuera la más pequeña de las tres, prescindiendo de los signos. Como lo revelan las matemáticas, podemos incluso pronunciarnos en forma más categórica. Es posible, en efecto, demostrar tanto para (*g*) como para (*h*), en las que la relación entre Y y Z es indirecta, que:

$$\varrho_{YZ} = \varrho_{XY}\varrho_{XZ}$$

Nos hemos servido de las ϱ para indicar que estas relaciones exactas sólo pueden esperarse en la población, y que los valores de las r de la muestra se apartarán por lo regular de esa relación estimada, a causa de las fluctuaciones de la muestra. Y en forma análoga puede demostrarse que para los casos (*e*) y (*f*) tendremos:

$$\varrho_{XY} = \varrho_{XZ}\varrho_{YZ}$$

Toda vez que los valores absolutos de los coeficientes de correlación no pueden ser mayores que la unidad, está claro que en

el primer caso el valor numérico de ρ_{YZ} ha de ser menor que el de cualesquiera de los otros coeficientes, a menos que uno de estos valores acontezca ser la unidad. En este caso especial, por supuesto, una de las variables puede predecirse exactamente a partir de una de las otras, y tenemos así esencialmente un problema de sólo dos variables.

Fijándonos con mayor detenimiento en la primera de estas ecuaciones, que se aplica a las figuras XIX.5*g* y *h*, vemos inmediatamente que si esta ecuación se verificara, la correlación parcial (en la población) entre *Y* y *Z*, controlando en relación con *X*, desaparecería, ya que el numerador de la fórmula de la parcial sería en tal caso cero. Así, pues, si (*g*) o (*h*) se verificaran, el valor de $r_{YZ \cdot X}$ debería ser cero o muy cerca de cero, habida cuenta de los errores de muestreo. Y en forma semejante, puede esperarse tanto para (*e*) como para (*f*) que la parcial entre *X* y *Y*, controlando respecto de *Z*, sea aproximadamente cero. ¿Qué indican estos hechos? Si limitamos nuestra atención a una comparación entre (*e*) y (*f*), ya que la relación entre (*g*) y (*h*) puede compararse directamente si se intercambian *X* y *Z*, vemos que en (*f*) interpretaríamos la relación entre *X* y *Y* como espuria, toda vez que *Z* actúa en el sentido de producir variación tanto en *X* como en *Y*. Esta situación es exactamente la misma que se describió en el ejemplo del consumo de helados *X* y las tasas de delincuencia *Y*. Debido a que sospechamos que la relación entre estas dos variables se deba a otra, o sea al nivel de ingreso *Z*, controlamos en relación con ésta para ver si la correlación entre *X* y *Y* se reduce a casi cero. Si (*f*) es de hecho el modelo correcto, acabamos de ver matemáticamente que tal será el caso.

Vimos también, sin embargo, que la parcial habría sido cero si el modelo correcto fuera el de la figura XIX.5*e*. En (*e*), en efecto, tenemos que *Z* actúa como variable interventora, en el sentido de que *X* causa *Z*, la cual a su vez causa *Y*. Pero, ¿tiene algún objeto controlar en relación con *Z* en estas condiciones? Probablemente no. Porque si *X* es efectivamente causa de *Z*, ¿cómo podemos concebir que mantengamos a *Z* constante mientras *X* sigue variando? No tiene sentido ciertamente pensar obtener residuales tomando aquella porción de la variación de *X* que es "debida a" *Z* cuando *Z* es un efecto de *X*. Puede, sin embargo, tener sentido el controlar para *Z* si lo que tratamos de demostrar es la ausencia de una conexión causal entre *X* y *Y*, excepto a través de la variable interventora *Z*. La manipulación de fórmulas estadísticas no constituye sustituto alguno del conocimiento de lo que se está haciendo. En este caso, saber lo que se está haciendo consiste en estar en condiciones de elegir entre los modelos (*e*) y (*f*), yendo más allá de la información estadística disponible y haciendo un supuesto acerca de la dirección de la flecha entre *X* y *Z*.

Hasta aquí hemos prescindido de la situación (*d*) de la figura XIX.5, en la que las flechas van a *Y* tanto de *X* como de *Z*, pero en la que no se da relación directa alguna entre *X* y *Z*. ¿Qué sucede en este caso si controlamos en relación con *Z*? Observamos en primer lugar que tiene objeto controlar aquí en relación con *Z* porque ésta se concibe como variable totalmente independiente que afecta también a *Y*. Desde el punto de vista de la relación entre *X* y *Y*, opera como influencia perturbadora. Es una variable "extraña" que produce esencialmente en *Y* variaciones fortuitas con respecto a las variaciones de *X*. Por lo tanto, esperaríamos que, controlando en relación con *Z*, aumentara la magnitud de la relación entre *X* y *Y*. Puede demostrarse matemáticamente que si establecemos los supuestos requeridos a propósito de otras variables, la correlación en la población entre *X* y *Z* será cero. Señalemos de paso que este hecho nos permitirá distinguir (*d*) empíricamente de cada una de las situaciones que hemos venido examinando. Ésta es, pues, la situación en la que la variable de control no se relaciona con una de las otras variables, y ya vimos que en tal caso la parcial será mayor en valor absoluto que la correlación total, lo que concuerda con el sentido común. Es asimismo la situación a la que nos enfrentamos en el análisis por dos métodos de la variancia, en la que la condición de subceldillas iguales suponía una independencia completa entre las variables de fila y de columna, y en la que también vimos que un control para una de las variables reducía la suma inexplicada de los cuadrados, sin reducir la variación explicada por la otra variable independiente.

Hay otro tipo de situación de control que no se ha examinado, pero que puede tratarse brevemente, ya que son pocos los casos, si los hay, en que podríamos vernos inducidos a servirnos de un control.

Supóngase, en efecto, en una de las situaciones (*e*) o (*h*), que iba a relacionar las variables dependientes que intervienen, controlando en relación con la variable independiente. En (*h*), por ejemplo, ¿qué sucedería si fuéramos a obtener la parcial entre *X* y *Y* controlando en relación con *Z*? Puede demostrarse algebraicamente que la parcial resultante sería menor en magnitud que la correlación total. Esto concuerda con la noción intuitiva de que, manteniendo constante la variable independiente, se reduce necesariamente la variación de la variable interferente, con lo que se debilita la relación con la variable dependiente. Una vez más, tendría poco objeto llevar a cabo semejante operación. Por lo regular, en efecto, nuestro interés se centrará en la cuestión de saber si existe o no una relación directa entre *X* y *Y*, y no en el problema de las causas antecedentes de *X*. Puede demostrarse, sin embargo, que si hubiéramos controlado inadvertidamente para *Z* en (*h*), *no* hubiéramos afectado sistemáticamente

el *declive* estimado b_{xy}, excepto en el sentido de que habríamos aumentado el valor del error de muestreo.

Las extensiones a cuatro o más variables son directas, con tal de que nos restrinjamos a una causación en sólo un sentido. Puede demostrarse que en los casos en que no hay lazo directo entre dos variables, se dará una parcial de orden más elevado entre estas variables, la que será aproximadamente igual a cero, excepto por los errores de muestreo. En general, debemos controlar para todas las variables antecedentes e interventoras, con objeto de hacer desaparecer la apropiada correlación parcial, pero habremos de tener cuidado, evitando controlar para variables que se supone son dependientes de las dos que están siendo consideradas. Por ejemplo, en el modelo

$$\begin{array}{ccc} X_1 & \rightarrow & X_2 \\ \downarrow & & \downarrow \\ X_3 & \rightarrow & X_4 \end{array}$$

será necesario controlar *tanto* para X_2 *como* para X_3, con el fin de reducir a cero la parcial $r_{14 \cdot 23}$. De manera análoga, el modelo predice para $r_{23 \cdot 1} = 0$ (excepto por errores de muestreo), pero *no* deberemos esperar que $r_{23 \cdot 14}$ sea igual a cero, ni tendría sentido alguno controlar en este caso para X_4. (Véase [2] para más amplia discusión.)

