localidad grande, etcétera). ¿Cuántas comparaciones mutuamente ortogonales serían posibles? Hállese un grupo específico del anterior número de comparaciones que sean mutuamente ortogonales, comprobando que así es el caso.

Bibliografía

1. Anderson, R. L., y T. A. Bancroft: *Statistical Theory in Research,* McGraw-Hill Book Company, Nueva York, 1952, caps. 17 y 18.
2. Blalock, H. M.: "Theory Building and the Statistical Concept of Interaction", *American Sociological Review,* vol. 30, pp. 374-380, 1965.
3. Bradley, J. V.: *Distribution-free Statistical Test,* Prentice-Hall, Inc., Englewood Cliffs, N. J., 1968, cap. 5.
4. Dixon, W. J., y F. J. Massey: *Introduction to Statistical Analysis,* 2ª ed., McGraw-Hill Book Company, Nueva York, 1957, cap. 10.
5. Haggard, E. A.: *Intraclass Correlation and the Analysis of Variance,* The Dryden Press, Inc., Nueva York, 1958, caps. 1-5.
6. Hagood, M. J., y D. O. Price: *Statistics for Sociologists,* Henry Holt and Company, Inc., Nueva York, cap. 22.
7. Hays, W. L.: *Statistics,* Holt, Rinehart and Winston, Inc. Nueva York, 1963, caps. 11-14.
8. Johnson, P. O.: *Statistical Methods in Research,* Prentice-Hall, Inc. Englewood Cliffs, N. J., 1949, caps. 10 y 11.
9. Kirk, R. E.: *Experimental Design: Procedures for the Behavioral Sciences,* Brooks/Cole Publishing Company, Belmont, Cal., 1968, cap. 3.
10. Siegel, S.: *Nonparametric Statistics for the Behavioral Sciences,* McGraw-Hill Book Company, Nueva York, 1956, pp. 166-172, 184-193.
11. Walker, H. M., y J. Lev: *Statistical Inference,* Henry Holt and Company, Inc., Nueva York, 1953, cap. 14.

XVII. CORRELACIÓN Y REGRESIÓN

En el presente capítulo y en el siguiente examinaremos la relación entre dos escalas de intervalo. La extensión a tres o más variables de escala de intervalo se verá en el capítulo XIX, al tratar de la correlación múltiple y parcial. De momento, consideramos situaciones en las que tenemos dos medidas de escala de intervalo por cada individuo. Así, por ejemplo, podemos conocer el número de años de enseñanza completados y el ingreso anual de los varones adultos de una localidad determinada. O puede interesarnos relacionar el porcentaje de mano de obra empleado en la industria con el crecimiento demográfico de una población.

En algunos problemas de esta índole nos interesamos a menudo no sólo en las pruebas de significación y las medidas de grados de relación, sino que podemos también querer describir la *naturaleza* de la relación entre las dos variables, de modo que, conociendo una de ellas, podamos anticipar la otra. Así, por ejemplo, podemos querer predecir el ingreso futuro de una persona sobre la base de su instrucción, o la tasa de crecimiento de una ciudad a partir del porcentaje de su mano de obra empleada en la industria. Cuando el interés se centra ante todo en la tarea exploradora de encontrar *cuáles* variables se relacionan con una variable determinada, nos interesamos por lo regular principalmente por las medidas de grados o fuerza de las relaciones, tales como los coeficientes de correlación. Por otra parte, una vez halladas las variables significativas, propendemos a dirigir nuestra atención al análisis de regresión, en el que intentamos predecir el valor exacto de una variable a partir de la otra.

Si bien el lector ya está familiarizado con las pruebas de significación y las medidas de asociación, recomiéndase, con todo, empezar nuestro examen estudiando el problema de la predicción. Esto se debe a que la noción de regresión es a la vez anterior lógicamente y más importante teóricamente que la de correlación. La razón de ello se irá viendo más clara a medida que vayamos avanzando. Después de haber examinado el problema de la predicción, dirigiremos nuestra atención a la medición de la fuerza de la relación. En el capítulo XVIII, que de hecho representa la continuación del presente, examinaremos diversas pruebas de significación, así como la correlación del orden de lugares, que pueden emplearse para relacionar dos escalas ordinales.

XVII.1. *Regresión lineal y mínimos cuadrados*

En cierto sentido, el objetivo último de todas las ciencias es el de la predicción. Esto no implica, por supuesto, que sólo secun-

dariamente estemos interesados en comprender o suministrar explicaciones causales de por qué dos o más variables se relacionan como lo hacen. Tal vez sea más acertado decir que la comprensión constituye el objetivo final y que, en la medida en que la comprensión se va perfeccionando, la predicción se hace cada vez más precisa. Es posible que si la comprensión fuera completa la predicción perfecta sería también posible siempre que se conociera asimismo cierta información factual necesaria. Por ejemplo: si uno conoce las leyes del movimiento de los planetas, el campo gravitatorio dentro del sistema solar, *y* la posición y la velocidad de Venus en determinado momento, podría predecir su movimiento futuro. Sin embargo, independientemente de las implicaciones filosóficas de semejante punto de vista determinista, lo cierto es que la predicción constituye el objetivo de toda ciencia.

En sociología y en otras ciencias sociales, los enunciados predictivos se formulan a menudo, por necesidad, en forma relativamente burda. Por lo regular esto se debe a que no hemos alcanzado el nivel de medición de la escala de intervalo. Así, por ejemplo, podríamos predecir que cuanto más elevada sea la posición de una persona en el grupo, tanto mayor será su conformación a las normas de éste. Semejante enunciado no necesita implicar causalidad en una sola forma, sino que afirma simplemente que la posición y la conformidad se relacionan de modo positivo. Estableciendo una analogía con una terminología matemática que no es estrictamente correcta, decimos que la posición es una *función* de la conformidad, o que la conformidad es una función de la posición, eludiendo la cuestión de la causalidad. Obsérvese, sin embargo, que hemos dicho muy poco acerca de la *forma* de esta relación, aparte de indicar que es positiva. Y a menos que tengamos un nivel de medición de escala de intervalo para ambas variables, resulta efectivamente muy difícil decir mucho más.

Supóngase, sin embargo, que tenemos dos escalas de intervalo. Se hace entonces posible describir más exactamente de qué modo una de las variables varía con la otra. Así, por ejemplo podríamos estar en condiciones de decir que, por cada año de instrucción recibida, el ingreso aumentará en $1 000. Si esto fuera efectivamente así, tendríamos en realidad una relación muy simple, o sea una relación lineal o en línea recta. Sin embargo, la mayoría de las relaciones no son ni con mucho tan sencillas, pese a que, según veremos, resulta a menudo posible obtener una aproximación muy buena de la verdadera relación suponiendo linealidad. La forma más elegante y sencilla de expresar una relación entre dos (o más) variables es por medio de una ecuación matemática. Así, por ejemplo, el lector estará familiarizado con ciertas leyes físicas que enuncian una relación entre la presión, el volumen y la temperatura ($PV/T = k$), o que indican una relación entre la razón de aceleración de un cuerpo al caer, la distancia recorrida y la duración del tiempo en que ha estado cayendo. Podemos también representar cada una de estas ecuaciones matemáticas como alguna clase de curva geométrica. Afortunadamente, en sociología solemos por lo regular operar con ecuaciones muy simples y con las curvas más simples posibles (rectas).

Cuando añadimos más variables, no podemos representar tan fácilmente las ecuaciones como figuras geométricas, ya que nos salimos de las dimensiones, de lo cual, sin embargo, no necesitamos preocuparnos por el momento.

Supóngase que hay una variable dependiente Y que ha de predecirse a partir de una variable independiente X. En algunos problemas, X precederá obviamente a Y en el tiempo. Por ejemplo: por lo regular una persona completa su instrucción antes de obtener un ingreso. En tales casos, semejante manera de representar las cosas resulta muy adecuada, pese a que hemos de poner cuidado en no implicar una relación necesaria o causal, o que X es la única variable que influye sobre el valor de Y. Si la dirección de la causa es ambigua, o si se piensa que cada variable es causa de la otra, necesitaremos, si es que deseamos suministrar una *explicación* teórica de la relación, usar un método de ecuaciones simultáneas que escapa a este texto. (Véanse [1], [2] y [6]. Si nuestro objetivo es una simple estimación o una predicción a plazo breve de Y a partir de X, no se presentarán tales ambigüedades, aunque deba señalarse una vez más que no hay nada en las operaciones estadísticas que nos impida realizar operaciones matemáticas teóricamente carentes de sentido. En éste y en los capítulos sucesivos supondremos que la variable Y, seleccionada como dependiente en sentido matemático, es asimismo causalmente dependiente, de manera que la interpretación teórica puede resultar relativamente directa.

Ya vimos que si X y Y son estadísticamente independientes, no podemos predecir Y a partir de X o, más exactamente, el conocimiento de X no mejora en nada nuestra predicción de Y. Presumiblemente, pues, cuando las variables no son estadísticamente independientes, el conocimiento de X sí nos ayuda a predecir Y. Cuanto más fuerte sea la dependencia, tanto más precisa será nuestra predicción. Más adelante mediremos la fuerza de esta relación por medio de coeficientes de correlación. Nos concentramos de momento en la cuestión acerca de *cómo* predecimos Y a partir de X. Así, por ejemplo, podemos querer estimar el ingreso futuro de un individuo, sabiendo que ha completado tres años de escuela secundaria. Sin este conocimiento relativo a la instrucción, nuestra mejor estimación (suponiendo que no hay inflación) sería la del ingreso medio de todos los varones adul-

tos. En cambio, el hecho de conocer su instrucción debería permitirnos obtener una predicción mejor.

La ecuación de regresión. Representémonos el problema de la siguiente manera. Nos imaginamos que para cada valor fijo de la variable independiente X (instrucción) tenemos una distribución de Y (ingresos). En otros términos: para cada nivel educacional

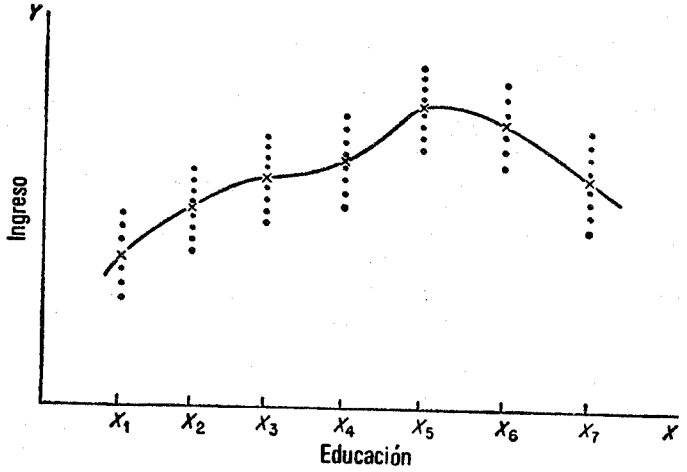


FIG. XVII.1. *Forma general de la regresión de* Y *sobre* X, *o curso de las medias de los valores de* Y *para valores fijos de* X.

habrá cierta distribución de ingresos en la población. No todas las personas que han terminado la escuela secundaria tendrán exactamente los mismos ingresos, por supuesto, pero dichos ingresos estarán con todo distribuidos alrededor de alguna media. Y habrá distribuciones de ingresos similares para los egresados de la escuela primaria, los de la universidad, los posgraduados, etcétera. Cada una de estas distintas distribuciones de ingresos (para X determinadas) tendrá una media, y podemos hacer una gráfica de la posición de dichas medias sirviéndonos del sistema familiar de las coordenadas rectangulares. Designamos el *curso* resultante *de estas medias* de las Y para X fijas como *ecuación de regresión de* Y *a* X. Semejante ecuación de regresión puede verse ilustrada en la figura XVII.1.

