Capítulo 9

MEDIDAS DE ASOCIACION PARA VARIABLES DE INTERVALO: REGRESION Y CORRELACION

9.1. PLANTEAMIENTO GENERAL

Una vez estudiadas las medidas de asociación para variables nominales y ordinales, pasamos al estudio de las medidas de asociación para variables de intervalo, estudio que presenta aspectos estadísticos y matemáticos inéditos hasta ahora, por lo que llevamos visto en los capítulos precedentes. Al tratar de estudiar el tipo de relación existente entre dos variables de intervalo aparecen dos conceptos que conviene diferenciar desde un principio. Por un lado, se trata de analizar el grado de *correlación* entre las dos variables, lo que nos remite al estudio de la variación conjunta de dos variables, su intensidad y dirección o sentido. Por otro lado, se presenta el problema de la *regresión* o predicción de los resultados en una de las dos variables, conocidos los resultados en la otra.

Al tratarse de variables de intervalo, el concepto de media recobra de nuevo su importancia. Como se recordará del estudio de la estadística univariable, la media aritmética de una variable es una predicción útil, porque la media goza de la propiedad de que la suma algebraica de las desviaciones de cada puntuación en relación a la media es cero. A partir de tales desviaciones se puede saber cuán acertada resulta la predicción, y para ello se suele utilizar la varianza (o su raíz cuadrada, la desviación típica) como medida del grado de dispersión de las puntuaciones alrededor de la media.

De este modo, pues, vemos que se puede predecir la media de una variable y medir los «errores» cometidos en la predicción por medio de la varianza (s^2), y ésta sería, de hecho, la predicción realizada con el mínimo de información.

Para realizar una predicción con mayor información, vamos a tener en cuenta la forma en que las puntuaciones de la variable independiente influyen en la distribución de las puntuaciones de la variable dependiente. Y ahora tenemos que introducir una línea de argumentación diferente a la seguida en el capítulo anterior, cuando estudiamos las me-

didas de asociación entre variables nominales y ordinales. Supongamos que somos capaces de obtener una fórmula que pueda describir la forma en que varía la media de la variable dependiente Y al trasladarnos de un extremo al otro de los valores de la variable independiente X. Mediante dicha fórmula lograríamos describir matemáticamente la naturaleza del tipo de relación entre las dos variables y, al mismo tiempo, nos permitiría «calcular» una estimación de una puntuación individual en la variable dependiente, a partir de la información de su puntuación en la variable independiente. Comparando las puntuaciones resultantes de realizar la predicción mediante la ecuación con las puntuaciones realmente observadas, podemos preguntarnos entonces por el grado de exactitud de la ecuación de predicción. Esto se puede expresar mediante una medida de asociación, llamada *coeficiente de correlación* (para el caso de las variables de intervalo), que expresaría la proporción en que se pueden reducir los errores predictivos mediante la ecuación de predicción, en lugar de utilizar como criterio predictivo la media global de la variable dependiente.

Este es el criterio que vamos a seguir a continuación para desarrollar la medida de asociación llamada coeficiente de correlación lineal de Pearson, que se designa mediante r_{xy}. Con el fin de desarrollar esta idea resulta conveniente comenzar nuestro análisis estudiando el problema de la predicción, ya que la noción de la regresión es, desde un punto de vista lógico y teórico, previa a la de correlación.

9.2. Ecuaciones de regresión lineal

Tal como se ha señalado repetidamente (ver, por ejemplo, Blalock, 1979, pág. 382), el fin último de toda ciencia es el de realizar predicciones. También trata el científico de lograr explicaciones en términos causales, pero las explicaciones, cuando alcanzan un alto grado de perfección, son las que permiten predecir mejor a partir del conocimiento de una información suficiente. Albert Einstein consiguió explicar la actuación de todas las fuerzas que actúan en el sistema solar mediante su teoría de la relatividad. A partir de los conocimientos aportados por la teoría de la relatividad, formalizados en las correspondientes expresiones matemáticas, ha sido posible hasta ahora predecir, entre otras cosas, el movimiento de los planetas y los eclipses solares.