Son de nuevo necesarias varias advertencias. Como en el caso de las tres variables, habrá siempre modelos alternativos que predigan exactamente las mismas intercorrelaciones empíricas, y habrá que confiar en el conocimiento de las secuencias temporales, o supuestos *a priori*, cuando haya que escoger entre tales alternativas. Por otra parte, la existencia de errores de medición aleatorios y no aleatorios invalidará las predicciones de cualquier modelo dado. Como observamos en el capítulo anterior, el error aleatorio de medición en una variable independiente atenuará las correlaciones entre ésta y otras variables. En el caso de regresión *múltiple*, y cuando las variables independientes estén altamente intercorrelacionadas, los errores aleatorios de medición, en algunas de ellas, tenderá a *aumentar* los efectos visibles de aquellas variables con las que estén más altamente intercorrelacionadas. Se ve de esta manera que los errores de medición en presencia de una alta intercorrelación entre variables independientes se prestan a conducirnos a deducciones erróneas.

Resultará claro de las observaciones anteriores que si uno suma variables a una ecuación de regresión podrá esperar que las correlaciones parciales cambien según sea la naturaleza de las intercorrelaciones entre las variables independientes. Esto es aplicable a los declives parciales y estandarizados que se examinan en la sección siguiente. Suponemos que el error, o término residual para la ecuación de regresión, no está relacionado con cada una de las variables independientes de la propia ecuación. En términos causales, esto hace suponer que los factores que son causa mayor de la variable dependiente no están sistemáticamente relacionados con las variables independientes.

Si somos capaces de localizar las variables que contribuyen a este factor de error y si las hacemos figurar de manera explícita en la ecuación, tales variables deberán *no* estar relacionadas con las variables independientes originales, aparte los errores de muestreo, no resultando afectados sistemáticamente los declives parciales. Las correlaciones parciales, por otra parte, aumentarán en su valor numérico, debido a que habrá sido eliminado algo de la variancia no explicada. Sin embargo, si las variables adicionales llevadas a la ecuación están relacionadas sistemáticamente con las variables independientes originales, podrá darse por seguro que todos los coeficientes resultarán afectados.

XIX.4. *Mínimos cuadrados múltiples y los coeficientes beta*

Nos hemos servido de las correlaciones parciales para indicar el grado de relación entre una variable dependiente y una variable independiente, controlando en relación con una o varias variables independientes más. Si tenemos un número grande de variables independientes, podemos obtener una indicación de su importancia relativa asociando la variable dependiente con cada una de las variables independientes sucesivamente y controlando en cada caso con las variables restantes. Anteriormente, en nuestro examen de la regresión múltiple y de los mínimos cuadrados, ya observamos también que las b y las β que figuran en nuestras ecuaciones y relacionan a Y con las variables independientes podrían interpretarse en cierto sentido como parciales. Se recordará que representan las pendientes de las ecuaciones de regresión o de los mínimos cuadrados en la dimensión de la variable independiente apropiada, esto es, con todas las demás variables independientes mantenidas constantes. Por lo tanto, cada coeficiente representa la cantidad de variación de Y que puede asociarse con un cambio determinado de las X, manteniendo fijas las demás variables independientes. Teniendo en cuenta esta similitud con los coeficientes de la correlación parcial, no debería sorprender que las fórmulas empleadas en el cálculo de esas b parciales resultaran muy semejantes a las que se emplearon en obtener las r parciales y que, además, esas pendientes pudieran emplearse para dar una indicación de la importancia relativa de cada una de las variables independientes en la determinación de la variación de Y.

Hemos de modificar nuevamente nuestra notación, con objeto de distinguir entre el gran número de combinaciones posibles de

las pendientes. Designando nuestras variables simplemente como 1, 2, 3, etcétera, nos servimos del símbolo $b_{12\cdot3}$ si anticipamos la variable uno a partir de las variables 2 y 3 con referencia al coeficiente de la segunda variable. El coeficiente $b_{12\cdot3}$ ha de distinguirse de $b_{21\cdot3}$, que emplearíamos si la segunda variable se tomara como variable dependiente. En ambos casos, el hecho de que el número tres se haya colocado a la derecha del punto indica que se ha controlado la tercera variable. Y en forma análoga, $b_{12\cdot34}$ se emplea para indicar el coeficiente de la segunda variable en una ecuación de predicción en la que la primera variable se toma como variable dependiente y que comporta dos variables de control. En este último caso, la ecuación de los mínimos cuadrados se daría en la siguiente forma:

$$X_1 = a_{1\cdot234} + b_{12\cdot34}X_2 + b_{13\cdot24}X_3 + b_{14\cdot23}X_4$$

en donde el subíndice de a indica que estamos anticipando en relación con la variable uno a partir de las variables 2, 3 y 4. La razón de que hayamos considerado conveniente apartarnos de la práctica consistente en designar la variable dependiente con Y está en servirnos de una combinación más sencilla de subíndices para seguir la traza de las distintas b.

Las fórmulas de cálculo de $a_{i\cdot jk}$ y $b_{ij\cdot k}$ son como sigue:

$$a_{i\cdot jk} = \overline{X}_i - b_{ij\cdot k}\,\overline{X}_j - b_{ik\cdot j}\,\overline{X}_k \qquad \text{(XIX.7)}$$

y

$$b_{ij\cdot k} = \frac{b_{ij} - (b_{ik})(b_{kj})}{1 - b_{jk}b_{kj}} \qquad \text{(XIX.8)}$$

Obsérvese que si bien el denominador de (XIX.8) difiere en cuanto a la forma del de la fórmula de $r_{ij\cdot k}$, el numerador, en cambio, es esencialmente similar en carácter.

En efecto, recordando que

$$r_{jk}^2 = b_{jk}b_{kj}$$

vemos que incluso los denominadores no son demasiado dispares en cuanto a la forma. Con todo, al emplear esta fórmula para obtener las b parciales, hay que poner cuidado en distinguir entre b_{jk} y b_{kj}, ya que los subíndices ya no pueden intercambiarse.

La extensión a parciales de orden superior es directa (véase [5]). Las ecuaciones de $a_{i\cdot jkl}$ y $b_{ij\cdot kl}$ pueden escribirse:

$$a_{i\cdot jkl} = \overline{X}_i - b_{ij\cdot kl}\,\overline{X}_j - b_{ik\cdot jl}\,\overline{X}_k - b_{il\cdot kj}\,\overline{X}_l \qquad \text{(XIX.9)}$$

y

$$b_{ij\cdot kl} = \frac{b_{ij\cdot k} - (b_{il\cdot k})(b_{lj\cdot k})}{1 - b_{jl\cdot k}(b_{lj\cdot k})} \qquad \text{(XIX.10)}$$

Igualmente cierto en el cálculo de correlaciones parciales de orden superior a medida que el número de variables aumenta, el empleo de estas fórmulas puede comportar acaso considerablemente más trabajo que el que requieren los métodos abreviados de Doolittle o de la raíz cuadrada de Dwyer. Normalmente será, por supuesto, más conveniente utilizar programas de computación, cuando se trate de obtener estos coeficientes.

Se puede interpretar un declive parcial como el cambio hipotético que ocurriría en la variable dependiente *si* una de las variables independientes hubiera de cambiar en una unidad y *si* las demás variables permanecieran constantes. Esto puede ser interpretado como una medida del efecto directo de la variable independiente sobre la variable dependiente; si un declive parcial es igual a cero, ello no implicaría un efecto directo. Pero no habiendo especificado las conexiones causales entre las propias variables independientes y teniendo en cuenta únicamente sus intercorrelaciones, no nos es posible decir nada en relación con el efecto total de cada variable. Si, por ejemplo, la primera variable independiente es una causa de la segunda, un cambio en la primera variable produciría un cambio también en la segunda, produciéndose efectos tanto directos como indirectos. De esta manera no podemos valorar la importancia relativa de cada variable, a menos que conozcamos más acerca de la estructura causal del sistema en su totalidad. Esto requeriría trabajar con todo un *grupo* de ecuaciones, una por cada variable que sea tomada como dependiente de cualesquiera de las otras. Por desgracia, los mínimos cuadrados ordinarios no son en general adecuados para tal sistema de ecuaciones (véanse [4] y [12]).

En tanto no estemos interesados en generalizar más allá de los límites de una sola población, en ocasiones es deseable obtener una medida asimétrica de los efectos directos de cada variable independiente, que no dependa de las unidades de medida utilizadas. Obtenemos así, en efecto, una medida del efecto directo real en el caso particular de la población que estudiamos, dado que algunas variables independientes varían más que otras. Una variable puede ser medida en dólares, otra en años. Carecería de sentido comparar la unidad de cambio en una con la unidad de cambio en la otra.