Estas ecuaciones de regresión son las "leyes" de la ciencia. En algunos casos hay muy poca dispersión alrededor de la ecuación de regresión. En tales casos, pueden hacerse predicciones muy precisas, y las desviaciones respecto de la ley se consideran a menudo como error de medición o como resultado de influencias menores no controladas. La "ley" puede formularse así como si existiera una perfecta relación entre Y y X. En el caso ideal,

se consideraría que todos los puntos caen exactamente en la curva, y la relación se abstraería como una función matemática perfecta en la que no hay más que una sola Y para cada X. En las ciencias sociales no podemos ser ni con mucho tan exigentes. En efecto, esperamos una variabilidad considerable alrededor de la ecuación de regresión, y preferimos pensar en términos de medias y de variancias de una distribución de Y para cada X. Sin embargo, el procedimiento es en principio el mismo en todas las ciencias, pese a que las leyes de las ciencias sociales no sean tan precisas como las de la física.

En la figura XVII.1 hemos indicado el carácter general de las ecuaciones de regresión, que comportan los cursos de las medias de los valores de Y para determinados valores de X. Vamos a tener que proceder ahora a algunos supuestos simplificadores, con objeto de poder tratar el problema estadísticamente. Si bien la idea de regresión es perfectamente general, la mayoría de la labor estadística sólo se ha realizado con los más simples de los modelos. En particular, vamos a suponer de momento: *1)* que la forma de la ecuación de regresión es lineal, *2)* que las distribuciones de los valores de Y para cada X son normales, y *3)* que las variancias de las distribuciones de Y son las mismas para cada valor de X. Podemos ahora hacer un examen de estos diversos supuestos uno por uno, prestando la mayor atención al primero de ellos.

Si la regresión de Y a X es lineal, o sea una relación en línea recta, podemos escribir una ecuación como sigue:

$$Y = \alpha + \beta X \qquad \text{(XVII.1)}$$

en la que α y β son constantes. La ecuación (XVII.1) indica que la relación entre X y Y es exacta, pero en breve hemos de introducir en la ecuación un término de error. Una forma alternativa de escribir la ecuación es la siguiente: $E(Y|X) = \alpha + \beta X$; en la que $E(Y|X)$ pone de relieve que estamos preocupados con el valor esperado de Y, el que depende de X. Hemos utilizado letras griegas, ya que de momento tratamos de la población total. En una ecuación de esta clase, tanto α como β tienen interpretaciones geométricas definidas. Si ponemos X igual a cero, vemos que $Y = \alpha$. Por consiguiente, α representa el punto en donde la línea de la regresión corta el eje de las Y (o sea, allí donde $X=0$).

La inclinación de la línea de la regresión está dada por β, ya que esta constante indica la magnitud del cambio de Y para una unidad de cambio en X. El hecho de que la relación sea lineal significa que todo cambio de X, digamos en 5 unidades, produce siempre el mismo cambio en Y (esto es, 5β unidades, independientemente de la posición sobre el eje de X (véase fig. XVII.2). El lector ha de convencerse por sí mismo que si $\beta = 1$ y si las uni-

dades de X y Y están indicadas por distancias iguales a lo largo de los respectivos ejes, la línea de regresión formará un ángulo de 45 grados con el eje de las X. Una β mayor que la unidad indica una pendiente más rápida. Cuanto más rápida sea la pendiente, tanto mayor es el cambio de Y para un cambio dado de X. Y en forma análoga, si β es menor que la unidad pero mayor que cero, se requerirá un cambio mayor de X para producir un cambio

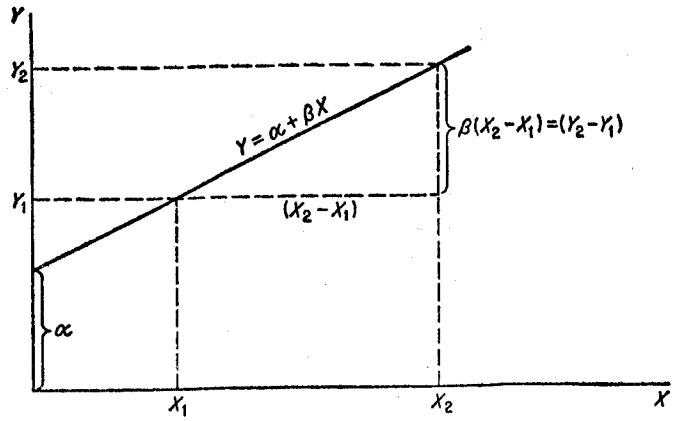


FIG. XVII.2. *La ecuación lineal de regresión, mostrando interpretaciones geométricas de* α *y* β.

dado en Y. En el caso límite, en que la línea es horizontal, β se hace cero, y los cambios de X no producen cambios de Y. En otros términos, si $\beta = 0$, no existe relación lineal entre X y Y. El conocimiento de X no nos ayuda a predecir Y, si se supone un modelo lineal.[1] Si β es negativa, sabemos que se da una relación negativa entre las dos variables, y que mientras X crece, Y decrece.

Una línea recta puede determinarse siempre por completo si conocemos ya sea dos puntos de la línea o un punto y la pendiente. Por lo tanto, no hay más que una sola línea de ecuación $Y = \alpha + \beta X$, a condición, por supuesto, que se considere a α y β como cantidades fijas (pero generales). Si α y β están dadas, podemos trazar la recta tomando simplemente dos puntos de la misma. Sabemos que cuando $X = 0$, $Y = \alpha$. Por consiguiente, el punto $(0, \alpha)$ se sitúa en la recta. Y así también, cuando $Y = 0$, tenemos $0 = \alpha + \beta X$ o $X = -\alpha/\beta$. Este punto $(-\alpha/\beta, 0)$ es, por supuesto, el punto en donde la línea corta el eje de las X. Si no

[1] Según veremos más adelante, la independencia estadística asegura que β sea cero, pero no se sigue necesariamente de ahí que si β es cero tengamos independencia.

conviene servirse de dichos dos puntos, pueden determinarse otros dos puntos cualesquiera por el mismo procedimiento.[2]

Supuestos acerca de X *y el término de perturbación.* Hasta ahora no hemos tratado en forma explícita el hecho de que, puesto que habrá dispersión alrededor de la ecuación de regresión, habremos de representar el valor real de Y para cada individuo mediante una ecuación que contenga un término de perturbación o de error que es único para cada individuo. Si suponemos que Y_i y X_i se refieren a las puntuaciones correspondientes al i-ésimo individuo, podremos representar la relación (lineal), como sigue:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

en la que ε_i representa el término de perturbación, cuyo comportamiento necesitamos estudiar. Podemos concebir este término como si contuviera el error de medición en Y (pero no en X), y como resultante de todas las varias causas de Y que no han sido llevadas a la ecuación de una manera explícita. Si la mayor parte de estas causas omitidas tienen individualmente un efecto menor, y si además están operando casi independientemente entre ellas, será razonable suponer que el valor esperado correspondiente al factor de perturbación $E(\varepsilon_i)$ será igual a cero, y que ε_i estará distribuido en forma aproximadamente normal. Lo que resulta muy importante es el hecho de que el factor de perturbación será estadísticamente independiente de X. Resulta que al usar mínimos cuadrados para estudiar los coeficientes de regresión α y β, es necesario suponer que $E(\varepsilon) = 0$, y que X_i y ε_i no están relacionados. La suposición de normalidad, más la suposición de homoscedasticidad, de que σ_ε^2 es constante a través de todos los niveles de X será necesaria en las pruebas de significancia y para la determinación de los límites de confianza.

El supuesto fundamental que subraya el uso del análisis de regresión es el de que X sea independiente del factor de error. En aplicaciones experimentales nos encontramos con frecuencia en la posibilidad de elegir niveles fijos de X (como, por ejemplo, cuando mantenemos constantes de temperatura a intervalos de 50 grados). En tales casos, puesto que el nivel de X está bajo nuestro control y se presume que no es manipulado en forma que varíe sistemáticamente con el factor de perturbación, será raro preocuparse con este supuesto concreto. Un momento de reflexión nos convencería, sin embargo, de que en muchas situaciones experimentales incluso este supuesto es inocente, ya que al manipular X uno puede inadvertidamente afectar otros factores que se quedaron fuera de la ecuación y contenidos por lo tanto en el factor de perturbación.

En la investigación no experimental se toma tanto a las X como

[2] Véase un ejemplo numérico en la página 392.

a las Y como observadas y no como manipuladas, siendo por lo tanto X y Y variables aleatorias, o lo que se denomina variables *estocásticas*, las que tienen una distribución de probabilidad. En algunos casos la distribución de X será aproximadamente normal, aunque esto no es necesario en el caso del análisis de regresión. Lo que *resulta* esencial, sin embargo, es el formular algunos supuestos acerca de la distribución conjunta de X_i y el factor de perturbación ε_i. Si tuviéramos *a priori* razones sólidas para especificar alguna distribución particular, esto resultaría suficiente, pero en la práctica se carece siempre de tal información. Con mucha frecuencia suponemos que X_i y ε_i son estadísticamente independientes, supuesto que resultará justificado si las causas de Y omitidas son, *1)* numerosas, aisladamente sin importancia, y no muy interrelacionadas, o *2)* sin relación con X en situaciones en las que predominan uno o dos de los factores omitidos. Si uno no está dispuesto a hacer tal suposición en algún caso particular, deberá tratar de identificar los mayores factores perturbadores que hayan sido omitidos, introduciéndolos explícitamente en la ecuación como variables adicionales. En el capítulo XIX examinaremos la regresión múltiple, en la que han sido incluidos tales factores causales adicionales.

Una de las ventajas de la teoría estadística del análisis de regresión consiste en que está lo suficientemente desarrollada como para que tales supuestos acerca del comportamiento de los factores de perturbación resulten explícitos. Resultará bien claro que lo que hemos dicho acerca del comportamiento de las variables omitidas se aplica igualmente bien a *todos* los procedimientos que hasta aquí hemos examinado. Si se encuentra, por ejemplo, una diferencia estadísticamente significativa en medias o proporciones, y si se desea atribuir una explicación causal a la variable independiente (por ejemplo, sexo) en esta relación, habrá que suponer también que los factores omitidos no están sistemáticamente relacionados con la escala nominal dicotomizada (por ejemplo sexo). No es posible soslayar supuestos acerca de variables omitidas cambiando simplemente el tipo del análisis y confiando en que así desaparecerá el problema.

Ya se indicó más arriba que para las pruebas de significación hemos de suponer que las Y están distribuidas normalmente alrededor de cada valor de X. Para las X estocásticas convendrá también suponer que para cada valor fijo de Y las X están asimismo distribuidas normalmente. Decimos que la distribución conjunta de X y Y es una distribución *normal bivariable*, lo que significa que hay dos variables, cada una de las cuales está distribuida alrededor de la otra en forma normal. Semejante distribución normal bivariable tiene una ecuación matemática definida y puede representarse como una superficie tridimensional, como en la figura XVII.3. La altura de la superficie en un punto

dado (X, Y) es proporcional al número de casos en el mismo. Así, pues, se requiere un diagrama tridimensional para representar la distribución conjunta entre X y Y, del mismo modo que necesitábamos dos dimensiones para representar la distribución de frecuencia de la X sola. La forma exacta de esta figura, que se

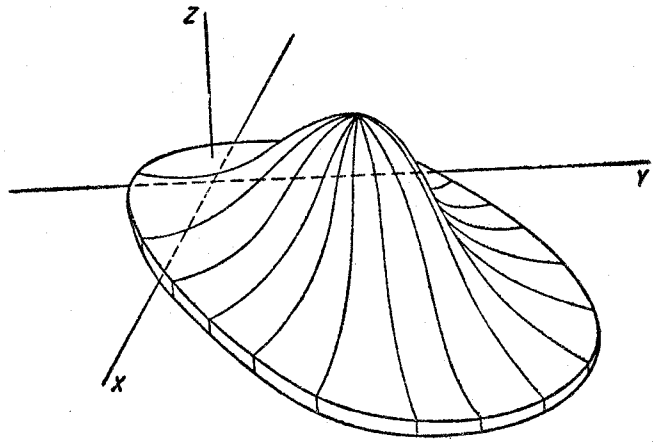


FIG. XVII.3. *La distribución normal bivariable.* (Con autorización de A. M. Mood, *Introduction to the Theory of Statistics*, McGraw-Hill Book Company, Inc., Nueva York, 1950, fig. 41, p. 165.)

parece mucho a un casco de bombero, dependerá de cuán cercanamente estén relacionadas las variables entre sí.