En sociología, al igual que en otras ciencias sociales, también se realizan predicciones, pero, a diferencia de las que se realizan en las ciencias físicas, no suelen ir acompañadas de ninguna precisión matemática. Y ello es debido a que, como ya señaló Homans, en sociología existen muchas teorías, pero ninguna explicación (Homans, 1967, pág. 28). Las teorías sociológicas, en lugar de ser sistemas deductivos de proposiciones empíricas que hagan posible la explicación de las mismas, son en

realidad matrices de definiciones operativas que, cuando establecen relaciones entre variables, lo hacen en términos meramente orientativos, con escaso o nulo poder explicatorio *. Además, al no haber alcanzado la mayoría de las variables sociológicas el nivel de medida de intervalo, los intentos por lograr sistemas deductivos formales se hacen extremadamente difíciles. Con todo, siempre que se disponga de dos variables medidas al nivel de intervalo debemos tratar de definir la función que relaciona a ambas variables, no sólo en términos verbales, sino tratando de especificar la forma y el significado de la misma.

Supongamos que disponemos de diversas observaciones referentes a dos variables de intervalo y tratamos de describir, de la forma más precisa posible, la forma en que varía una variable con la otra. Por ejemplo, se podría afirmar, a la vista de una serie de datos, que, por cada año de escolaridad recibida, los ingresos mensuales esperados se incrementarán en 10.000 pesetas. Si los datos confirman este hecho, podríamos decir que existe una relación lineal entre la variable educación y la variable ingresos. Ahora bien, no siempre el tipo de la relación entre dos variables es tan sencilla como la anterior, apareciendo entonces relaciones curvilíneas. Pero, como aproximación al verdadero tipo de relación, la relación lineal es con frecuencia una buena aproximación.

9.2.1. *Relación entre dos variables estadísticas: Ecuación de una recta*

Naturalmente, la forma más simple y clara de expresar una relación entre variables es a través de una ecuación matemática. Aunque existen relaciones muy complejas que se expresan por medio de ecuaciones matemáticas igualmente complejas, lo cierto es que en sociología, por las razones anteriormente apuntadas, la mayor parte de las relaciones empíricas conocidas entre variables son muy simples y del tipo lineal.

Veamos ahora, a través de unos datos ficticios, la forma en que se construye una ecuación matemática que exprese la relación lineal existente entre dos variables. Supongamos que disponemos de datos de seis individuos referentes a los años de escolaridad que han finalizado cada uno y el nivel de ingresos mensuales que alcanzan:

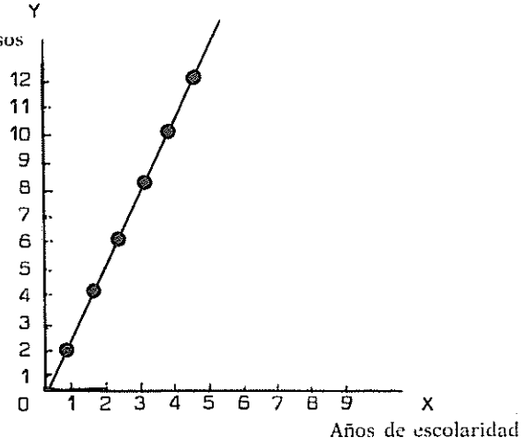
Individuo	*(X) Años de escolaridad*	*(Y) Ingresos (10.000 ptas.)*
A	1	2
B	2	4
C	3	6
D	4	8
E	5	10
F	6	12

* Para un tratamiento más detallado del problema de la explicación en sociología, ver mi trabajo, Manuel García Ferrando: *Sobre el Método*, Madrid, CIS, 1979, especialmente las páginas 143-150.

Estas puntuaciones se pueden representar en un sistema cartesiano de ejes coordenados, distribuyendo a lo largo del eje X las puntuaciones referentes a «años de escolaridad» y a lo largo del eje Y los ingresos. Obtendríamos así seis puntos para cada par de observaciones o puntuaciones, en el sistema cartesiano, como sigue:

Datos

X	Y
1	2
2	4
3	6
4	8
5	10
6	12
$\overline{X}=3,5$	$\overline{Y}=7$



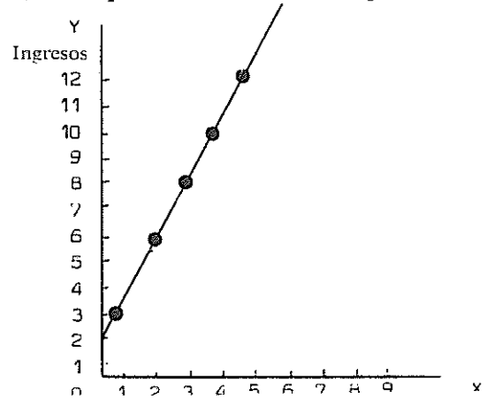
Resulta evidente de la observación de este gráfico que la relación entre ambas variables es muy simple. En realidad, se puede predecir la puntuación en Y a partir del conocimiento de la correspondiente puntuación en X, mediante la multiplicación por dos de cada puntuación de X. Esta relación se expresa mediante la ecuación siguiente:

$$Y=2X$$

Como se puede observar en la representación efectuada en el sistema cartesiano de coordenadas, las predicciones se distribuyen a lo largo de una línea recta, por lo que se dice que las variables X e Y están relacionadas linealmente. Veamos ahora otro conjunto de datos como los anteriores e, igualmente, los representamos en dos ejes coordenados:

Datos

X	Y
1	3
2	6
3	8
4	10
5	12
$\overline{X}=3$	$\overline{Y}=7,8$



Las puntuaciones de Y para los cinco casos se pueden predecir también por una fórmula simple, como la que sigue:

$$Y=2+2X$$

Es decir, dada una puntuación para X, podemos predecir el correspondiente valor de Y simplemente multiplicando por dos la puntuación de X y sumando una constante, 2. Como en el caso anterior, la ecuación describe una simple línea recta, que representa la relación lineal entre las dos variables. Pero ahora la fórmula que relaciona a X e Y incorpora un término constante, que representa el punto en el que la línea recta corta el eje Y. Pues bien, como se sabe, este tipo de ecuación con término constante responde a la forma más general de ecuación de una recta:

$$Y=a+bX \qquad [9.1]$$

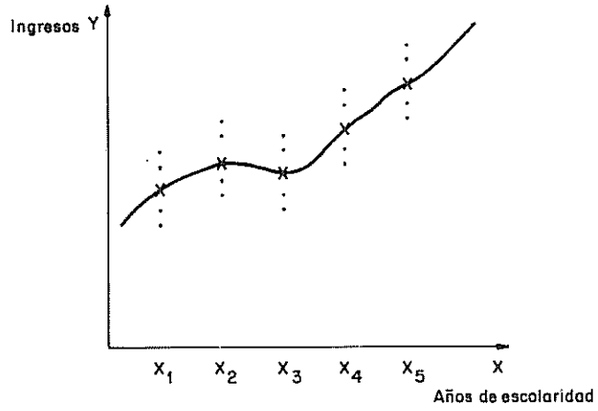
Cualquier relación lineal entre dos variables X e Y se puede expresar mediante la expresión [9.1]. El significado de los parámetros a y b es también sencillo. Cuando $X=0$, la expresión $Y=a+bX$ se convierte en $Y=a$, por lo que el parámetro a recibe el nombre de *ordenada en el origen*, ya que representa aquel punto de la recta cuya abscisa es el origen de coordenadas. En el ejemplo anterior, $a=2$.

El otro parámetro b representa la cuantía en que varía Y cuando X varía en una unidad. En el ejemplo anterior, cuando X aumenta un año de escolaridad, los ingresos se duplican, es decir, $b=2$. Al parámetro b se le denomina *coeficiente angular o pendiente de la recta.* Cuando b es un número positivo, la recta es creciente; esto es, al aumentar el valor de X crece también el valor de Y, mientras que si b es un número negativo la recta es decreciente, ya que al crecer el valor de la variable independiente X disminuye el valor que toma la variable dependiente Y.