Si cada variable es estandarizada, dividiéndola por su desviación estándar, en la misma forma que se aplicó para obtener la curva normal estándar obtendremos declives ajustados, comparables de una variable a la siguiente. Medimos así los cambios en la variable dependiente en función de unidades de desviación estándar para cada una de las otras variables, lo que nos asegura

una misma variabilidad en cada una de estas variables. Estos declives parciales ajustados resultan así bes (b) estandarizadas, llamadas frecuentemente *ponderaciones beta,* siendo denominados *coeficientes de curso* en los más simples modelos causales, en los que está implicada una determinante de causa en una sola dirección (véase [14]).

Por desgracia, una vez más nos vemos envueltos en incongruencias de notación. En efecto, estas ponderaciones de beta no son las mismas que las de las β en la ecuación de regresión, que se refieren a características de la *población* y no han sido ajustadas en relación con las diferencias de variabilidad. Las ponderaciones de beta se obtienen de los datos de la *muestra* y son simples funciones de las b parciales. Las fórmulas generales de $\beta_{ij\cdot k}$ y $\beta_{ij\cdot kl}$ son:

$$\beta_{ij\cdot k} = b_{ij\cdot k}\frac{s_j}{s_i} \qquad \text{(XIX.11)}$$

y

$$\beta_{ij\cdot kl} = b_{ij\cdot kl}\frac{s_j}{s_i} \qquad \text{(XIX.12)}$$

Así, pues, la ponderación de beta puede obtenerse multiplicando la b comparable por la razón de la desviación estándar de la variable independiente (no controlada) a la de la variable dependiente.

La comparabilidad de las ponderaciones de beta y los coeficientes de correlación parcial puede verse en la fórmula:

$$\beta_{ij\cdot k} = \frac{r_{ij} - r_{ik}r_{jk}}{1 - r_{jk}^2} \qquad \text{(XIX.13)}$$

Las dos medidas sólo difieren en los denominadores. De hecho, vemos inmediatamente que:

$$r_{ij\cdot k}^2 = (\beta_{ij\cdot k})(\beta_{ji\cdot k})$$

ya que $\beta_{ji\cdot k}$ sólo difiere de $\beta_{ij\cdot k}$ en que el denominador de r_{jk}^2 será remplazado por r_{ik}^2. Ya que las ponderaciones de beta y las correlaciones parciales representan tipos de medida de asociación algo distintos, no darán exactamente los mismos resultados, aunque por lo regular comprendan variables del mismo orden de importancia. En efecto, la correlación parcial es una medida de la *cantidad de variación explicada* por una de las variables independientes después que las otras han explicado todo lo que podían. Las ponderaciones de beta, en cambio, indican *cuánto cambio* se produce en la variable dependiente por un cambio estandarizado en una de las variables independientes al controlar en relación con las otras.

XIX.5. *Correlación múltiple*

Como quiera que nuestro interés puede acaso residir en el poder explicativo de cierto número de variables independientes tomadas juntas más que en la relación entre la variable dependiente y cada una de las variables independientes tomadas separadamente, preferiremos tal vez servirnos del *coeficiente de correlación múltiple,* que es una medida de la bondad de ajuste de la superficie de mínimos cuadrados a los datos. Al igual que el cuadrado del coeficiente de la correlación de orden cero indicaba el porcentaje de variación explicada por la recta de mejor ajuste, el cuadrado del coeficiente de correlación múltiple puede emplearse para dar el porcentaje de variación explicado por la ecuación de mejor ajuste de la forma:

$$Y_p = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k$$

Otra manera de concebir la correlación múltiple está en que representa la *correlación de orden cero* entre los *valores reales* obtenidos para la variable dependiente y los *valores anticipados a partir de la ecuación de mínimos cuadrados.* Si todos los puntos se encuentran exactamente en la superficie de mínimos cuadrados, los valores real y anticipado coincidirán, y la correlación múltiple será la unidad. Y cuanto mayor sea la dispersión alrededor de la ecuación de mínimos cuadrados tanto más baja será la correlación entre los valores real y predicho.

La fórmula de la correlación múltiple puede desarrollarse fácilmente sirviéndose del hecho de que el cuadrado del múltiple será igual al porcentaje de la variación explicada por todas las variables independientes. Conviene recalcar una vez más que se supone un modelo de tipo lineal. Al escribir la fórmula de la correlación múltiple, dejamos primero que una de las variables independientes explique todo lo que puede. Dejamos luego que la segunda variable independiente haga lo propio en relación con la porción de la variación no explicada por la primera. Sin embargo, con objeto de evitar duplicación, hemos de controlar en relación con esta primera variable independiente. Dejamos a continuación que la tercera explique todo lo que puede del resto, controlando ahora en relación con las dos primeras variables independientes. El procedimiento puede prolongarse de manera indefinida.

De momento, sin embargo, nos limitamos al caso de tres variables, que sólo comporta dos variables independientes. Si tomamos la primera variable como la variable dependiente, y designamos el coeficiente de correlación múltiple por $R_{1\cdot 23}$, podremos escribir:

$$R^2_{1\cdot 23} = r^2_{12} + r^2_{13\cdot 2}(1 - r^2_{12})$$

$$\begin{pmatrix}\text{Proporción}\\ \text{explicada}\\ \text{por 2 y 3}\end{pmatrix} = \begin{pmatrix}\text{proporción}\\ \text{explicada}\\ \text{por 2}\end{pmatrix} + \begin{pmatrix}\text{proporción}\\ \text{adicional}\\ \text{explicada}\\ \text{por 3}\end{pmatrix}\begin{pmatrix}\text{proporción}\\ \text{no explica-}\\ \text{da por 2}\end{pmatrix} \qquad \text{(XIX.14)}$$

Obsérvese que las correlaciones múltiples sólo tienen una cifra a la izquierda del punto, cifra que indica la variable dependiente. Los números de la derecha, en cambio, indican aquellas variables independientes que se están empleando para explicar la variación de la variable dependiente. Así, pues, la fórmula general (para tres variables) puede escribirse como sigue:

$$\begin{aligned} R^2_{i\cdot jk} &= r^2_{ij} + r^2_{ik\cdot j}(1 - r^2_{ij}) \\ &= r^2_{ik} + r^2_{ij\cdot k}(1 - r^2_{ik}) \end{aligned} \qquad \text{(XIX.15)}$$

No importa, por supuesto, cuál de las dos variables independientes se emplee primero en la fórmula, a condición que dicha variable se controle en los términos siguientes de la ecuación.

Operamos con los cuadrados tanto de la correlación total como de las correlaciones parciales, ya que obtenemos los porcentajes de la variación explicada. Por lo tanto, no tenemos por qué preocuparnos por los signos de estas correlaciones. Y de hecho, la dirección de la múltiple carece de significado, ya que comporta correlaciones con cierto número de variables, algunas de las cuales son positivas y otras posiblemente negativas. Por convención, al designar el coeficiente de correlación múltiple, tomamos siempre la raíz cuadrada positiva de R^2.

Si resolvemos la ecuación (XIX.14) en relación con la parcial $r^2_{13\cdot 2}$, obtenemos:

$$r^2_{13\cdot 2} = \frac{R^2_{1\cdot 23} - r^2_{12}}{1 - r^2_{12}} \qquad \text{(XIX.16)}$$

Esto nos permite ver la relación entre los coeficientes de las correlaciones múltiples y parciales bajo una perspectiva algo distinta. En el numerador hemos sustraído la proporción de la variación de 1 explicada por la 2 sola, de la proporción explicada por 2 y 3 actuando juntas ($R^2_{1\cdot 23}$). El resultado es el incremento explicado por 3, después de haber permitido actuar a 2. Si dicho incremento se divide entre la proporción de variación dejada sin explicar por 2, obtenemos la parcial entre 1 y 3 controlando en relación con 2. Esto concuerda con nuestra interpretación anterior del coeficiente de la correlación parcial.

De la ecuación (XIX.14) pueden derivarse diversas fórmulas alternativas pero equivalentes de $R^2_{1\cdot 23}$. Sustrayendo ambos miembros de dicha ecuación de la unidad, obtenemos:

$$\begin{aligned} 1 - R^2_{1\cdot 23} &= 1 - r^2_{12} - r^2_{13\cdot 2}(1 - r^2_{12}) \\ &= (1 - r^2_{12})(1 - r^2_{13\cdot 2}) \end{aligned} \qquad \text{(XIX.17)}$$

Esta ecuación indica que podemos escribir la proporción de variación dejada sin explicar por 2 y 3 juntas, como producto de la proporción inexplicada por 2 y de aquella inexplicada por 3, controlando en relación con 2.