Si ambas variables se han expresado en términos de unidades de desviación estándar, entonces, cuanto más relacionadas estén las variables tanto más angosto será el casco. En el caso extremo, en el que Y puede predecirse exactamente a partir de X y, por consiguiente, todos los puntos están exactamente en la ecuación de regresión, las desviaciones estándar de las Y para cada X serían cero, y el casco no tendría grueso alguno. Por otra parte, si no existiera relación alguna entre X y Y, la base del casco sería más aproximadamente circular. Cualquier plano perpendicular al plano XY cortaría la superficie en una curva normal. En tanto que un plano paralelo al plano XY cortará el casco en una elipse. La distribución normal bivariable posee la propiedad de que la regresión de Y a X sea lineal. Por lo tanto, si tenemos una distribución normal bivariable, sabemos que, si trazamos las medias de las Y para cada X, el resultado será una recta. No se sigue de ahí, sin embargo, que si la regresión es lineal, la distribución conjunta sea necesariamente normal bivariable.

En el caso de las pruebas de significancia necesitaremos tam-

bién suponer que las desviaciones estándar de las Y para cada X son las mismas, independientemente del valor de X. Este supuesto se examinará en conexión con el tema de la correlación, ya que ésta es esencialmente una medida de dispersión alrededor de la línea de regresión. De momento basta, con todo, señalar que si la distribución conjunta es normal bivariable, las desviaciones estándar de las Y para cada X serán de hecho todas idénticas. Esta propiedad de variancias iguales se designa como *homoscedasticidad* y es análoga al supuesto hecho en el análisis de variancia de que $\sigma_1 = \sigma_2 = \ldots = \sigma_k$.

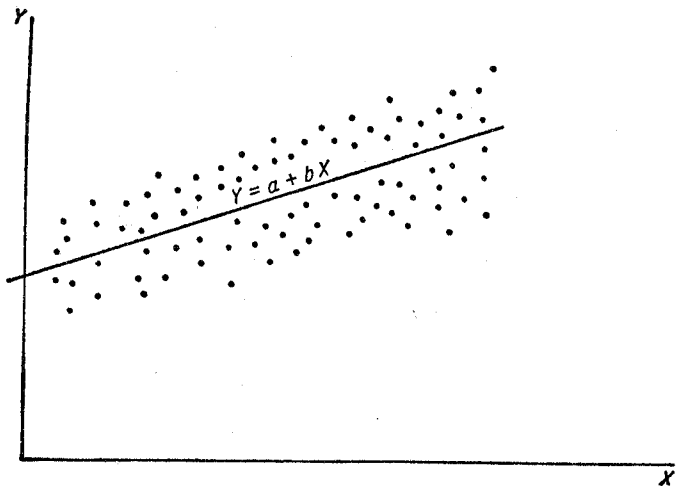


FIG. XVII.4. *Diagrama de dispersión y recta de mínimos cuadrados.*

Mínimos cuadrados lineales. El modelo de regresión que hemos estado examinando es más bien sencillo en sus conceptos, pero no es por desgracia directamente útil en su forma teórica. Es raro, en efecto, que tengamos suficientes casos para examinar la distribución de las Y para valores fijos sucesivos de X. Con mayor frecuencia encontramos que hay relativamente pocos casos en los que las X sean idénticas o aproximadamente tales. Si hacemos una gráfica de la distribución de los casos alrededor de los ejes de las X y las Y en la forma convencional, encontramos por lo regular una dispersión de puntos como la que se indica en la figura XVII.4. Y si hacemos una gráfica de la distribución de los puntos en esta forma, obtenemos lo que se designa como *esquedograma* o diagrama de dispersión. El estudiante ha de acostumbrarse a dibujar un diagrama de dispersión antes de proceder al análisis ulterior. La mera inspección del diagrama en cuestión, en efecto, puede acaso indicar que no tiene objeto seguir adelante. Así, por ejemplo, si los puntos aparecen en el diagrama como si estuvieran distribuidos al azar, resulta claro que no existe relación, o sólo una relación muy débil, entre las dos variables.

Una vez fijadas las marcas en un diagrama de dispersión, podemos querer acercarnos a dichos puntos por medio de alguna clase de curva que sea la más adecuada. Una de las maneras de hacerlo es trazar una curva (en el presente caso una recta) por inspección. Sin embargo, existen para ello métodos más precisos. Uno de éstos es el método de los mínimos cuadrados, que se examinará en la presente sección. Nuestro objetivo es ahora algo distinto del objetivo del análisis de regresión, en el que trazábamos el curso de la media de las Y. Aquí, en efecto, queremos aproximarnos a cierto número de puntos por medio de una curva de mejor adaptación.

Con objeto de servirnos de la teoría de los mínimos cuadrados, hemos de postular la forma de la curva a utilizar en la adaptación de los datos. En el caso del análisis de regresión, la forma de la curva se hallaría propiamente determinada por el curso de las medias, suponiendo que se dispone de datos relativos a la población entera. Vamos a tomar una vez más la curva más simple posible, la recta, como curva de nuestros mínimos cuadrados. Esto significa que hemos de adaptar los datos a una recta de mejor ajuste, conforme al criterio de los mínimos cuadrados, obteniendo una ecuación de la forma:

$$Y = a + bX \qquad \text{(XVII.2)}$$

Resultará así que la a y la b obtenidas con este método son las apreciaciones insesgadas más eficaces de los parámetros de la población, α y β, si la ecuación de regresión es efectivamente una recta y si suponemos: *1)* Muestreo al azar, *2)* Que $E(\varepsilon_i) = 0$, y *3)* Que X_i y ε_i son estadísticamente independientes.

Nuestro criterio de los mínimos cuadrados comporta hallar la única recta que posee la propiedad de que la suma de los cuadrados de las desviaciones de los valores reales de Y respecto de dicha recta sea mínima. Así, por ejemplo, si trazamos líneas verticales de los puntos a la línea de los mínimos cuadrados, y si elevamos al cuadrado dichas distancias y las sumamos, la suma resultante será menor que la suma correspondiente de cuadrados a cualquier otra recta posible (véase la figura XVII.5). Obsérvese que son las distancias verticales, y no las perpendiculares o las horizontales las que aquí se consideran. Sería posible minimizar la suma de los cuadrados de las distancias perpendiculares (designada como suma ortogonal de los mínimos cuadrados), pero las ecuaciones de ello resultantes no son ni

con mucho tan prácticas. Y si se emplearan las distancias horizontales, la recta de mínimos cuadrados resultante podría utilizarse para apreciar la regresión de X a Y. El lector ha de convencerse por sí mismo que minimizar la suma de cuadrados de las distancias verticales no minimiza necesariamente la suma de cuadrados de las distancias horizontales. Así, pues,

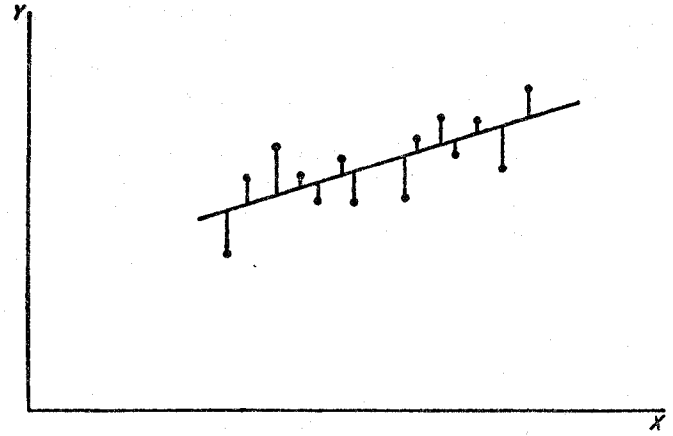


FIG. XVII.5. *Ecuación de mínimos cuadrados, que minimiza las sumas de los cuadrados de las distancias verticales y estima la regresión de* Y *sobre* X.

podemos obtener varias líneas de mínimos cuadrados distintas. Pero éstas sólo coincidirán si todos los puntos quedan exactamente en una sola línea. Resulta asimismo que, al minimizar la suma de los cuadrados de las distancias verticales, encontramos de hecho la recta que posee la propiedad de que la suma de las distancias verticales positivas y negativas sea cero *y* la desviación estándar de los puntos respecto de aquélla sea mínima. Este concepto de la desviación estándar de las Y se examinará con mayor detalle más adelante.

Con objeto de obtener la línea de mínimos cuadrados, pues, necesitamos calcular la a y la b que determinan la línea provista de la propiedad deseada. Esta clase de problemas puede resolverse fácilmente por medio del cálculo y conduce a las siguientes fórmulas de cálculo de a y b.[3]

[3] Para los estudiantes familiarizados con el cálculo elemental vamos a delinear la naturaleza de la derivación. Comenzaremos con la ecuación $Y_i = a + bX_i + e_i$, en la que e_i es un término residual que puede ser utilizado para estimar el residual ε_i de la ecuación de regresión. Deseamos minimizar la suma de los cuadrados de estos residuales, es decir: la cantidad $\Sigma e_i^2 = \Sigma(Y_i - a - bX_i)^2$ con respecto a las dos cantidades a y b, a las que aquí se trata como desconocidas. Tomamos derivativos parciales con respecto a a y b; las hacemos igual a cero, y resolvemos las dos ecuaciones resultantes (a las que se denomina *ecuaciones normales*) para a y b. Este mismo procedimiento es de aplicación al caso multivariado.

$$a = \frac{\sum_{i=1}^{N} Y_i - b \sum_{i=1}^{N} X_i}{N} = \bar{Y} - b\bar{X} \qquad \text{(XVII.3)}$$

y

$$b = \frac{\sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N} (X_i - \bar{X})^2} = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2} \qquad \text{(XVII.4)}$$

en donde $x_i = X_i - \bar{X}$ y $y_i = Y_i - \bar{Y}$. Obsérvese que en estas ecuaciones a y b son las incógnitas, hallándose las otras cantidades determinadas a partir de los datos. Una vez que se haya obtenido b, a puede calcularse fácilmente a partir de la primera de las dos fórmulas. Podemos, pues, centrar nuestra atención en el cálculo de b.

El numerador de b comporta la expresión $\sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})$ que se designa como *covariación* de X y Y. Esta cantidad es directamente análoga a las sumas de cuadrados tanto de X como de Y, excepto que, en lugar de elevar al cuadrado $(X - \bar{X})$ o $(Y - \bar{Y})$, tomamos el producto de estos dos términos. Obtenemos en esta forma una medida de cómo X y Y varían juntas, y de ahí el nombre de *covariación*. Si dividimos esta expresión entre N, obtenemos, por analogía, lo que se designa como *covariancia*. Veremos inmediatamente que b puede ponerse igual a la razón de la covariancia a la variancia en X.