9.2.2. *La ecuación de regresión y el ajuste por mínimos cuadrados*

Si en lugar de disponer de datos referentes a los años de escolaridad y nivel de ingresos de un grupo de individuos dispusiéramos de los correspondientes datos para grupos diferentes de población, el problema de la predicción se hace más significativo. Supongamos, por ejemplo, que para cada nivel de educación tenemos la distribución de los ingresos para cada uno de los individuos que se encuentran en el mismo nivel educativo. Naturalmente, no todos los individuos del mismo nivel educativo disfrutarán de idéntico nivel de ingresos, pero tales ingresos se distribuirán alrededor de una media. Pues bien, para cada nivel de escolaridad (valores de la variable X) tendremos una distribución de ingresos (variable Y) alrededor de una media. De este modo, representando los valores de X y las medias de Y en unos ejes coordenados, obten-

dremos una representación, lineal o curvilínea, de las medias de Y para cada valor de X como una *ecuación de regresión* * de Y en X, tal como se ilustra a continuación:



Como destaca Blalock (*op. cit.*, pág. 384), estas ecuaciones de regresión son las «leyes» de la ciencia. Conocida la expresión matemática que describe la forma y dirección de la línea o curva de las medias se pueden realizar predicciones muy exactas. Así, conociendo el nivel de escolaridad de un individuo y la ecuación matemática que describe la anterior relación, podemos predecir con bastante exactitud su nivel de ingresos. Ahora bien, a diferencia de otras ciencias más «exactas», en sociología usualmente no se conoce con precisión la curva o línea que relaciona a ambas variables. Al no disponer de mediciones precisas para sus variables, el sociólogo suele conceder cierta variabilidad a la ecuación de regresión y prefiere pensar en términos de medias y varianzas de la distribución de Y para cada X, en lugar de considerar la distribución precisa de los valores de Y en X.

Para hacer más manejable estadísticamente el problema de la predicción mediante la ecuación de regresión se hace necesario considerar un modelo lo más sencillo posible. Es por esta razón por lo que se presupone que la forma de la ecuación de regresión es lineal, que las distribuciones de los valores de Y en cada valor de X son del tipo normal, y que las varianzas de las distribuciones de Y son las mismas para cada valor de X (Blalock, *op. cit.*, pág. 385). De todos estos supuestos simplificadores, el que más nos interesa destacar para seguir nuestro hilo argumental es el de la linearidad. En efecto, si la regresión de Y en X es lineal, su ecuación será de la forma [9.1], es decir, se representará matemáticamente como la ecuación de una recta, $Y=a+bX$, en donde los parámetros a y b tienen el significado que anteriormente hemos visto, es decir, a es la ordenada en el origen y b es el coeficiente angular de la recta.

* En realidad, el verbo «regresar» no tiene definición matemática, aunque como señala Guttman (1979), pág. 112), el nombre de «regresión» desgraciadamente va unido a él. Una regresión es sencillamente una serie de medias condicionales, usualmente de medias aritméticas, tal como estamos viendo aquí. En sus orígenes, las «ecuaciones de regresión» se desarrollaron con los estudios genéticos que trataban de conseguir líneas genéticas puras, es decir, trataban de «regresar» de los tipos reales genéticos impuros, a los tipos originales puros. Desde entonces, el nombre de «ecuaciones de regresión» ha permanecido, aunque como ecuaciones matemáticas se aplican a la investigación empírica en ciencias, como la sociología, bastante alejadas de la genética.

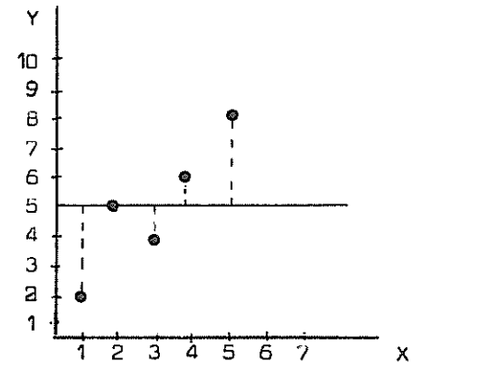
Insistamos una vez más en el hecho de que no todas las asociaciones entre dos variables pueden describirse bien por medio de una línea recta, ya que con frecuencia es curvilínea la forma geométrica que describe la asociación. No obstante, dadas las dificultades que plantea la búsqueda de una fórmula adecuada que se ajuste a la descripción de la relación curvilínea, se suele utilizar el modelo más simplificado y, por tanto, aproximado de la relación lineal, como el criterio «óptimo» de ajuste de una línea de regresión. En la realidad de la investigación empírica, los datos que obtiene el sociólogo suelen encontrarse bastante dispersos, aunque el conjunto de todos ellos se adapte bastante bien alrededor de la línea de la regresión. El problema entonces radica en situar la línea de regresión de tal forma que se ajuste lo mejor posible a los datos.