La fórmula de la múltiple puede escribirse también totalmente en términos de correlaciones de orden cero. En efecto, sirviéndonos de la ecuación (XIX.3) de $r_{13\cdot 2}$ en términos de coeficientes de orden cero y simplificando la expresión algebraica resultante, obtenemos:

$$R^2_{1\cdot 23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23}}{1 - r^2_{23}}$$

o bien, en general:

$$R^2_{i\cdot jk} = \frac{r^2_{ij} + r^2_{ik} - 2r_{ij}r_{ik}r_{jk}}{1 - r^2_{jk}} \qquad \text{(XIX.18)}$$

En particular, si la correlación entre las dos variables independientes j y k acontece ser cero, obtenemos:

$$R^2_{i\cdot jk} = r^2_{ij} + r^2_{ik}$$

Pueden observarse ahora algunas relaciones entre la múltiple R y las diversas correlaciones totales. Es obvio que R no puede ser menor en magnitud que cualesquiera de las correlaciones totales, ya que es imposible explicar *menos* variación añadiendo más variables. Normalmente, por supuesto, la múltiple R será mayor que una cualquiera de las r totales. Su valor máximo en relación con los coeficientes totales suele producirse cuando las intercorrelaciones entre las variables independientes son todas cero. Como acabamos de ver, el cuadrado de la correlación múltiple será en este caso igual a la suma de los cuadrados de las demás correlaciones. Por otra parte, si las intercorrelaciones entre las variables independientes son muy grandes en magnitud,

la múltiple R no será por lo regular mucho mayor que la correlación total mayor con la variable dependiente. En otros términos: si deseamos explicar lo más posible de la variación de la variable dependiente, hemos de buscar variables independientes que tengan relativamente poca relación unas con otras, pero que tengan por lo menos correlaciones moderadamente altas con la variable dependiente. O expresado en otra forma: si tenemos dos variables independientes de interrelación alta, la segunda explicará esencialmente la misma variación que la primera, ya que las dos se traslaparán considerablemente. Y si no están correlacionadas, entonces cada una explicará una porción diferente de la variación total.

Hay otra razón importante para preferir las variables independientes que no estén altamente intercorrelacionadas. No sólo habrá menos superposiciones en la variancia explicada, y por ello menos ambigüedad en nuestra interpretación causal de sus supuestos efectos, sino que en la medida en que las variables independientes estén altamente intercorrelacionadas, tanto las correlaciones parciales como las estimaciones de declives se harán cada vez más sensibles a los errores de muestreo y medición. Esta dificultad se denomina *multicolinearidad* en la bibliografía econométrica (véanse [4] y [12]). El problema resulta especialmente serio cuando se utilizan bloques de variables independientes altamente intercorrelacionadas, y cuando dichos bloques difieren en cuanto al número de variables que contienen. (Véase [10]). Puede demostrarse, por ejemplo, que con muy pequeñas diferencias en las correlaciones totales con la variable dependiente se producen diferencias considerables en las correlaciones parciales y en la estimación de los declives, de tal manera que si se confía en las magnitudes relativas de estos coeficientes parciales, cabe esperar encontrar diferencias considerables de una muestra a la siguiente, o bien entre réplicas en las que se utilicen instrumentos de medición algo distintos. La conclusión es que en cuantas ocasiones las variables independientes estén altamente intercorrelacionadas, resultará necesario contar *tanto* con muestras grandes *como* con las mediciones exactas.

A título de ejemplo numérico del cálculo de la múltiple R, veamos cuánta variación en materia de discriminación puede explicarse por el porcentaje de negros y el porcentaje urbano. Sirviéndonos de la ecuación (XIX.14) obtenemos:

$$R^2_{1\cdot23} = r^2_{12} + r^2_{13\cdot2}(1 - r^2_{12}) = (.536)^2 + (.332)^2[1 - (.536)^2]$$
$$= .2873 + (.1102)(.7127) = .3658$$

De ahí: $$R_{1\cdot23} = .605$$

Por consiguiente, el porcentaje urbano explica muy poca variación por encima y por debajo de aquella explicada por el porcentaje de negros.

A título de control de nuestros cálculos, observamos que el mismo resultado deberá obtenerse si dejamos que actúe primero el porcentaje urbano.

Obtenemos en este caso:

$$r_{12\cdot3} = \frac{r_{12} - r_{13}(r_{23})}{\sqrt{1 - r^2_{13}}\sqrt{1 - r^2_{23}}} = \frac{.536 - (.139)(-.248)}{\sqrt{1 - (.139)^2}\sqrt{1 - (-.248)^2}} = .595$$

Así pues, $$R^2_{1\cdot23} = r^2_{13} + r^2_{12\cdot3}(1 - r_{13})$$
$$= (.139)^2 + (.595)^2[1 - (.139)^2] = .3667$$

y por lo tanto: $$R_{1\cdot23} = .605$$

Las fórmulas del coeficiente de correlación múltiple pueden extenderse fácilmente asimismo a un número cualquiera de variables independientes. Al introducir una tercera variable independiente, designada como X_4, no hacemos más que añadir a la fórmula de $R^2_{1\cdot23}$ un término que comporta el producto del cuadrado de la parcial entre 1 y 4, controlando en relación con 2 y 3, y la proporción de variación queda inexplicada por 2 y 3. Así, pues:

$$R^2_{1\cdot234} = r^2_{12} + r^2_{13\cdot2}(1 - r^2_{12}) + r^2_{14\cdot23}[1 - r^2_{12} - r^2_{13\cdot2}(1 - r^2_{12})]$$
$$= R^2_{1\cdot23} + r^2_{14\cdot23}(1 - R^2_{1\cdot23}) \qquad \text{(XIX.19)}$$

Podemos, pues, ir añadiendo a la proporción de la variación explicada, siempre que controlemos en relación con todas las variables previamente empleadas y a condición que permitamos que la nueva parcial sólo actúe sobre aquella porción de variación dejada inexplicada por las variables que la han precedido. Obsérvese, de paso, el paralelo con lo que hicimos en el análisis de la variancia. Según veremos a continuación, podemos servirnos de este hecho en las pruebas de significación tanto de la correlación múltiple como de la parcial. Si procediéramos a añadir una quinta variable, obtendríamos:

$$R^2_{1\cdot2345} = R^2_{1\cdot234} + r^2_{15\cdot234}(1 - R^2_{1\cdot234})$$

Podemos resolver de nuevo estas ecuaciones en relación con los

coeficientes parciales. Así, por ejemplo, tenemos (de XIX.19):

$$r^2_{14\cdot 23} = \frac{R^2_{1\cdot 234} - R^2_{1\cdot 23}}{1 - R^2_{1\cdot 23}} \qquad \text{(XIX.20)}$$

indicando que la parcial entre 1 y 4, controlando en relación con 2 y 3, puede interpretarse como la razón de la proporción de variación adicional explicada por 4, además de la explicada por 2 y 3, a la proporción de variación dejada sin explicar por estas dos últimas variables. Podemos también extender la ecuación (XIX.17) para comprender más variables. Así, por ejemplo:

$$1 - R^2_{1\cdot 234} = (1 - r^2_{12})(1 - r^2_{13\cdot 2})(1 - r^2_{14\cdot 23})$$

y, en general,

$$1 - R^2_{1.234\ldots k} = (1 - r^2_{12})(1 - r^2_{13.2}) \ldots (1 - r^2_{1k.234\ldots(k-1)}) \qquad \text{(XIX.21)}$$