Si examinamos más de cerca la covariación de X y Y, vemos que, a diferencia de una suma de cuadrados, la covariación puede tomar valores tanto positivos como negativos. Si X y Y se relacionan positivamente, entonces valores grandes de X se asociarán por lo regular con valores grandes de Y. Así, pues, si $X > \bar{X}$, será por lo regular cierto que $Y > \bar{Y}$. Y asimismo, en el caso de una relación positiva, si $X < \bar{X}$, tendremos generalmente $Y < \bar{Y}$. Por consiguiente, el producto de $(X - \bar{X})$ y $(Y - \bar{Y})$ será normalmente positivo, y la suma de estos productos será asimismo positiva. Y en forma análoga, si X y Y se relacionan negativamente, esperaríamos que, si $X > \bar{X}$, entonces Y será menor que $\bar{Y}$, y la suma de productos resultante será negativa. Si no existe relación, entonces aproximadamente la mitad de los productos serán positivos y la otra mitad negativos, ya que X y Y variarán indepen-

dientemente. En este caso, b será cero, o vecino de cero. Por lo tanto, cuanto mayor sea el valor numérico de la relación, independientemente de la dirección, tanto mayor será el valor numérico de la covariación. Como habremos de ver en breve, la covariación figura también en el numerador del coeficiente de correlación, que es nuestra medida del grado de asociación. En el caso de b, tomamos la covariación y la dividimos entre la suma de los cuadrados en X, con objeto de obtener nuestra estimación de la pendiente de la ecuación de regresión.

Es más conveniente servirse para la covariación de una fórmula que es directamente análoga a la fórmula de cálculo de la suma de los cuadrados y puede derivarse en forma similar. Podemos escribir la fórmula de cálculo de b como sigue:

$$b = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2} \qquad \text{(XVII.5)}$$

En la ecuación (XVII.5), tanto el numerador como el denominador se han multiplicado por N, con objeto de redondear los errores debidos a la división y con objeto de facilitar el cálculo con una calculadora.[4]

Problema. Supóngase que tenemos los datos del cuadro XVII.1, en donde X representa el porcentaje de negros en las grandes ciudades del Medio Oeste, y Y indica la diferencia entre las medianas de los ingresos de los blancos y los negros, como medida de discriminación económica.[5]

CUADRO XVII.1. *Datos para un problema de correlación*

Porcentaje de negros X	*Diferencia de ingresos* Y	*Porcentaje de negros* X	*Diferencia de ingresos* Y
2.13	$ 809	4.62	$ 859
2.52	763	5.19	228
11.86	612	6.43	897
2.55	492	6.70	867
2.87	679	1.53	513
4.23	635	1.87	335
		10.38	868

4 En esta y las fórmulas posteriores hemos prescindido de los subíndices, ya que se opera siempre la suma total de los casos, del cuadro N.

5 Aunque la palabra "negro" puede resultar ofensiva para algunos lectores, resulta necesario mantener esta terminología al referirse a los datos del censo, como contraste con otros datos hipotéticos o los obtenidos de otras fuentes.

A partir de los datos podemos calcular cinco sumas que, junto con N, son todo lo que necesitamos para tratar los problemas de regresión y correlación. Todas estas sumas menos una se emplearán en los cálculos de a y b. Los cálculos pueden resumirse como sigue:

$$N = 13 \qquad \Sigma Y = 8\,557$$
$$\Sigma X = 62.88 \qquad \Sigma Y^2 = 6\,192\,505$$
$$\Sigma X^2 = 432.2768 \qquad \Sigma XY = 43\,943.32$$

Aquí la única cantidad nueva es ΣXY. Si ponemos estos valores en las fórmulas de a y b, tenemos ahora:

$$b = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2}$$

$$= \frac{13(43\,943.32) - (62.88)(8\,557)}{13(432.2768) - (62.88)^2} = \frac{33\,199.0}{1\,665.7} = 19.931$$

y

$$a = \frac{\Sigma Y - b\Sigma X}{N}$$

$$= \frac{8\,557 - (19.931)(62.88)}{13} = 561.83$$

Por lo tanto, la ecuación lineal resultante es:

$$Y_p = a + bX = 561.83 + 19.931X$$

en donde hemos utilizado Y_p para indicar que los valores de Y se han estimado a partir de una ecuación de mínimos cuadrados. Como ya se indicó anteriormente, las a y b obtenidas por este método son las estimaciones insesgadas más eficaces de α y β, o sea los coeficientes de regresión reales a condición de que el factor de perturbación ε_i en la ecuación $Y_i = \alpha + \beta X_i + \varepsilon_i$ tenga un valor esperado de cero no relacionado con X, y siempre, por otra parte, de que tengamos una muestra al azar de la población que estudiamos. Por consiguiente, la línea de mínimos cuadrados será la mejor apreciación de la verdadera regresión, si la ecuación de regresión es efectivamente lineal.

La ecuación de los mínimos cuadrados posee asimismo la propiedad de pasar por el punto $(\overline{X},\overline{Y})$, que representa las medias de X y de Y. Esto puede verse en la ecuación (XVII.3). Ya que

$$a = \overline{Y} - b\overline{X}$$

tenemos:

$$\overline{Y} = a + b\overline{X}$$

lo que indica que estos valores de X y Y satisfacen la ecuación. Por consiguiente, el punto $(\overline{X},\overline{Y})$ queda exactamente sobre la línea.

En el problema anterior, si sabemos el valor de X (porcentaje de negros) para cualquier ciudad dada del Medio Oeste, nuestra mejor estimación del valor de Y sería aquel valor de Y que co-

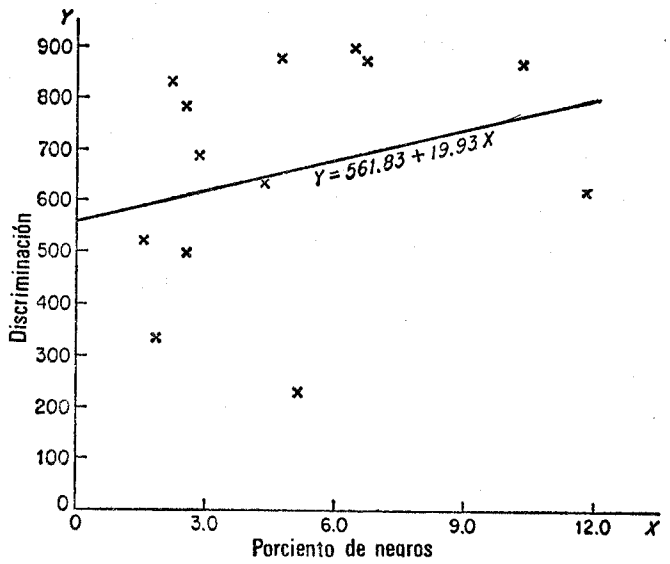


FIG. XVII.6. *Diagrama de dispersión y recta de mínimos cuadrados para los datos del cuadro XVII.1.*

rresponde en la ecuación de los mínimos cuadrados a la X dada. Como quiera que las marcas de discriminación indican diferencias (en dólares) entre los ingresos (en medianas) de los blancos y los negros, vemos que un aumento del 1 por ciento de los negros corresponde a una diferencia de $ 19.93 en dichos ingresos. En la figura XVII.6 se han trazado un diagrama de dispersión y la ecuación de los mínimos cuadrados. Con objeto de ilustrar el empleo de semejante ecuación de predicción, si supiéramos que había un 8 por ciento de negros en una ciudad determinada, la diferencia estimativa del ingreso mediano sería:

$$Y_p = a + b(8) = 561.83 + (19.931)(8) = \$\,721.28$$

Vemos en la figura que se habría obtenido aproximadamente el mismo resultado con la gráfica. Observemos de paso que, haciendo $X = 8$ y resolviendo en relación con Y, hemos localizado un

segundo punto de la línea, que puede utilizarse a continuación con objeto de trazar la línea en el diagrama de dispersión.

XVII.2. *Correlación*

Supongamos a partir de ahora que X es estocástica, y no sometida por tanto al control del investigador. No sólo deseamos conocer la *forma* o la naturaleza de la relación entre X y Y, de modo que una de las variables pueda predecirse a partir de la otra, sino que es necesario al propio tiempo conocer el *grado* o fuerza de la relación. Es obvio que si la relación es muy débil, no tiene objeto tratar de predecir Y a partir de X. Los sociólogos tienen a menudo interés ante todo en descubrir *cuáles* de un gran número de variables se relacionan más de cerca con una variable dependiente determinada. En los estudios de exploración de esta clase, el análisis de regresión reviste importancia secundaria. A medida que una ciencia va madurando y que se descubren variables importantes, la atención puede centrarse en métodos de predicción exacta. Algunos estadígrafos son del parecer que en conjunto se ha prestado demasiada atención a la correlación y casi ninguna al análisis de regresión. Que esto sea así o que no lo sea depende, por supuesto, del estado del conocimiento en la ciencia considerada.

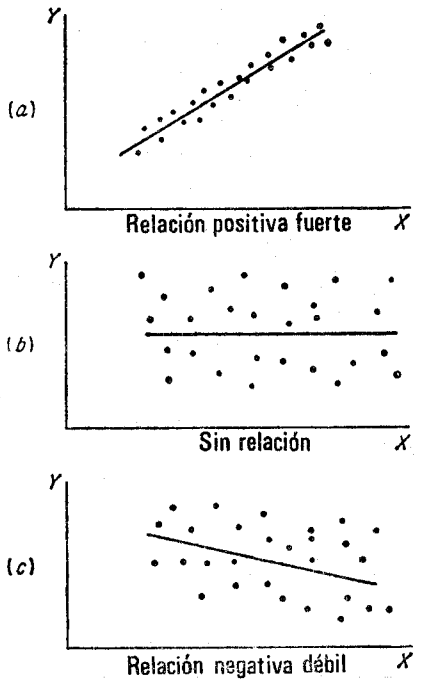


FIG. XVII.7. *Diagrama de dispersión que muestra las diferentes fuerzas y direcciones de las relaciones entre* X *y* Y.

El coeficiente de correlación r, que vamos a examinar en esta sección, fue introducido por Karl Pearson y se designa a menudo como correlación momento-producto, con objeto de distinguirla de otras medidas de asociación. Este coeficiente mide la cantidad de dispersión alrededor de la ecuación *lineal* de los mínimos cuadrados. Hay un coeficiente correspondiente de población rho (ρ), que mide la bondad del ajuste a la verdadera ecuación de

regresión. Obtenemos una estimación r de dicho parámetro midiendo las desviaciones respecto de la línea calculada por medio de los mínimos cuadrados.

Como quiera que la ecuación de regresión representa el curso de las medias de las Y para unas X dadas, sería también posible medir la dispersión respecto de esa línea tomando una desviación estándar de la misma.[6] Sin embargo, los investigadores de la mayoría de los campos de aplicación se han acostumbrado al coeficiente de correlación; es probable, con todo, que el coeficiente de correlación se mantenga. Posee la ventaja de ser de fácil interpretación, y su recorrido va de -1.0 a 1.0, hecho que resulta atractivo para la mayoría de los prácticos. Según veremos, en efecto, la relación entre el coeficiente de correlación y la desviación estándar respecto de la línea de los mínimos cuadrados es muy sencilla, hecho que puede utilizarse para proporcionar una interpretación de r.

Se acaba de indicar que r tiene un límite superior de 1.0. Si todos los puntos se hallan exactamente sobre la recta, r será 1.0 o -1.0, según que la relación sea positiva o negativa. Y si los puntos están dispersados al azar, r será cero. Cuanto mejor sea el ajuste, tanto mayor será la magnitud de r. Es lo que se indica en la figura XVII.7.