En último término, el criterio de ajuste de una línea de regresión responde al grado en que la variable dependiente puede predecirse a través de la ecuación que representa a dicha línea. Vamos a desarrollar esta idea mediante otro ejemplo ficticio, y para ello partiremos de unos pocos datos referentes a la relación que venimos estudiando entre escolaridad e ingresos:

Datos

X	Y
1	2
2	5
3	4
4	6
5	8
$\bar{X}=3$	$\bar{Y}=5$

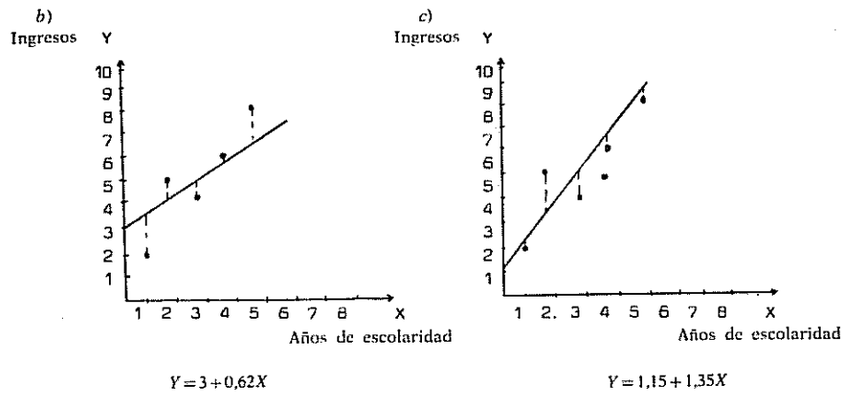
a) Ingresos



$Y=\bar{Y}+0X$

El conjunto de las cinco puntuaciones puede predecirse con formas diferentes. La forma más sencilla de hacerlo es mediante el uso de la media de Y, $\bar{Y}$, tal como se ha representado en el sistema de coordenadas *a)*. La línea de regresión sería en tal caso una recta horizontal, como se observa en dicha figura. Del mismo modo se podría pensar en for-

mular otras predicciones; por ejemplo, mediante las ecuaciones $Y=3+0{,}62X$ o $Y=1{,}15+1{,}35X$. En tal caso, las correspondientes representaciones gráficas serían las siguientes:



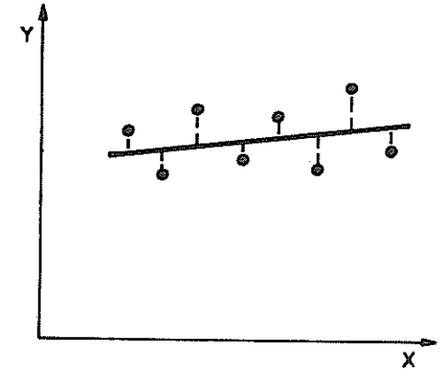
Ambas ecuaciones de ajuste se han elegido arbitrariamente, con fines ilustrativos. Con el fin de comprobar cuál de las tres ecuaciones predice con mayor exactitud los valores de Y en X podemos seguir el criterio de la varianza, que consistirá simplemente en restar de cada valor real de Y el resultante de la ecuación, se eleva al cuadrado la diferencia, se suman todos los casos y se divide por N. Es decir, mediante la *estimación de la varianza* $s^2_{yx}=\frac{\Sigma (Y-Y')^2}{N}$, en donde Y' representa el valor de Y calculado mediante la aplicación de la ecuación de predicción. Volviendo a los datos de nuestros ejemplos ficticios, tenemos que:

			Predicción a)			*Predicción b)*			*Predicción c)*		
	Datos		$y'=y+ox$			$y'=3+0{,}62x$			$y'=1{,}1+1{,}3x$		
	x	y	y'	$(y-y')$	$(y-y')^2$	y'	$(y-y')$	$(y-y')^2$	y'	$(y-y')$	$(y-y')^2$
	1	2	5	−3	9	3,62	−1,62	2,62	2,40	0,40	0,16
	2	5	5	0	0	4,24	0,76	0,58	3,70	1,30	1,69
	3	4	5	−1	1	4,86	−0,86	0,74	5,00	−1,00	1,00
	4	6	5	1	1	5,48	0,52	0,27	6,30	−0,30	0,09
	5	8	5	3	9	6,10	1,90	3,61	7,60	0,40	0,16
Suma:	15	25			20			7,82			3,10
Promedio:	3	5			4			1,56			0,62

En este cuadro se puede comprobar que las tres ecuaciones de predicción utilizadas ofrecen estimaciones de la varianza que van disminuyendo según vamos utilizando una ecuación más precisa. Así, para la ecuación *a)* la estimación de la varianza es bastante elevada, 4; mientras que para la ecuación de predicción *b)* la estimación de la varianza desciende a 1,52; quedando tan sólo el valor de s^2_{yx} en 0,64 para la ecuación *c)*, que es, naturalmente, la que mejor se ajusta a la distribución real de los datos.

La raíz cuadrada de la estimación de la varianza, $\sqrt{s^2_{yx}}$, se denomina *error típico de la estimación*. En las representaciones gráficas, en el eje de coordenadas de las tres ecuaciones se han trazado las distancias entre cada punto real y la línea de regresión. Tales distancias son máximas en la representación de la ecuación *a)* y mínimas en la representación de la ecuación *c)*. Esto se ha reflejado, tal como se ha dicho anteriormente, en una menor varianza de la estimación, es decir, que la ecuación de predicción $Y'=1{,}15+1{,}35X$ produce la menor varianza, o, dicho en otros términos, representa la línea de regresión de Y en X que produce el mejor ajuste. El criterio de «mejor» se basa en que la suma de las desviaciones al cuadrado de las puntuaciones alrededor de la recta es la más pequeña de todas las rectas consideradas, por lo que se le denomina *línea de regresión de mínimos cuadrados* de Y en X.

Así, pues, el criterio de los mínimos cuadrados consiste en encontrar la línea recta que tenga la propiedad de que la suma de los cuadrados de las desviaciones de los valores reales de Y en relación a dicha recta sea mínima. De este modo, si trazamos las líneas verticales que unen a cada uno de los puntos con la línea de mínimos cuadrados, y si se elevan al cuadrado tales distancias, la suma resultante será la menor posible de todas las sumas de cuadrados que se puedan calcular en relación a cualquier otra recta, tal como se observa en la siguiente figura:



Obsérvese que si en lugar de trazar las distancias verticales trazáramos las distancias horizontales obtendríamos una recta de regresión de

Y en X. Es decir, permutaríamos los papeles de las variables dependientes e independientes. El criterio sería, pues, el mismo, sólo que con los papeles de las variables cambiados.

Con el fin de obtener la línea de los mínimos cuadrados, se hace preciso calcular el valor de los parámetros a y b. Se puede demostrar que:

$$b=\frac{\Sigma\,(X-\overline{X})\,(Y-\overline{Y})}{\Sigma\,(X-\overline{X})^2} \qquad [9.2]$$

$$a=\overline{Y}-b\overline{X}=\frac{\Sigma Y-b\,(\Sigma X)}{N} \qquad [9.3]$$

en donde $\overline{X}$ e $\overline{Y}$ son las medidas aritméticas de las variables X e Y, respectivamente *. El numerador de b está formado por la expresión $\Sigma\,(X-\overline{X})\,(Y-\overline{Y})$, que se denomina la *covariación* de Y en X. Esta cantidad es análoga directamente a las sumas de los cuadrados para X o Y, con la diferencia de que, en lugar de elevar al cuadrado $(X-\overline{X})$ o $(Y-\overline{Y})$, se realiza el producto de ambos términos. En realidad, lo que se consigue de este modo es obtener una medida de cómo varían conjuntamente X e Y, de donde proviene el término de *covarianza*.