El coeficiente parcial-múltiple. En ocasiones resulta deseable calcular una correlación múltiple entre una variable dependiente y algunas variables independientes, controlando en relación con una o varias de éstas. Supóngase, por ejemplo, que se está tratando de predecir el tamaño real de la familia a partir de cierto número de variables independientes. Es obvio que ciertas variables como la duración del matrimonio y la edad de la esposa en el momento de celebrarlo han de tomarse en consideración. Por otra parte, estas variables son tan obvias, que el hecho de conjuntarlas al coeficiente general múltiple podría oscurecer los efectos de las variables restantes. Así, pues, el interés podría fijarse en la variación del tamaño de la familia después que dichas variables teóricamente poco importantes han explicado de la variación todo lo que podían. Siguiendo a Croxton y Cowden [6], indicamos la parcial-múltiple entre la variable 1 (dependiente) y 2 y 3, controlando en relación con 4, por medio de $r_{1(23)\cdot 4}$. La fórmula se convierte en tal caso en:

$$r^2_{1(23)\cdot 4} = \frac{R^2_{1\cdot 234} - r^2_{14}}{1 - r^2_{14}}$$

La fórmula anterior de la parcial-múltiple es una simple extensión de las fórmulas que hemos utilizado en las correlaciones múltiple y parcial. Dejamos primero que la variable de control 4 explique todo lo que puede. Observamos luego que $R^2_{1\cdot 234}$ representa la proporción de variación explicada por las tres variables independientes tomadas juntas. La diferencia, pues, ha de deberse a las variables 2 y 3. De este modo, el numerador representa la proporción de variación explicada por 2 y 3, además de aquella explicada por 4. Pero, como quiera que sólo hemos de operar con la variación no explicada por la variable de control, dividimos entre la cantidad $1 - r^2_{14}$. Sirviéndonos del principio consistente en dejar actuar primero todas las variables de control, podemos escribir la fórmula general de la parcial-múltiple como:

$$r^2_{i(jk\ldots n)\cdot tu\ldots w} = \frac{R^2_{i\cdot jk\ldots w} - R^2_{i\cdot tu\ldots w}}{1 - R^2_{i\cdot tu\ldots w}} \qquad \text{(XIX.22)}$$

Por ejemplo:

$$r^2_{3(56)\cdot 124} = \frac{R^2_{3\cdot 12456} - R^2_{3\cdot 124}}{1 - R^2_{3.124}}$$

La parcial-múltiple no parece haberse utilizado con mucha frecuencia en la investigación sociológica, debido tal vez al hecho de que las personas del ramo están poco familiarizadas con ella. Sin embargo, como medida que permite tratar problemas de correlación múltiple y parcial simultáneamente, su empleo potencial parece ser grande. En la próxima sección examinaremos otro tipo de empleo de esta medida.

XIX.6. *Regresión múltiple y no linealidad*

Hasta aquí toda nuestra labor se ha basado en el supuesto de modelos lineales. En el capítulo anterior vimos una prueba de no linealidad, pero sólo pudimos decir muy poco a propósito de la forma de la relación no lineal, excepto en el caso de transformaciones logarítmicas. En otros términos: no hicimos más que verificar en relación con la existencia de una *desviación* respecto de la linealidad, pero no efectuamos prueba alguna por lo que se refiere a la forma de la curva. Si bien el problema general de la no linealidad rebasa el objetivo de este texto, podemos, con todo, examinar brevemente de qué modo las técnicas de la regresión múltiple y de los mínimos cuadrados se dejan modificar ligeramente para permitirnos tratar algunos tipos de problemas que comportan no linealidad.

Como ya se señaló en el capítulo anterior, el número de formas que la relación no lineal puede adoptar es sumamente grande. Consideremos ecuaciones del tipo:

$$Y = a + b_1X + b_2X^2 + b_3X^3 + \ldots + b_kX^k \qquad \text{(XIX.23)}$$

Si todos los coeficientes b_2, b_3, ..., b_k son cero, la ecuación se reduce a la forma lineal familiar. En otros términos: la recta puede considerarse como caso particular de este tipo general de curva, que se designa como *polinomial*. Y en forma análoga, si todos los coeficientes, excepto a, b_1 y b_2, son cero, obtenemos una

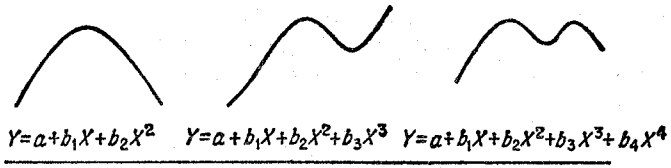

FIG. XIX.6. *Formas de polinomios de segundo, tercero y cuarto grados.*

polinomial de segundo grado. El grado de la polinomial se refiere al exponente más elevado de X que tenga un coeficiente no cero.

Las polinomiales tienen una propiedad muy importante, la que nos permite decir cuál es el grado de la ecuación que puede resultar más apropiada para nuestros datos. Obsérvese que una polinomial de primer grado es una línea recta sin desviaciones. Sucede que una ecuación de segundo grado contará con una desviación, describiendo de hecho la curva que llamamos parábola. Una polinomial de tercer grado tendrá dos desviaciones; la de cuarto grado, tres, y así sucesivamente. Si ignoramos ciertos casos degenerados en los que las "desviaciones" no se comportan adecuadamente, podremos dibujar las ecuaciones de segundo, tercero y cuarto grados, como se ve en la figura XIX.6. La dirección en que la parábola o curva de más alto grado "se abre", dependerá del signo de los coeficientes. Lo importante es observar que siempre habrá una desviación menos que lo que indica el grado de la ecuación.

Algunas veces obtenemos curvas empíricas que se parecen a una u otra de esas polinomiales, aunque raras veces, si es que alguna, necesitamos ir más allá de una ecuación de tercer grado. La parábola simple proporciona a menudo una adaptación razonablemente buena a los datos, sobre todo si nos damos cuenta de que nuestra curva puede ser perfectamente plana y que nuestros datos no necesitan extenderse lo suficiente para completar la flexión. Así, por ejemplo, los datos podrían ser similares a los que se indican en la figura XIX.7. Aquí, aunque no exista acaso razón teórica alguna para esperar que las marcas vuelvan a bajar una vez que hayamos avanzado cierta distancia a lo largo del eje de las X, la parábola puede constituir con todo una adaptación razonable, *dentro de los límites de variación dados en el problema*. Es, pues, perfectamente posible que una parábola de

mínimos cuadrados se adapte a los datos mucho mejor que la recta.

Supóngase que sea efectivamente así. ¿Cómo puede tratarse el problema? El lector se habrá ya dado cuenta, sin duda, de la semejanza entre la fórmula de la polinomial general y la de la ecuación de los mínimos cuadrados de más de una variable independiente. La única diferencia, en efecto, está en que hemos escrito X^2 en lugar de X_2, etcétera. Supóngase ahora que hubiéramos de representarnos X^2 como variable separada y distinta de X. Mientras nos servimos de símbolos abstractos esto es perfectamente posible, aunque, admitiéndolo, esta práctica no tendría mucho sentido en términos de una variable concreta. Con todo, las matemáticas del caso resultan ser idénticas. Así, por ejemplo, si sospechamos que la relación entre la discriminación y el porcentaje de negros pueda representarse acaso más adecuadamente por medio de una curva de segundo grado, tomamos el porcentaje de negros como una de las variables independientes X_1 y (el porcentaje de negros)^2 como segunda variable independiente X_2. Por consiguiente, la ecuación de segundo grado:

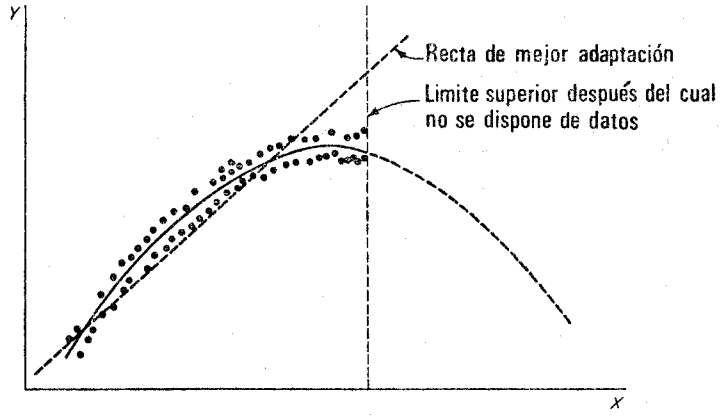

FIG. XIX.7. *Datos hipotéticos con una parábola de mejor ajuste.*

$$Y = a + b_1X + b_2X^2$$

difícil de tratar por medio de los mínimos cuadrados, se reduce a la ecuación familiar:

$$Y = a + b_1X_1 + b_2X_2$$

Para obtener una medida de bondad de ajuste a la parábola, podemos servirnos ahora de la correlación múltiple entre Y y X_1

y X_2. La diferencia entre el cuadrado de esta correlación múltiple y el cuadrado de la r total (suponiendo linealidad) nos dará una medida del grado en que hemos aumentado nuestra habilidad en cuanto a predecir la discriminación sirviéndonos, sin embargo, de una ecuación de segundo grado en lugar de una recta.