Obsérvese que r es una medida de relación *lineal*, ya que es una medida de la bondad de ajuste de la línea de los mínimos cuadrados. El lector no debe caer en el error de suponer que si $r = 0$ (o si $\rho = 0$) no existe relación alguna. En efecto, si no hay relación, síguese que r será aproximadamente cero y habrá una dispersión de puntos al azar. Sin embargo, puede haber una relación perfectamente curvilínea y, con todo, ser r cero, indicando que no se da *recta* alguna que satisfaga los datos. Éste es el caso en la figura XVII.8, por ejemplo. Por lo tanto, si el investigador encuentra una correlación de cero, habrá de precaverse contra la deducción de que no existe relación entre las variables. Por lo regular, la inspección del diagrama de dispersión indicará si hay o no relación de hecho, o si la relación es suficientemente no lineal para producir una correlación de cero. En la mayoría de los problemas sociológicos, las relaciones pueden aproximarse razonablemente por medio de rectas. Sin embargo, esto no significa que no se deba estar bastante alerta contra excepciones eventuales.

Hasta el presente no hemos definido todavía el coeficiente de correlación, pero podemos hacerlo fácilmente en los términos de la fórmula:

[6] La naturaleza exacta de semejante medida se examinará más adelante. De momento podemos señalar simplemente que representa una extensión del concepto de la desviación estándar, en la que la media de las Y ya no se toma como fija, sino que se considera función de X.

$$r = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{[\Sigma(X-\overline{X})^2][\Sigma(Y-\overline{Y})^2]}} = \frac{\Sigma xy}{\sqrt{(\Sigma x^2)(\Sigma y^2)}} \quad \text{(XVII.6)}$$

U oralmente: el coeficiente de correlación es la razón de la covariación a la raíz cuadrada del producto de la variación de X y la variación de Y. Dividiendo el numerador y el denominador entre N y poniendo esta cantidad como N^2 bajo el radical, vemos

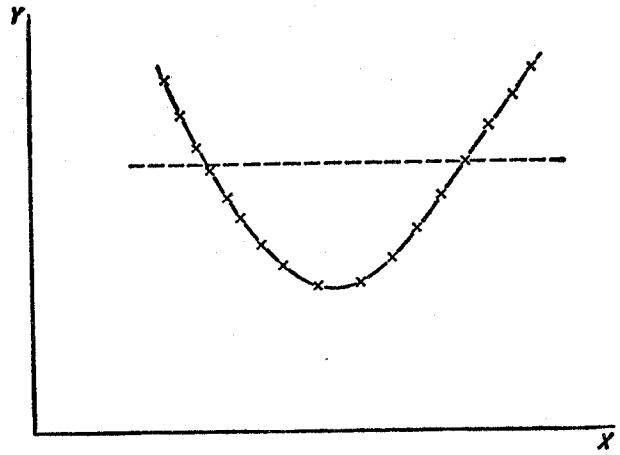


FIG. XVII.8. *Diagrama de dispersión de una relación no lineal perfecta, en que* r = 0.

que r puede también definirse como la razón de la covariancia al producto de las desviaciones estándar de X y Y. La covariancia es la medida de la variación conjunta de X y Y, pero su magnitud depende de la cantidad total de variabilidad de las dos variables. Como quiera que el valor numérico de la covariancia puede ser considerablemente mayor que la unidad, no resulta conveniente emplearlo directamente como medida de asociación. En lugar de ello, estandarizamos dividiendo entre el producto de las dos desviaciones estándar, con lo que obtenemos una medida que varía entre -1.0 y 1.0.

Ya vimos que la covariancia será cero siempre que X y Y no estén relacionadas. Puede demostrarse también fácilmente que el límite superior de r es la unidad. Tomemos, por ejemplo, el caso en que b es positiva y todos los puntos se encuentran exactamente sobre la recta. En tal caso, para cada Y podemos escribir $Y = a + bX$. Y como quiera que el punto $(\overline{X},\overline{Y})$ se encuentra tam-

bién sobre la recta, tenemos $\overline{Y} = a + b\overline{X}$. Por consiguiente, para todos los puntos sobre la recta tenemos:

$$Y - \overline{Y} = (a + bX) - (a + b\overline{X}) = b(X - \overline{X})$$

De donde: $$\Sigma(X - \overline{X})(Y - \overline{Y}) = b\Sigma(X - \overline{X})^2$$

y $$\Sigma(Y - \overline{Y})^2 = b^2\Sigma(X - \overline{X})^2$$

La inspección del numerador y el denominador de r indica ahora que, en estas condiciones, $r = 1.0$. Y en forma análoga, puede demostrarse que si todos los puntos se encuentran exactamente sobre una línea de pendiente negativa, la r resultante será -1.0.

Conviene observar asimismo la relación entre el coeficiente de correlación y las pendientes de las dos ecuaciones de los mínimos cuadrados. Si hacemos que b_{yx} sea la pendiente de la ecuación de mínimos cuadrados estimando la regresión de Y sobre X, y dejamos que b_{xy} indique la pendiente de la estimación de la regresión de X sobre Y, tenemos, por simetría, que:

$$b_{xy} = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{\Sigma(Y - \overline{Y})^2}$$

en donde $$X = a_{xy} + b_{xy}Y$$

Así, pues, r tiene el mismo numerador que las dos b. Si éstas son cero, síguese que r ha de ser también cero y viceversa.

Para sumas de cuadrados en X y Y dadas, el valor de b_{yx} (o de b_{xy}) será proporcional a r. Esto parecería conducir a la conclusión de que la fuerza de la relación sea proporcional a la pendiente de la línea de los mínimos cuadrados. Sin embargo, esto sólo será así si el denominador permanece fijo. Así, pues, b es una función no sólo de la fuerza de la relación, sino también de las desviaciones estándar.[7] Si hay bastante variabilidad en X, en relación con Y, el valor de b será relativamente pequeño, indicando que se requiere un gran cambio de X para producir un cambio moderado de Y. Como lo veremos después, los valores numéricos de las b dependen, por consiguiente, de la magnitud de las unidades de medida.

El valor de r se ha estandarizado de modo que sea hasta cierto punto independiente de las magnitudes relativas de las desviaciones estándar en X y Y. Sería en efecto desdichado que no fuera así, ya que difícilmente deseamos una medida que variara

[7] Excepto en los casos en que ello pudiera dar lugar a confusión, seguiremos sirviéndonos de b sin subíndice para representar b_{yx}.

según que escogiéramos como unidad monetaria dólares o centavos. Se observará en las fórmulas de r y las b que r^2 puede expresarse en términos de estas últimas. Así, pues:

$$r^2 = b_{yx}b_{xy} = \frac{[\Sigma xy]^2}{\Sigma x^2 \Sigma y^2} \qquad \text{(XVII.7)}$$

El lector hará bien en verificar que cuando r es 1.0 (o -1.0), $b_{yx} = 1/b_{xy}$, lo que significa que las dos ecuaciones de mínimos cuadrados coinciden. Por lo regular, a medida que r se acerca a cero, el ángulo entre las dos líneas se va haciendo cada vez mayor, hasta que, $r = 0$, las líneas se hacen perpendiculares.

Finalmente, podemos introducir una fórmula de cálculo para r que comporta las cinco sumas previamente obtenidas en conexión con los cálculos de a y b. La fórmula es:

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}} \qquad \text{(XVII.8)}$$

El numerador, por supuesto, ha sido ya calculado, lo mismo que una parte del denominador. Así, pues, la correlación entre el porcentaje de negros y el índice de discriminación es:

$$r = \frac{13(43\,943.32) - (62.88)(8\,557)}{\sqrt{[13(432.2768) - (62.88)^2][13(6\,192\,505) - (8\,557)^2]}}$$

$$= \frac{33\,199}{110\,120} = .301$$

Conviene observar que se pueden adicionar valores tanto a X como a Y, o sustraerlos, sin afectar el valor del coeficiente de correlación. De forma análoga, r no se verá afectado por un cambio de escala en cualquiera de las variables. Esto equivale a decir, de hecho, que la correlación entre el ingreso y la educación es la misma, ya sea que se mida el ingreso en dólares o en centavos. Sin embargo, aunque el coeficiente de correlación sea invariante en transformaciones de esta clase, la ecuación de los mínimos cuadrados, en cambio, no lo es. En efecto, la adición o sustracción de valores afecta el valor numérico de a. Y un cambio de escala afecta la pendiente de la línea. Así, por ejemplo, si cada X se divide entre 10 manteniendo a la Y fija, la b resultante se verá multiplicada por 10. El lector hará bien en verificar que estas propiedades se mantienen, examinando las fórmulas de r, a y b. Estos hechos pueden utilizarse con objeto de simpli-

ficar los cálculos. Así, por ejemplo, si X comporta un número muy grande o un decimal muy pequeño, un cambio de escala puede reducir el riesgo de errores de cálculo. O bien, si la variable X consta de valores tales como 1 207, 1 409, 1 949 y 1 568, se recomendará probablemente sustraer 1 000 de cada marca. Algunas rutinas de cálculo requieren que todos los valores sean positivos.

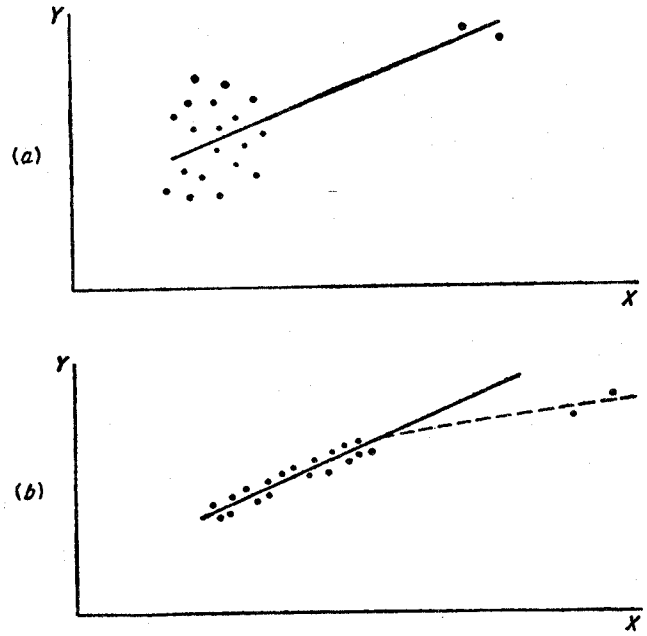


FIG. XVII.9. *Diagramas de dispersión que muestran los efectos posibles de valores extremos de* X.

Por lo tanto, al calcular r puede resultar necesario añadir a cada valor un número ligeramente superior a la marca negativa mayor.

Hay que tener presente, en este punto, otro hecho relativo a la correlación. Y es que, como quiera que esta medida comporta variancias y covariancias a la vez, se ve sumamente afectada por unos pocos valores extremos de cualquiera de las dos variables. Por otra parte, la magnitud de r depende del grado de variabilidad general de la variable independiente. Es lo que ilustra la figura XVII.9. En la figura XVII.9*a*, el efecto de uno o dos valores extremos produce una correlación moderadamente alta cuando no se da ninguna en los casos restantes. En la figura XVII.9*b*, tenemos una relación lineal moderadamente elevada, excepto en cuanto al hecho de que los casos extremos no quedan en línea recta con los demás. En este último caso tenemos probablemente un ejemplo de relación no lineal. El diagrama de dispersión resultará siempre útil para indicar la naturaleza de la situación en un problema determinado. Veamos ahora lo que puede hacerse cuando se presenta una u otra de estas situaciones.