En principio, el procedimiento indicado puede extenderse de varios modos. Podrían emplearse ecuaciones de tercer grado y mayores con objeto de conseguir una adaptación algo mejor. Además, pueden añadirse al cuadro otras variables. Respecto de algunas de estas variables independientes, puede suponerse un modelo no lineal, y uno lineal respecto de otras. Al anticipar tasas de discriminación a partir de cierto número de variables independientes, podremos acaso encontrar que cabría obtener ecuaciones de predicción algo mejores suponiendo modelos no lineales para algunas de las variables. En particular, tal vez la relación entre la discriminación y el porcentaje de negros pueda ser de forma parabólica, en tanto que las variables independientes restantes se relacionan con la discriminación en forma lineal. Por lo tanto, la ecuación múltiple de los mínimos cuadrados adoptará la siguiente forma:

$$Y = a + (b_1X_1 + b_2X_2) + b_3X_3 + \ldots + b_kX_k$$

en la que los dos términos al interior del paréntesis se necesitan para describir la relación (no lineal) entre la discriminación y el porcentaje de negros. En este caso también, la variable X_2 vuelve a representar el porcentaje de negros al cuadrado: (porcentaje de negros)^2. Se concibe que también alguna de las otras X de la ecuación pueda emplearse asimismo para indicar relaciones no lineales entre la discriminación y las demás variables.

En el ejemplo anterior, supóngase que deseábamos obtener la correlación parcial entre la discriminación y el porcentaje de negros controlando en relación con las variables restantes. Como quiera que X_1 y X_2 se han utilizado para referirse a la primera y la segunda potencias del porcentaje de negros, no tendría sentido referir Y a X_1 controlando en relación con todas las demás "variables", incluida X_2. Antes bien, necesitamos obtener la correlación múltiple entre Y y tanto X_1 como X_2, controlando en relación con X_3, X_4, ..., X_k. Para lograr dicho propósito, podemos servirnos del coeficiente parcial-múltiple.

Manejo de la interacción como productos cruzados. En el análisis por dos métodos de la variancia, en el de la covariancia (véase capítulo XX), y en relación con las variables, dependientes de escala nominal, concebíamos la interacción estadística como si implicara *cualquier* diferencia de la simple adición. Una alternativa obvia para un modelo aditivo la constituye una relación multiplicativa del tipo que podría ser sugerido mediante argumentos verbales orientados a señalar que, al objeto de tener "presente" la Y, deberán tenerse "presentes" *tanto* la X_1 *como* la X_2. Cuando se avanza más allá de las simples dicotomías, la idea, generalizada, nos dice que Y puede ser una función multiplicativa de X_1 y X_2. La ecuación que sigue puede constituir una formulación general de tal relación.

$$Y = (\alpha_1 + \beta_1X_1)^{\gamma_1}(\alpha_2 + \beta_2X_2)^{\gamma_2}$$

en las que los exponentes de gamma pueden ser, o bien positivos, en cuyo caso estará implícita la multiplicación, o negativos, con división implicada. La función puede desde luego ser convertida en aditiva, haciendo una transformación logarítmica de todas las variables, pudiendo extender fácilmente el principio general a más de dos variables independientes.

Supongamos, como aproximación razonable, que ambos exponentes fuesen la unidad, lo que reduciría la ecuación a:

$$Y = (\alpha_1 + \beta_1X_1)(\alpha_2 + \beta_2X_2) = \alpha_1\alpha_2 + \alpha_2\beta_1X_1 + \alpha_1\beta_2X_2 + \beta_1\beta_2X_1X_2$$

Vemos inmediatamente que mediante la suma de un factor que abarca X_1X_2 podremos manejar este tipo de modelo simple multiplicativo, conservando el formato aditivo. Nos limitamos a denominar X_1X_2 como X_3, construyendo en consecuencia nuestra medida de X_3, y continuamos adelante. Deseamos, por ejemplo, medir el grado en que X_3 agrega a la variancia explicada, y podríamos probar la significancia de este factor adicional como se indica en la sección siguiente. Si hubiéramos comenzado con tres variables independientes, podríamos haber formado tres factores con los productos X_1X_2, X_1X_3 y X_2X_3 para determinar las tres interacciones de primer orden, y un triple producto $X_1X_2X_3$ para manejar la interacción de orden superior.

Es necesario formular varias advertencias. En primer lugar, el uso de factores de productos cruzados está justificado con base en que la relación "verdadera" sea multiplicativa y no aditiva, en tanto que la "no aditividad" se refiera a *cualquier* tipo de separación de la aditividad. Tenemos así una medida de interacción algo más restrictiva que la que se obtuvo en relación con el análisis de la variancia, y es posible que otros factores de interacción hubieran funcionado mejor (por ejemplo: $X_1 \log X_2$, $X_1 \cos X_2$, o $e^{X_1} \log X_2$). Segundo: si tomamos $X_3 = X_1X_2$, debemos tener presente que X_3 es una función no lineal exacta de X_1 y X_2, y por tanto las correlaciones momento-producto de X_3 tanto con X_1 como con X_2 serán de ordinario muy altas. Tendremos así entre manos un problema de multicolinearidad, y no podremos tener mucha fe en nuestras estimaciones de los coeficientes de

los factores X_i. Este problema resulta particularmente serio cuando se comienza con cinco o seis variables independientes y se desea tener en cuenta todas las posibles interacciones. Si las propias variables originales están altamente intercorrelacionadas, o bien forman parte de bloques, los factores de productos cruzados se relacionarán con tales bloques en formas peculiares (véase [1]). En tales casos puede resultar razonable medir hasta qué punto el grupo completo de factores de productos cruzados aumenta significativamente la variancia explicada, mediante el uso del coeficiente parcial-múltiple, o comparando los múltiples, con y sin los factores de los productos. La determinación de los efectos de determinados factores de los productos cruzados puede, sin embargo, resultar demasiado arriesgada, por razón de un gran volumen de errores de muestreo en los que pudiera haberse incurrido.

Hay evidentemente muchos más usos y más posibles extensiones de las técnicas de correlación y regresión múltiples, de los que pueden ser examinados en un texto general. Hemos visto, sin embargo, algunos de los principios básicos más elementales, los que permitirán consultar inteligentemente con los especialistas en caso de que se plantearan problemas más complicados.

XIX.7. *Pruebas de significación e intervalos de confianza*

En relación con la significación será necesario verificar, por supuesto, tanto el coeficiente múltiple como el parcial. La hipótesis nula y los supuestos serán similares a los que se establecieron en el caso de la correlación total. Una muestra aleatoria será supuesta como de costumbre. El supuesto de una distribución normal multivariable nos asegurará que cada variable está normalmente distribuida alrededor de las otras, que las variancias son iguales, y que la ecuación de regresión tendrá la forma indicada por la ecuación (XIX.1).[4] Hechos estos supuestos, podemos servirnos de las pruebas de análisis de variancia para la significación de varios coeficientes parciales y múltiples. Veremos primero pruebas de significancia de correlaciones múltiples, ya que éstas son más sencillas desde el punto de vista de los conceptos que las de las correlaciones parciales.

Como quiera que el cuadrado de la correlación múltiple representa siempre la proporción del total de la variación explicada por las variables independientes actuando juntas, hemos dividido

[4] Debe recalcarse una vez más que no todas las X_i necesitan tener distribuciones normales, en tanto la variable dependiente esté normalmente distribuida alrededor de todas las combinaciones de niveles fijos de las variables independientes con la misma variancia σ^2. Suponemos, con otras palabras, que el factor de perturbación ε_i se encuentra distribuido normalmente con la variancia constante.

de hecho esta variación total en dos porciones: las sumas explicada e inexplicada de cuadrados. Por lo tanto, el cuadro del análisis de variancia será siempre similar al cuadro XIX.1.