La figura XVII.9*a* ilustra el punto anteriormente señalado de que la magnitud del coeficiente de correlación depende del mar-

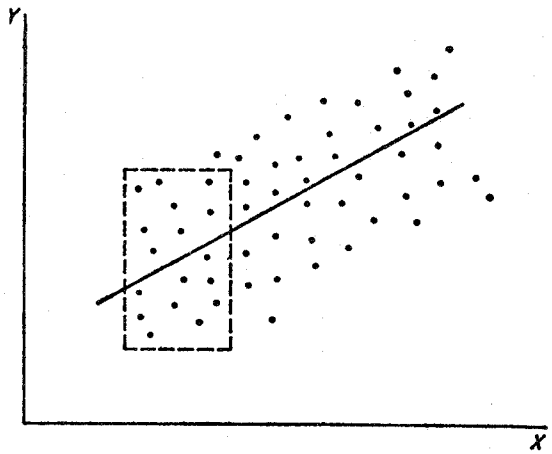


FIG. XVII.10. *Diagrama de dispersión que no muestra relación alguna dentro de un recorrido limitado de variación de* X, *pero con relación positiva sobre el recorrido total.*

gen de variabilidad de ambas variables. Si hubiera habido un número mayor de casos extremos, la distribución resultante habría podido ser como en la figura XVII.10. En este caso, la correlación conjunta podría ser alta, pero en el interior de cualquier recorrido limitado de las X la correlación puede ser vecina de cero. Esto indica de hecho que hay insuficiente variabilidad de X en el interior de dicho recorrido limitado para contrarrestar los efectos de las numerosas variables incontroladas. En realidad, X está siendo mantenida casi constante. Por consiguiente, si el diagrama de dispersión resulta ser semejante al de la figura XVII.9*a*, habría que tratar de extender el recorrido de variabilidad de X hallando más casos extremos.

Si la extensión del recorrido de variabilidad no resulta prácticamente posible, o si el interés del investigador se centra ante todo en casos menos extremos, será tal vez más razonable prescindir totalmente en el análisis de los casos extremos. Así, por ejemplo, supongamos que X es el tamaño de las ciudades y que la ciudad de Nueva York figura en la muestra. A menos que haya un gran número de ciudades de tamaño correspondiente, y no las hay, puede resultar necesario limitar la atención a ciudades de

menos de 500 000 habitantes. En algunos casos podrá parecer indicado calcular r tanto con los casos extremos como sin ellos. Es obvio que la decisión dependerá de la naturaleza del problema y del interés del sociólogo. El lector ha de percatarse bien del hecho de que una o dos marcas extremas pueden eventualmente ejercer un efecto muy pronunciado sobre el tamaño de r, hecho que en alguna forma debe tenerse siempre en cuenta. De ahí que el recorrido de variabilidad debiera consignarse juntamente con los coeficientes de correlación. Esto constituye otra ilustración del punto importante relativo a que una simple medida de resumen, por muy superior que sea respecto de otras, puede ser a menudo desorientadora.

Si los datos se presentan como en la figura XVII.9*b*, sospecharemos, por supuesto, que no existe linealidad. Aquí también, pues, habría que obtener, de ser posible, más casos extremos. Si éstos son sólo uno o dos, resultará tal vez preferible excluirlos del análisis. Las situaciones de esta índole ilustran el hecho de que, al interior de cierto recorrido una relación de variación puede ser aproximadamente lineal, resultando en cambio inapropiada si se extiende el modelo lineal. De ahí, pues, que se imponga prudencia en cuanto a generalizar más allá de los límites de los datos. Un enunciado por el estilo de "dentro los límites de ———y——— la relación resulta ser aproximadamente lineal" será más apropiado.

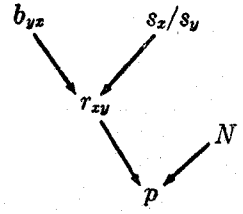
Comparación entre correlaciones y declives. Las observaciones anteriores acerca de la sensibilidad de los coeficientes de correlación ante las diferencias en la cantidad de variación de X, relativa a la dispersión producida por factores extraños, apunta uno de los problemas fundamentales con *cualquier* medida del grado de asociación. Nuestra atención debe estar centrada en la naturaleza de la ley que relaciona X y Y, de si la relación es o no es lineal, y, si lo es, en la magnitud del declive. Al comparar los resultados de dos estudios o de varias submuestras, debemos reconocer que es perfectamente posible obtener diferencias sustanciales entre los coeficientes de correlación, aun cuando se apliquen las mismas leyes (medidas por los declives). Es decir, que las r pueden diferir aunque no los declives, lo que puede ser debido únicamente a diferencias en la cantidad de variación en la variable independiente X, o a diferencias en la amplitud con que han sido sometidos a control otros factores extraños que producen variaciones aleatorias en Y. Como veremos al tratar del análisis de la covariancia, al buscar la interacción estamos en efecto buscando una diferencia entre declives, y no correlaciones. En el capítulo siguiente estudiaremos brevemente las pruebas para encontrar diferencias entre correlaciones, pero el lector debe estar prevenido acerca del peligro de que tales diferencias, una vez encontradas, puedan ser fácilmente mal interpretadas.

Puede ser útil concebir un coeficiente de correlación r_{xy} como función de dos tipos variables, con el declive b_{yx} y un factor s_x/s_y abarcando la razón de las dos desviaciones estándar que se aplican a la muestra o submuestra particular que nos ocupa. Así:

$$r_{xy} = b_{yx}\,(s_x/s_y)$$

El valor numérico de b_{yx} es, por supuesto, determinado no sólo por la ley que une a X con Y, sino también por la elección que el investigador hace entre las unidades de medida. El factor s_x/s_y es también una función de tales unidades, las que son por supuesto conocidas antes que los datos de la población o la muestra. Pero la razón s_x/s_y será también única para cada muestra (y σ_x/σ_y para cada población), y se utiliza para obtener la medida estandarizada r_{xy}. Un coeficiente de correlación tiene la ventaja de ser estandarizado, independizándolo así de la elección que se haga de unidades de medida, pero lamentablemente tiene que ser estandarizado en función de algo que resulta ser una cantidad no invariable en relación con muestras o poblaciones. Esta circunstancia debe ser claramente entendida, debiendo ser señalados *siempre* los declives no estandarizados, de modo que las réplicas no resulten desorientadoras a este respecto.

Planteando el asunto en forma algo diferente, podemos reconocer que en la inferencia y estimación estadísticas se da una jerarquía de metas científicas. Probamos buscando primero la significancia, para decidir si se ha encontrado una relación que no pueda ser fácilmente explicable por mecanismos casuales. Observamos a este respecto que el nivel de probabilidad o significación es función del grado de relación y del tamaño de la muestra. Si ésta es muy grande podremos obtener un pequeño nivel de probabilidad, incluso con una relación muy débil y tal vez sin importancia práctica. Pero habiendo encontrado al menos una relación moderadamente fuerte, se nos plantea de nuevo una tarea más importante, a saber: la de estimar la naturaleza de tal relación, medida por un coeficiente de regresión en el caso lineal. Cuando las correlaciones son moderadamente fuertes, en lugar de comparar estas r directamente, estimamos los declives, y los comparamos en nuestras pruebas de interacción. El proceso puede presentarse diagramáticamente así:



donde la dirección de las flechas representa el "curso causal" (por ejemplo: probabilidades influidas por magnitudes de relaciones y tamaños de muestras), lo que va frecuentemente en dirección opuesta a la que siguen los pasos del procedimiento empleado en un análisis estadístico. El diagrama indica que p es una función de dos variables, una de las cuales (el tamaño de la muestra) no es de interés inherente, y que la correlación r_{xy} es asimismo una función de dos factores, uno de los cuales (s_x/s_y), no es de interés. Nuestro objetivo consiste en llevar el análisis hacia arriba en el diagrama hasta la estimación de los coeficientes de regresión, en lugar de detenernos en los niveles de probabilidad, o formulando declaraciones en relación con los coeficientes de correlación.

Resulta que en cuantas ocasiones se manejan medidas ordinales de asociación, tales como las que se verán en el capítulo siguiente, desaparece la distinción entre declives y medidas de asociación. En el caso de dicotomías, sin embargo, puede demostrarse que si se sigue la regla de colocar la variable independiente al través de la parte alta del cuadro, y se computan las proporciones (o porcentajes) de modo que sumen 1.00 (o 100) hacia abajo, comparando a continuación de izquierda a derecha, la diferencia de proporciones resultante puede ser considerada como un caso especial del declive b_{yx}, en tanto que ϕ pasa a ser un caso especial de r_{xy}. Si se computan las proporciones en la otra dirección, la diferencia de proporciones pasa a ser un caso especial de b_{xy}, de modo que tendremos una justificación más para seguir la regla empírica previamente sugerida. Pueden obtenerse estos resultados por el simple procedimiento de asignar puntuaciones de 0 y 1 tanto a X como a Y, utilizando a continuación las fórmulas básicas para el cálculo de r_{xy} y b_{yx}.

* *Cálculos a partir de datos agrupados.* Si el número de casos es grande o si no se dispone de una calculadora moderna, el cálculo de los coeficientes de correlación puede resultar extremadamente laborioso. En tal caso será tal vez más indicado servirse de datos agrupados, aun a riesgo de introducir eventualmente algunas imprecisiones. En principio, estos cálculos de datos agrupados no son más que aplicaciones abreviadas de los procedimientos empleados para obtener la media y la desviación estándar. Tenemos ahora dos variables que han de clasificarse cruzadamente como en el cuadro XVII.2. Hemos de anticipar una media para cada variable, tomando desviaciones graduales de cada una de las medias y sirviéndonos de factores de corrección en cada caso. Además, necesitaremos un término de producto cruzado equivalente a Σxy. Como que las desviaciones tanto de X como de Y se tomarán de las medias *adivinadas* respectivas, necesitamos servirnos de un factor de corrección a sustraer del término del producto cruzado apreciado. Podemos modificar así las fórmulas

de cálculo de r y b de modo que se tenga en cuenta que nos hemos servido de medias adivinadas en lugar de las correctas.

Se recordará que una de las fórmulas de s sirviéndose de datos agrupados era (dejando de lado los subíndices):

$$s = \frac{i}{N}\sqrt{N\Sigma fd'^2 - (\Sigma fd')^2}$$

Como quiera que tenemos ahora dos variables, X y Y, nos serviremos de subíndices con objeto de distinguir las frecuencias y las desviaciones graduales de X (esto es, f_x y d'_x) de las de Y (o sea, f_y y d'_y). Al calcular el término del producto cruzado, necesitamos obtener también las frecuencias f_{xy} de cada subcasilla. Estas últimas serán por lo regular más pequeñas que f_x o f_y. Así, pues, si bien hay 24 casos en la categoría de 40.0 a 49.9 para la variable X y 30 casos en la categoría de 15.0 a 19.9 de Y, sólo hay 6 casos en la subcasilla correspondiente a ambas categorías. El lector ha de convencerse por sí mismo de que la fórmula de cálculo de r (ecuación XVII.8) puede modificarse como sigue:

$$r = \frac{N\Sigma f_{xy}d'_x d'_y - (\Sigma f_x d'_x)(\Sigma f_y d'_y)}{\sqrt{[N\Sigma f_x d'^2_x - (\Sigma f_x d'_x)^2][N\Sigma f_y d'^2_y - (\Sigma f_y d'_y)^2]}} \quad \text{(XVII.9)}$$

Y en forma análoga, la fórmula de b se convierte en:

$$b = \frac{N\Sigma f_{xy}d'_x d'_y - (\Sigma f_x d'_x)(\Sigma f_y d'_y)}{N\Sigma f_x d'^2_x - (\Sigma f_x d'_x)^2}\frac{i_y}{i_x} \quad \text{(XVII.10)}$$

en donde i_y e i_x representan las amplitudes de intervalos de Y y X respectivamente. El valor de a puede calcularse ahora a partir de la ecuación:

$$a = \frac{\Sigma Y - b\Sigma X}{N} = \overline{Y} - b\overline{X}$$

en donde $\overline{X}$ y $\overline{Y}$ pueden obtenerse sirviéndonos de la fórmula usual de los datos agrupados.

Calculemos ahora los valores en esos coeficientes en relación con los datos de 150 distritos del Sur consignados en el cuadro XVII.2. Tomaremos como variable dependiente Y, o sea el porcentaje de mujeres de la clase trabajadora, siendo la variable independiente el porcentaje de la población clasificada como granjas rurales. Convendrá servirse de una fórmula de cálculo como la que se da en el cuadro XVII.3. En ésta, los límites de

las clases y los puntos medios se indican horizontalmente en la parte superior (para Y) y de arriba abajo, a mano izquierda, para X. Obsérvese el área cerrada en el interior del cuadro. Se verá que hay tres números en cada subcasilla. En cada casilla, el número de arriba representa el número de casos de la subcasilla, tal como se da en el cuadro XVII.2. Los números restantes de la

CUADRO XVII.2. *Datos clasificados cruzados para obtener correlaciones de datos agrupados*

Porcentaje de granjas rurales, X	*Porcentaje de mujeres de la clase trabajadora, Y*							*Totales*
	10.0–14.9	15.0–19.9	20.0–24.9	25.0–29.9	30.0–34.9	35.0–39.9	40.0–44.9	
0.0– 9.9	0	0	0	1	8	4	0	13
10.0–19.9	1	2	0	2	4	1	3	13
20.0–29.9	2	5	1	2	3	3	0	16
30.0–39.9	2	0	5	5	7	3	0	22
40.0–49.9	4	6	6	7	1	0	0	24
50.0–59.9	3	10	9	6	2	0	0	30
60.0–69.9	2	4	3	7	4	0	0	20
70.0–79.9	2	3	4	1	0	0	0	10
80.0–89.9	1	0	1	0	0	0	0	2
Totales	17	30	29	31	29	11	3	150

FUENTE: Censo de los Estados Unidos de 1950.

subcasilla se emplean para calcular el término del producto cruzado. La cifra central de cada subcasilla representa el producto de las desviaciones graduales $d'_x d'_y$. Así, por ejemplo, en la subcasilla más baja de la izquierda (correspondiente a las categorías de 80.0 a 89.9 y de 10.0 a 14.9), la cifra − 12 es el producto de 4 por −3. En otros términos: la categoría de 80.0 a 89.9 se halla 4 desviaciones graduales *por encima* de la media anticipada de X, y la categoría de 10.0 a 14.9 se encuentra 3 desviaciones graduales *por debajo* de la media anticipada de Y. Finalmente, el número inferior en cada subcasilla representa el producto de los dos números que tiene arriba y puede por consiguiente representarse simbólicamente como $f_{xy} d'_x d'_y$. Por lo tanto, la suma de estas cifras inferiores de todas las subcasillas nos da el término del producto cruzado, sin corrección de los errores introducidos sirviéndose de medias estimadas. Esta suma se empleará en el primer término del numerador de r; es numéricamente igual a − 200, y se ha dispuesto en el ángulo inferior derecho del cuadro.

Las cantidades restantes necesitadas en el cálculo de r y b pueden obtenerse en la forma usual. Las cuatro últimas columnas

CUADRO XVII.3. *Cálculos de la correlación de datos agrupados* *

Límites de clase	Y	10.0-14.9	15.0-19.9	20.0-24.9	25.0-29.9	30.0-34.9	35.0-39.9	40.0-44.9	f_x	d'_x	$f_x d'_x$	$f_x (d'_x)^2$
X	Puntos medios	12.45	17.45	22.45	27.45	32.45	37.45	42.45				
0.0-9.9	4.95				1 0 0	8 −4 −32	4 −8 −32		13	−4	−52	208
10.0-19.9	14.95	1 +9 9	2 +6 12		2 0 0	4 −3 −12	1 −6 −6	3 −9 −27	13	−3	−39	117
20.0-29.9	24.95	2 +6 12	5 +4 20	1 +2 2	2 0 0	3 −2 −6	3 −4 −12		16	−2	−32	64
30.0-39.9	34.95	2 +3 6		5 +1 5	5 0 0	7 −1 −7	3 −2 −6		22	−1	−22	22
40.0-49.9	44.95	4 0 0	6 0 0	6 0 0	7 0 0	1 0 0			24	0	0	0
50.0-59.9	54.95	3 −3 −9	10 −2 −20	9 −1 −9	6 0 0	2 +1 2			30	1	30	30
60.0-69.9	64.95	2 −6 −12	4 −4 −16	3 −2 −6	7 0 0	4 +2 8			20	2	40	80
70.0-79.9	74.95	2 −9 −18	3 −6 −18	4 −3 −12	1 0 0				10	3	30	90
80.0-89.9	84.95	1 −12 −12		1 −4 −4					2	4	8	32
f_y		17	30	29	31	29	11	3	N = 150		−37	643
d'_y		−3	−2	−1	0	1	2	3				
$f_y d'_y$		−51	−60	−29	0	29	22	9	−80		$\Sigma f_{xy} d'_x d'_y = -200$	
$f_y (d'_y)^2$		153	120	29	0	29	44	27	402			

* Esta forma de cálculo se ha tomado, con ligeras adaptaciones, de [1], cuadro XIX.4 de la p. 476, con la amable autorización del editor.

del cuadro se emplean para obtener f_x, d'_x, $f_x d'_x$ y $f_x(d'_x)^2$, las sumas de las dos últimas de estas cantidades utilizándose directamente en la fórmula de r. Obsérvese que al calcular los valores numéricos de estas cuatro columnas prescindimos por completo de los valores de Y. Así, pues, si dejamos totalmente de lado el área encerrada, tenemos exactamente la misma clase de tabla de la que nos servimos al calcular la media y la desviación estándar de datos agrupados. Y en forma análoga, las cuatro hileras inferiores pueden emplearse para obtener sumas correspondientes en relación con la variable Y. Todas las cantidades necesitadas en las fórmulas de r y b pueden ponerse ahora en las casillas inferiores de la derecha de la tabla mayor.

Obtenemos ahora los valores de r y b como sigue:

$$r = \frac{150(-200) - (-37)(-80)}{\sqrt{[150(643) - (-37)^2][150(402) - (-80)^2]}} = \frac{-32\,960}{71\,590} = -.460$$

$$b = \frac{150(-200) - (-37)(-80)}{150(643) - (-37)^2} \frac{5.0}{10.0} = \frac{-32\,960}{95\,081} \frac{1}{2} = -.1733$$

Como quiera que los valores de $\overline{X}$ y $\overline{Y}$ son 42.48 y 24.78, respectivamente, obtenemos:

$$a = \overline{Y} - b\overline{X} = 24.78 - (-.1733)(42.48) = 32.14$$

y la ecuación de los mínimos cuadrados puede escribirse como:

$$Y_p = 32.14 - .1733X$$

Interpretación del coeficiente de correlación. Con objeto de obtener una interpretación de r que tenga sentido cuando r no es ni cero ni 1.0, volvamos al concepto de variabilidad a propósito de la ecuación de regresión. Hemos definido la variancia respecto de la media de Y como:

$$\sigma_y^2 = \frac{\Sigma(Y - \mu_y)^2}{M}$$

en donde M representa la magnitud de la población (frente al tamaño de la muestra N) y donde nos servimos de los subíndices para recalcar el hecho de que tenemos ahora dos variables que han de distinguirse. Así, pues, el concepto corriente de la variancia comporta desviaciones respecto de una medida *fija* de tendencia central, o sea la media conjunta. Pero podemos obtener

también la media de las Y para una X fija, y estamos suponiendo que estos valores varían con X de manera que produzcan una regresión lineal. Podemos generalizar en esta forma el concepto de la media, obteniendo una especie de media condicional de Y para una X dada, que podemos simbolizar como $\mu_{y|x}$ o como $E(Y|X)$.

Si generalizamos el concepto de variancia en forma similar, podemos obtener una medida de dispersión respecto de la ecuación de regresión tal como:

$$\sigma^2_{y|x} = \frac{\Sigma(Y - \mu_{y|x})^2}{M} \qquad \text{(XVII.11)}$$

en donde el símbolo $\sigma^2_{y|x}$ se emplea para señalar el hecho de que la magnitud de la variabilidad respecto de la ecuación de regresión, lo mismo que la media de Y, depende del valor de X. En otros términos: para cada X se dan tanto una media de las Y como una variancia respecto de dicha media. La cantidad de dispersión alrededor de la línea no necesita ser siempre la misma para cada X, pese a que vamos a *suponer* la propiedad de homoscedasticidad o de variancias iguales.

Tenemos ahora dos medidas de variabilidad para Y. La primera mide la dispersión alrededor del valor de Y, la gran media μ_y, que sería el mejor valor anticipado de Y si no se conociera X. En otros términos: si se nos pidiera anticipar Y no conociendo X, la mejor anticipación sería μ_y (o $\overline{Y}$, si sólo se dispusiera de los datos de la muestra). En cambio, si conociéramos X, anticiparíamos el valor correspondiente de Y que se sitúa en la ecuación de regresión. A menos que no existiera relación entre X y Y, el conocimiento de X nos ayudará a predecir el valor de Y. Si la relación fuera perfecta, podríamos predecir Y exactamente, ya que todos los puntos quedarían exactamente sobre la línea. Por lo regular, no estaremos en condiciones de hacerlo así, pero, como quiera que estamos suponiendo una distribución normal de las Y y una desviación estándar $\sigma_{y|x}$ fija, podemos emitir enunciados de probabilidad acerca de los riesgos y de la magnitud del error. Y lo que es más importante todavía desde el punto de vista de nuestros propósitos, podemos comparar las dos desviaciones estándar (o variancias) y obtener una medida acerca de en qué proporción se ha mejorado la anticipación por el conocimiento de X. Al proceder en esta forma, podemos servirnos de procedimientos con los que estamos ya familiarizados a partir del análisis de la variancia.

En dicho análisis, en efecto, tomamos la variación total o suma de cuadrados y descompusimos dicha cantidad en porciones explicadas e inexplicadas. Vamos a servirnos ahora exacta-

mente del mismo procedimiento, obteniendo casi a manera de producto accesorio los valores de $\sigma^2_{y|x}$ y r^2. Con lo que estaremos en condiciones de dar una interpretación lógica del coeficiente de correlación. Primero, podemos expresar las desviaciones de cada Y respecto de $\overline{Y}$ como suma de dos cantidades $(Y - Y_p) + (Y_p - \overline{Y})$ (véase la figura XVII.11). La primera de estas cantida-

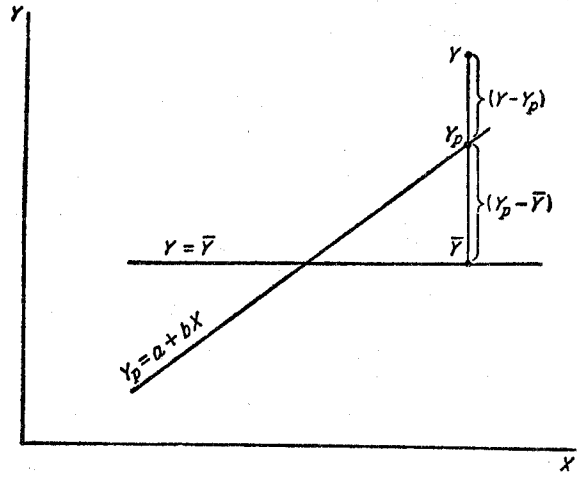


FIG. XVII.11. *Representación geométrica que muestra las desviaciones respecto de la media* $\overline{Y}$ *como una suma de desviaciones respecto de la recta de mínimos cuadrados y desviaciones de la recta de mínimos cuadrados respecto de la* $\overline{Y}$.

des representa la desviación del valor de Y respecto de la línea de los mínimos cuadrados e indica la cantidad de error que se comete cuando se emplea Y_p para predecir Y. La segunda expresión, en cambio, indica la desviación de la línea de mínimos cuadrados (para una X dada) respecto de $\overline{Y}$. En la mayoría de los casos, esta cantidad representará el monto en que se reduce el error al conocer Y_p. Si elevamos al cuadrado ahora ambos miembros de la ecuación y sumamos luego todos los casos, obtenemos:

$$\Sigma(Y - \overline{Y})^2 = \Sigma(Y - Y_p)^2 + 2\Sigma(Y - Y_p)(Y_p - \overline{Y}) + \Sigma(Y_p - \overline{Y})^2$$

Afortunadamente, el término central vuelve a desaparecer, y nos quedamos con:

$$\Sigma(Y - \overline{Y})^2 = \Sigma(Y - Y_p)^2 + \Sigma(Y_p - \overline{Y})^2 \quad \text{(XVII.12)}$$

$$\text{SC total} = \text{SC inexplicada} + \text{SC explicada}$$

La primera cantidad de la derecha de la ecuación representa la suma de los cuadrados de las desviaciones de los valores reales de Y respecto de la línea de los mínimos cuadrados. Esta cantidad es inexplicada, ya que indica la magnitud del error en la predicción. Y la cantidad restante indica lo que hemos ganado al servirnos de Y_p con preferencia a $\overline{Y}$, pudiendo designarse como la suma de cuadrados explicada. Por *explicada* no entendemos, por supuesto, una explicación causal, sino simplemente una asociación entre las dos variables. Consideremos ahora más de cerca cada una de estas cantidades.