CUADRO XIX.1. *Prueba de análisis de variancia para la significación de la correlación múltiple*

	Sumas de cuadrados	*Grados de libertad*	*Apreciación de la variancia*	*F*	
Total	Σx_1^2	$N-1$			
Explicada	$R^2\Sigma x_1^2$	k	$\dfrac{R^2\Sigma x_1^2}{k}$	$\dfrac{R^2}{1-R^2}$	$\dfrac{N-k-1}{k}$
Inexplicada	$(1-R^2)\Sigma x_1^2$	$N-k-1$	$\dfrac{(1-R^2)\Sigma x_1^2}{N-k-1}$		

En el cuadro XIX.1 hemos indicado la variable dependiente con X_1, dejando que k represente el número de las variables independientes. Si R tiene, por ejemplo, una variable dependiente y tres variables independientes, habrá en la ecuación de regresión cuatro parámetros que hay que apreciar. Por consiguiente, sirviéndonos de la ecuación de los mínimos cuadrados para apreciar la variable dependiente, deberíamos perder 4 o $(k+1)$ grados de libertad. Así, pues, los grados de libertad asociados al término de error serán por lo regular

$$N-(k+1)=N-k-1$$

Los grados de libertad asociados a la suma de cuadrados explicada puede obtenerse a continuación por sustracción. Toda vez que los grados de libertad para las sumas de cuadrados explicada e inexplicada resultarán ser siempre k y $N-k-1$, respectivamente, podemos escribir una fórmula general de F. Obvérvese que, al igual que en el caso de las correlaciones totales, el factor que representa la suma total de cuadrados se elimina. Obtenemos así una fórmula general para verificar la significación de una R múltiple, o sea:

$$F_{k,N-k-1}=\frac{R^2}{1-R^2}\,\frac{N-k-1}{k} \qquad \text{(XIX.24)}$$

No es necesario, por consiguiente, establecer la tabla del análisis de variancia en la forma convencional. Verificando la significación de la correlación múltiple que obtuvimos al explicar la discriminación a partir del porcentaje de negros y el porciento urbano (p. 476), obtenemos ahora:

$$F_{2,147} = \frac{.3658}{1-.3658}\,\frac{150-3}{2} = \frac{.3658}{.6342}\,\frac{147}{2} = 42.39$$

que es significativa al nivel de .001.

Al verificar la significación de coeficientes parciales, operamos sobre la base del principio de dejar que las variables de control expliquen primero todo lo que pueden. Tomamos a continuación la porción de la suma total de cuadrados que queda *inexplicada* por la variable de control, y nos servimos de ella como nuevo total. Esta última cantidad se descompone luego en dos componentes, las porciones explicada e inexplicada, y una prueba F efectuada tomando la razón de las apreciaciones de la variancia basadas en estas dos últimas componentes. El procedimiento se ilustra en el cuadro XIX.2, en el que verificamos la significación de $r_{13\cdot 2}$ (o sea, $H_0: \varrho_{13\cdot 2} = 0$).

CUADRO XIX.2. *Prueba de análisis de variancia para la significación de la correlación parcial* $r_{13\cdot 2}$

	Sumas de cuadrados	*Grados de libertad*	*Estimación de la variancia*	F
Total	Σx_1^2	$N-1$		
Explicada por 2	$r_{12}^2\Sigma x_1^2$	1		
Inexplicada por 2	$(1-r_{12}^2)\Sigma x_1^2$	$N-2$		
Explicada por 3	$r_{13\cdot 2}^2(1-r_{12}^2)\Sigma x_1^2$	1	$r_{13\cdot 2}^2(1-r_{12}^2)\Sigma x_1^2$	
Inexplicada por 3	$(1-r_{13\cdot 2}^2)(1-r_{12}^2)\Sigma x_1^2$	$N-3$	$\dfrac{(1-r_{13\cdot 2}^2)(1-r_{12}^2)\Sigma x_1^2}{N-3}$	$\dfrac{r_{13\cdot 2}^2(N-3)}{1-r_{13\cdot 2}^2}$

Obsérvese que los grados de libertad inexplicados decrecen en uno cada vez que se añade una nueva variable. Por otra parte, en la fórmula de F la expresión se simplifica de tal modo, que resulta innecesario escribir la tabla entera cada vez que deseamos efectuar una prueba. En el problema numérico del que nos hemos venido sirviendo (p. 456) el valor de F de la prueba de significancia de la relación entre la discriminación y el porcentaje urbano, controlándolo en relación con el porcentaje de negros, se convierte en:

$$F_{1,N-3} = \frac{r^2_{13\cdot 2}}{1-r^2_{13\cdot 2}}(N-3) \qquad \text{(XIX.25)}$$

$$= \frac{(.332)^2}{1-(.332)^2}(147) = 18.21$$

Así pues, la parcial es significativa al nivel de .001.

Si al extender este procedimiento deseamos verificar la significación de $r_{14\cdot 23}$, podemos tomar como nuevo total la porción no explicada por 2 y 3 combinadas. Esta cantidad puede luego descomponerse en porciones explicada e inexplicada, practicándose la prueba de F lo mismo que anteriormente. Una vez más, todas las cantidades tanto del numerador como del denominador de F se eliminarán, excepto en cuanto a los factores que comportan las parciales. Toda vez que los grados de libertad asociados al numerador serán siempre la unidad y, como quiera que los del denominador serán $N-k-1$, podemos escribir la fórmula general de la verificación de la parcial $r_{ij\cdot mn\ldots t}$ como sigue:

$$F_{1,N-k-1} = \frac{r^2_{ij\cdot mn\ldots t}}{1-r^2_{ij\cdot mn\ldots t}}(N-k-1) \qquad \text{(XIX.26)}$$

en donde el número total de variables es $k+1$.

Obsérvese que al comparar las pruebas de la significación de las correlaciones múltiples y las parciales el término final de error que comporta la suma de cuadrados inexplicada por todas las variables deberá ser el mismo en ambas tablas, a condición, por supuesto, que se empleen las mismas variables dependientes e independientes. Ya demostramos que era así, toda vez que sabemos que:

$$1-R^2_{1\cdot 23} = (1-r^2_{12})(1-r^2_{13\cdot 2})$$

De los cuadros XIX.1 y XIX.2 puede verse que estas expresiones son las que figuran en las hileras inferiores de las tablas respectivas.

El procedimiento que acabamos de describir para verificar las correlaciones parciales puede utilizarse asimismo para verificar la significación de la parcial-múltiple. A estas alturas el lector estará ya en condiciones de verificar que, con objeto de hacer la prueba de significación de $r_{1(23)\cdot 45}$ tomaremos la suma de cuadrados no explicada por 4 y 5, sirviéndonos luego del cuadrado de la parcial-múltiple para obtener la proporción de esta nueva suma de cuadrados, que resulta explicada por las variables 2 y 3.

Pueden calcularse asimismo intervalos de confianza para los coeficientes parcial y múltiple, mediante una ligera modificación del procedimiento de transformación de la z descrito en el capítulo anterior. Podemos convertir de nuevo los dos tipos de coefi-

cientes en z sirviéndonos de la tabla. El único cambio que se requiere es que el error estándar de z ya no nos venga dado por

$$\sigma_z = \frac{1}{\sqrt{N-3}}$$

En lugar de ello, en cambio, perdemos un grado más de libertad por cada variable añadida, de modo que el error estándar se convierte en general en:

$$\sigma_z = \frac{1}{\sqrt{N-k-2}} \qquad \text{(XIX.27)}$$

en donde k representa el número total de variables.

Obtenemos por consiguiente los intervalos de confianza del 95 por ciento para $R_{1\cdot23}$ y $r_{13\cdot2}$ de la manera siguiente:

$$1.96\sigma_z = 1.96\frac{1}{\sqrt{146}} = .1622$$

	z	$z_l = z - 1.96\sigma_z$	$z_u = z + 1.96\sigma_z$	r_l	r_u
$R_{1\cdot23} = .605$	.7010	.5388	.8632	.492	.698
$r_{13\cdot2} = .332$	.3451	.1829	.5073	.181	.468

Así, pues, el intervalo de confianza del 95 por ciento alrededor de $R_{1\cdot23}$ va de .492 a .698, en tanto que el de $r_{13\cdot2}$ va de .181 a .468.