Si tomamos una suma de cuadrados inexplicada y dividimos entre el número total de casos, obtenemos la variancia de la muestra $s^2_{y|x}$ respecto de la línea de los mínimos cuadrados. O sea:

$$s^2_{y|x} = \frac{\Sigma(Y - Y_p)^2}{N} \quad \text{(XVII.13)}$$

Si deseamos obtener una estimación insesgada de la variancia de la población $\sigma^2_{y|x}$ respecto de la regresión real, hemos de dividir no entre N sino entre los grados apropiados de libertad. En este caso hemos perdido 2 grados de libertad al calcular a y b como estimaciones de α y β. Por consiguiente, si deseamos estimar $\sigma^2_{y|x}$ nos serviremos de:

$$\hat{\sigma}^2_{y|x} = \frac{\Sigma(Y - Y_p)^2}{N - 2} \quad \text{(XVII.14)}$$

En esta forma, la suma de cuadrados inexplicada puede convertirse fácilmente en una estimación de la variancia respecto de la ecuación de regresión. El lector hará bien en convencerse por sí mismo de que lo que hemos hecho es directamente paralelo a nuestro tratamiento anterior del análisis de la variancia. La variabilidad respecto de la ecuación de mínimos cuadrados ha sustituido la noción de variabilidad en el *interior* de las categorías de X.

Volviendo ahora a la suma de cuadrados explicada $\Sigma(Y_p - \overline{Y})^2$, podemos mostrar fácilmente que esta cantidad es equivalente a $r^2[\Sigma(Y - \overline{Y})^2]$, o $r^2\Sigma y^2$. Como quiera que $Y_p = a + bX$ y $\overline{Y} = a + b\overline{X}$, tenemos:

$$(Y_p - \overline{Y}) = b(X - \overline{X})$$

Por consiguiente:

$$\Sigma(Y_p - \overline{Y})^2 = b^2\Sigma(X - \overline{X})^2 = b^2\Sigma x^2$$

$$= \frac{(\Sigma xy)^2}{(\Sigma x^2)^2}(\Sigma x^2) = \frac{(\Sigma xy)^2}{\Sigma x^2}$$

$$= \frac{(\Sigma xy)^2}{\Sigma x^2 \Sigma y^2}(\Sigma y^2) = r^2\Sigma y^2$$

$$= r^2\Sigma(Y - \overline{Y})^2$$

Hemos demostrado así que:

$$r^2 = \frac{\Sigma(Y_p - \overline{Y})^2}{\Sigma(Y - \overline{Y})^2} = \frac{\text{SC explicada}}{\text{SC total}}$$

Por medio de un razonamiento similar pudimos haber demostrado que r^2 representa la razón de la variación explicada en X a la variación total en X. Por lo tanto, el cuadrado del coeficiente de correlación puede interpretarse como la proporción de variación total en una de las variables explicada por la otra. La cantidad de $\sqrt{1 - r^2}$, designada a menudo como *coeficiente de alienación*, representa la raíz cuadrada de la proporción de la suma total de cuadrados que permanece sin explicar por la variable independiente.

Cabe observar que no se da interpretación directa y simple alguna de la r misma. De hecho, es posible dejarse desorientar por los valores de r, ya que estos valores serán numéricamente mayores que los de r^2 (a menos que r sea 0 o $\pm$ 1.0). Así, por ejemplo, podría parecer que una r de .5 sea la mitad de buena que una correlación perfecta, en tanto que vemos que, en este caso, sólo explicamos un 25 por ciento de la variación. Una correlación de .7 indica que algo menos de la mitad de la variación resulta explicada. Vemos asimismo que correlaciones de .3 o menores significan que sólo una fracción muy pequeña de la variación es explicada. El cuadro XVII.4 indica las relaciones entre las diversas cantidades.

Como quiera que $1 - r^2$ representa la proporción de variación inexplicada, tenemos:

$$(1 - r^2)[\Sigma(Y - \overline{Y})^2] = \Sigma(Y - Y_p)^2$$

Por consiguiente:

$$(1 - r^2)\frac{\Sigma(Y - \overline{Y})^2}{N} = \frac{\Sigma(Y - Y_p)^2}{N}$$

o bien:

$$(1 - r^2)s_y^2 = s_{y|x}^2$$

De donde:

$$s_{y|x} = \sqrt{1 - r^2}\, s_y$$

Este resultado nos proporciona una indicación acerca de en qué medida podemos reducir la desviación estándar conociendo X.

CUADRO XVII.4. *Relaciones numéricas entre* r, r^2, $1 - r^2$ *y* $\sqrt{1 - r^2}$

r	r^2	$1 - r^2$	$\sqrt{1 - r^2}$
.90	.81	.19	.44
.80	.64	.36	.60
.70	.49	.51	.71
.60	.36	.64	.80
.50	.25	.75	.87
.40	.16	.84	.92
.30	.09	.91	.95
.20	.04	.96	.98
.10	.01	.99	.995

(Véase la última columna del cuadro XVII.4.) Si r es cero, las dos desviaciones estándar son iguales. Este hecho es obvio, por supuesto, si nos percatamos de que la línea de los mínimos cuadrados será en tal caso una recta horizontal de ecuación $Y = \overline{Y}$. Si r^2 es igual a la unidad, $s_{y|x}$ será cero, por supuesto, ya que todos los puntos quedarán exactamente sobre la recta. Del cuadro XVII.4 se desprende que la magnitud de r ha de ser grande para que obtengamos una reducción sustancial de las desviaciones estándar. Para una r de .80, la desviación estándar respecto de la línea de los mínimos cuadrados es de .60 de la desviación estándar corriente; pero, con una r de .40, vemos que no hemos ganado mucho en cuanto a apreciar Y a partir de X.

GLOSARIO

Distribución normal bivariada
Coeficiente de alienación
Coeficiente de correlación
Covariancia
Intercepción
Ecuación de los mínimos cuadrados
Regresión de Y sobre X
Declive.

Ejercicios

1. Los siguientes datos relativos a 29 ciudades de 100 mil o más habitantes de regiones fuera del Sur están tomados del estudio de R. C. Angell sobre la integración moral de las ciudades norteamericanas. El índice de integración moral se ha derivado combinando los índices de tasas de criminalidad con los de la labor de mejoramiento. La heterogeneidad se midió en términos de los números relativos de los no blancos y los blancos nacidos en el extranjero contenidos en la población. Y se calculó asimismo, a título de segunda variable independiente, un índice de movilidad, que mide los números relativos de las personas que se establecen o dejan la ciudad.

Ciudad	*Índice de integración*	*Índice de heterogeneidad*	*Índice de movilidad*
Rochester	19.0	20.6	15.0
Syracuse	17.0	15.6	20.2
Worcester	16.4	22.1	13.6
Erie	16.2	14.0	14.8
Milwaukee	15.8	17.4	17.6
Bridgeport	15.3	27.9	17.5
Buffalo	15.2	22.3	14.7
Dayton	14.3	23.7	23.8
Reading	14.2	10.6	19.4
Des Moines	14.1	12.7	31.9
Cleveland	14.0	39.7	18.6
Denver	13.9	13.0	34.5
Peoria	13.8	10.7	35.1
Wichita	13.6	11.9	42.7
Trenton	13.0	32.5	15.8
Grand Rapids	12.8	15.7	24.2
Toledo	12.7	19.2	21.6
San Diego	12.5	15.9	49.8
Baltimore	12.0	45.8	12.1
South Bend	11.8	17.9	27.4
Akron	11.3	20.4	22.1
Detroit	11.1	38.3	19.5
Tacoma	10.9	17.8	31.2
Flint	9.8	19.3	32.2
Spokane	9.6	12.3	38.9
Seattle	9.0	23.9	34.2
Indianapolis	8.8	29.2	23.1
Columbus	8.0	27.4	25.0
Portland (Ore.)	7.2	16.4	35.8

Fuente: R. C. Angell, "The Moral Integration of American Cities" ("La integración moral de las ciudades norteamericanas"), *American Journal of Sociology*, vol. 57, 2ª parte, p. 17, julio de 1951, con la amable autorización del autor y el editor. (Copyright 1951 de la Universidad de Chicago).

a) Trácese un diagrama de dispersión que relacione la integración moral con la heterogeneidad.
b) Calcúlense r, a y b para las mismas variables, y trácese en el diagrama de dispersión la línea de mínimos cuadrados, tomando la integración moral como Y. Respuesta, $r = -.156$; $a = 13.9$; $b = -.049$.
c) ¿De cuánto es la desviación estándar respecto de la línea de los mínimos cuadrados comparada con la desviación estándar respecto de Y?

2. Con objeto de resolver los ejercicios del capítulo XIX, se necesitará obtener las correlaciones entre la integración moral y la movilidad, así como entre la heterogeneidad y la movilidad. Calcúlense las dos r. Respuesta, $r = -.456$; $r = -.513$.

3. Agrúpense los índices de integración moral y heterogeneidad en intervalos y calcúlense r, a y b sirviéndose de las fórmulas de datos agrupados. Compárense los resultados con los datos sin agrupar.

Bibliografía

1. Blalock, H. M.: *Causal Inferences in Nonexperimental Research.* University of North Carolina Press, Chapel Hill, 1964, caps. 2 y 3.
2. Christ, Carl: *Econometric Models and Methods,* John Wiley & Sons, Inc., Nueva York, 1966, Parte III.
3. Croxton, F. E., y D. J. Cowden: *Applied General Statistics,* 3ª ed., Prentice-Hall, Inc.: Englewood Cliffs, N. J. 1967, caps. 19 y 20.
4. Hagood, M. J., y D. O. Price: *Statistics for Sociologists,* Henry Holt and Company Inc., Nueva York, 1952, cap. 23.
5. Hays, W. L.: *Statistics,* Holt, Rinehart and Winston. Inc., Nueva York, 1963, cap. 15.
6. Johnston, J.: *Econometric Methods,* McGraw-Hill Book Company, Nueva York, 1963, Parte II.
7. McCollough, C., y L. Van Atta: *Introduction to Descriptive Statistics and Correlation,* McGraw-Hill Book Company, Nueva York, 1965, caps. 5-8.
8. Mueller, J. H., K. Schuessler, y H. L. Costner: *Statistical Reasoning in Sociology,* 2ª ed., Houghton Mifflin Company, Boston, 1970, cap. 11.
9. Wallis, W. A., y H. V. Roberts: *Statistics: a New Approach,* The Free Press of Glencoe, Ill., Chicago, 1956, cap. 17.
10. Weinberg, G. H., y J. A. Schumaker: *Statistics: An intuitive Approach,* Wadsworth Publishing Company, Inc., Belmont, Cal., 1962, caps. 16-18.