Antes de terminar el presente capítulo conviene observar un importante punto más. Cada vez que añadimos a la ecuación de los mínimos cuadrados otra variable, sólo perdemos un grado de libertad más. Podemos, por consiguiente, añadir variables, con una pérdida muy pequeña de eficacia, por lo que se refiere a las pruebas de significación. En ocasiones, la adición de más variables podrá bajar el nivel de significación, debido al hecho de que aquéllas no contribuyen a explicar bastante variación adicional para compensar la pérdida en grados de libertad. No obstante, tenemos en la correlación múltiple y parcial un instrumento que, si se aplica adecuadamente, es mucho más potente que cualquiera de los métodos que examinamos anteriormente. Sin embargo, si el número de variables utilizadas empieza a aproximarse al de los casos, podemos esperar obtener unas correlaciones múltiples muy grandes, debido simplemente a que estamos en condiciones de sacar partido de las fluctuaciones fortuitas. Con 15 casos y 15 variables, será posible pasar una superficie de mínimos cuadrados

exactamente entre todos los puntos, incluso si suponemos un modelo de tipo lineal. Por consiguiente, la múltiple R será automáticamente la unidad. De ahí que, lo mismo que las demás técnicas estadísticas, las de regresión y correlación múltiple deban emplearse con precaución. A estas alturas ya no será probablemente necesario señalar que, excepto con fines de exploración, no deberán emplearse, a menos que los supuestos requeridos se cumplan, si no totalmente, por lo menos aproximadamente.

Glosario

Ponderaciones de beta
Correlación múltiple
Correlación parcial-múltiple
Ecuación de regresión múltiple
Distribución normal multivariable
Correlación parcial
Ecuación polinomial

Ejercicios

1. Sirviéndose de los datos del ejercicio 1 del capítulo XVII.

   a) Obténgase la correlación parcial entre la integración moral y la heterogeneidad, controlando la movilidad. Calcúlese asimismo la parcial entre la integración moral y la movilidad, controlando la heterogeneidad. Respuesta, $-.51$; $-.63$.
   b) Obténgase la ecuación de mínimos cuadrados múltiple, tomando la integración moral como variable dependiente.
   c) ¿Qué son las ponderaciones beta? ¿Cómo se comparan con las parciales obtenidas en *a*)?
   d) Calcúlese la correlación múltiple, tomando la integración moral como variable dependiente. ¿Cómo pueden controlarse los cálculos? Respuesta, $R = .64$.
   e) Verifíquese la significación de las correlaciones múltiple y parcial calculadas en los apartados *a*) y *d*). Pónganse intervalos de confianza del 99 por ciento alrededor de cada una de estas correlaciones.

2. Escríbanse fórmulas para $r_{37\cdot12456}$, $R^2_{4\cdot1235}$ y $r^2_{5(23)\cdot1467}$. Respuesta, *b*) $R^2_{4\cdot1235} = R^2_{4\cdot123} + r^2_{45\cdot123}(1 - R^2_{4\cdot123})$.

3. Escríbanse las fórmulas de F que se emplearían para verificar el significado de cada una de las correlaciones del ejercicio 2 anterior. Respuesta, $(c)F = \dfrac{r^2_{5(23)\cdot1467}}{1 - r^2_{5(23)\cdot1467}} \dfrac{N-7}{2}$

Bibliografía

1. Althauser, R. P.: "Multicollinearity and Non-Additive Regression Models", en H. M. Blalock (ed.), *Causal Models in the Social Sciences*, Aldine Publishing Company, Chicago, 1971, cap. 26.

2. Blalock, H. M.: *Causal Inferences in Nonexperimental Research*, University of North Carolina Press, Chapel Hill, 1964, cap. 3.
3. Blalock, H. M.: "Per Cent Non-white and Discrimination in the South", *American Sociological Review*, vol. 22, pp. 677-682, 1957.
4. Christ, Carl: *Econometric Models and Methods*, John Wiley & Sons, Inc., Nueva York, 1966, Parte III.
5. Cowden, D. J.: "A Procedure for Computing Regression Coefficients", *Journal of the American Statistical Association*, vol. 53, pp. 144-150, 1958.
6. Croxton, F. E., y D. J. Cowden: *Applied General Statistics*, 3ª ed. Prentice-Hall, Inc., Englewood Cliffs, N. J., 1967, cap. 21.
7. Davis, J. A.: "A Partial Coefficient for Goodman and Kruskal's Gamma", *Journal of the American Statistical Association*, vol. 62, pp. 189-193, 1967.
8. Draper, N. R., y H. Smith: *Applied Regression Analysis*, John Wiley & Sons, Inc., Nueva York, 1966, caps. 5-10.
9. Dwyer, P. S.: *Linear Computations*, John Wiley & Sons, Inc., Nueva York, 1951.
10. Gordon, Robert: "Issues in Multiple Regression", *American Journal of Sociology*, vol. 73, pp. 592-616, 1968.
11. Hagood, M. J., y D. O. Price: *Statistics for Sociologists*, Henry Holt and Company, Inc., Nueva York, 1952, cap. 25.
12. Johnston, J.: *Econometric Methods*, McGraw-Hill Book Company, Nueva York, 1963.
13. Kendall, M. G.: *Rank Correlation Methods*, Hafner Publishing Company, Inc., Nueva York, 1955, cap. 8.
14. Land, K. C.: "Principles of Path Analysis", en Edgar Borgatta (ed.), *Sociological Methodology 1969*, Jossey-Bass, Inc., Publishers, San Francisco, 1969, cap. 1.
15. Morris, R. N., "Multiple Correlation and Ordinally Scaled Data", *Social Forces*, vol. 48, pp. 299-311, 1970.
16. Quade, Dana: *Nonparametric Partial Correlation*, University of North Carolina, Institute of Statistics Mimeo Series, núm. 526, 1967.
17. Reynolds, H. T.: *Making Causal Inferences with Ordinal Data*, University of North Carolina, Institute for Research in Social Science, Chapel Hill, 1971.
18. Simon, H. A.: "Spurious Correlation: A Causal Interpretation", *Journal of the American Statistical Association*, vol. 49, pp. 467-479, 1954.
19. Somers, R. H.: "An Approach to the Multivariate Analysis of Ordinal Data", *American Sociological Review*, vol. 33, pp. 971-977, 1968.
20. Wilson, T. P.: "A Critique of Ordinal Variables", *Social Forces*, vol. 49, pp. 432-444, 1971.

XX. ANÁLISIS DE COVARIANCIA Y VARIABLES SIMULADAS

HEMOS estudiado el análisis de variancia en que una sola escala de intervalo puede relacionarse con una o más escalas nominales. En el capítulo anterior vimos cómo las técnicas de la correlación podían emplearse para relacionar cualquier número de escalas de intervalo. En el análisis de covariancia combinamos ahora las ideas básicas del análisis de variancia y del análisis de correlación, con objeto de tratar problemas que comportan más de una escala de intervalo en combinación con cualquier número de escalas nominales. Así, pues, el análisis de covariancia es una extensión teórica de estos dos procedimientos, que nos pone idealmente en condiciones de tratar problemas que comporten diversas combinaciones de escalas de intervalo y nominales.

Por desgracia, según veremos en seguida, los cálculos requeridos por el análisis de covariancia son muy fastidiosos si se realizan a mano o con una calculadora de escritorio, pero no se plantean problemas especiales si se dispone de programas de computación. En un terreno ideal cabe ampliar el procedimiento hasta incluir el manejo de un gran número de variables independientes nominales y de escalas de intervalos, a condición de que la variable dependiente sea una escala de intervalo. En la práctica, sin embargo, uno se encuentra limitado a tres o cuatro variables independientes por razón de que las interacciones de más elevado orden resultan muy numerosas pasado aquel límite. El análisis de la covariancia es, en su forma, equivalente a un procedimiento denominado de análisis por "variable simulada", que será explicado al final del capítulo. Este procedimiento equivale a una simple ampliación del modelo de regresión, y el estudio de ambos sistemas suministra una buena apreciación intuitiva de la relación existente entre el análisis de la variancia y la regresión.

En este capítulo limitaremos nuestra atención al caso de tres variables, en el que tenemos una escala nominal y dos escalas de intervalo. El problema básico del que nos ocuparemos es el de relacionar dos de dichas variables controlando en relación con la tercera. Si bien semejante control podría efectuarse tomando categorías de la variable de control y llevando a cabo análisis separados en el interior de esas clases, es posible, con todo, obtener una eficacia mucho mayor mediante el empleo de las técnicas del análisis de covariancia, a condición de que la interacción no sea significativa. En otros términos: el control puede efectuarse sin necesidad de tener que tomar un número sumamente grande de casos. Efectivamente, nos servimos de promedios ponderados y de procedimientos de ajuste, como lo hicimos en el caso de