En realidad, la covariación puede ser positiva o negativa, según el sentido de la relación de X en Y. Cuando X e Y se encuentran relacionados positivamente, los valores superiores de X se encontrarán relacionados con los valores superiores de Y, y, viceversa, los valores inferiores de X se encontrarán relacionados con los valores inferiores de Y. Entonces, si $X>\overline{X}$, también será $Y>\overline{Y}$, o si $X<\overline{X}$, también $Y<\overline{Y}$. Con lo cual, el producto de $(X-\overline{X})$ por $(Y-\overline{Y})$ será positivo, y la suma de todos los productos también será positiva. E, inversamente, si X e Y se encuentran relacionados negativamente, cuando $X>\overline{X}$, será $Y<\overline{Y}$, con lo que el anterior producto será negativo.

Cuando no exista relación alguna entre X e Y, la mitad de los productos serán positivos y la otra mitad negativos, dado que X e Y varían independientemente. En tal caso, b valdrá cero o casi cero. De ahí que cuanto más alto sea el grado de relación entre las dos variables, mayor será el valor numérico de la covariación. Como se observa en la fórmula [9.2], el cálculo de b se realiza a partir de la covariación dividida por la suma de los cuadrados en X. Es de este modo como se calcula la pendiente de la ecuación de regresión, ya que ésta es la interpretación de b:

$$b=\frac{\text{Covarianza de } X \text{ en } Y}{\text{Varianza de } X}$$

* Para ver el fundamento matemático de estas fórmulas, se pueden consultar algunos de los libros de estadística que se citan en el apartado bibliográfico al final del capítulo, como, por ejemplo, Alcaide (1975), Amón (1978), Blalock (1979).

Existe otra fórmula para el cálculo de b que no necesita tener en cuenta el valor de las medias de X e Y, y sólo utiliza las puntuaciones individuales de ambas variables. Dicha fórmula operacional de b se escribe como sigue:

$$b=\frac{N\Sigma XY-(\Sigma X)\,(\Sigma Y)}{N\Sigma X^2-(\Sigma X)^2} \qquad [9.4]$$

Esta fórmula, en la que tanto el numerador como el denominador aparecen multiplicados por N, es de más fácil manejo que la [9.2].

Ejemplo: Supongamos de nuevo que estamos estudiando la relación que existe entre años de escolaridad e ingresos, y que hemos reunido los mismos datos que hemos utilizado en el ejemplo anterior. Lo que se trata ahora es de calcular los parámetros a y b y la consiguiente ecuación de regresión: $Y=a+bX$. Para ello prepararemos la siguiente tabla de datos y cálculos:

(X) *Años de escolaridad*	(Y) *Nivel de ingresos*	X^2	Y^2	XY
1	2	1	4	2
2	5	4	25	10
3	4	9	16	12
4	6	16	36	24
5	8	25	64	40
$\Sigma X=15$	$\Sigma Y=25$	$\Sigma X^2=55$	$\Sigma Y^2=145$	$\Sigma XY=88$

Sustituyendo estos cálculos en la fórmula [9.4], obtendremos el valor de b:

$$b=\frac{N\Sigma XY-(\Sigma X)\,(\Sigma Y)}{N\Sigma X^2-(\Sigma X)^2}=\frac{5\cdot 88-15\cdot 25}{5\cdot 55-225}=\frac{440-375}{275-225}=1{,}3$$

Y ahora, sustituyendo en [9.3], se puede obtener el valor de a:

$$a=\frac{\Sigma Y-b\,(\Sigma X)}{N}=\frac{25-1{,}3\cdot 15}{5}=1{,}1$$

Con lo que la ecuación de regresión queda como sigue:

$$Y=1{,}1+1{,}3X$$

A partir de esta fórmula se pueden predecir los niveles de ingresos

para los diferentes niveles de escolaridad. Así, para el nivel de escolaridad $X=9$, el nivel de ingresos sería el siguiente:

$$Y=1{,}1+1{,}3 \cdot 9=12{,}8$$

De este modo hemos establecido una fórmula simple que describe la *naturaleza* de la asociación entre dos variables de intervalo y que, al mismo tiempo, nos permite utilizar la información que disponemos sobre la variable independiente, con el objeto de lograr una predicción mejor de la variable dependiente. A continuación nos ocuparemos de desarrollar una medida del grado de asociación que expresará, en último término, la reducción proporcional en los errores predictivos que se logra con dicha fórmula.

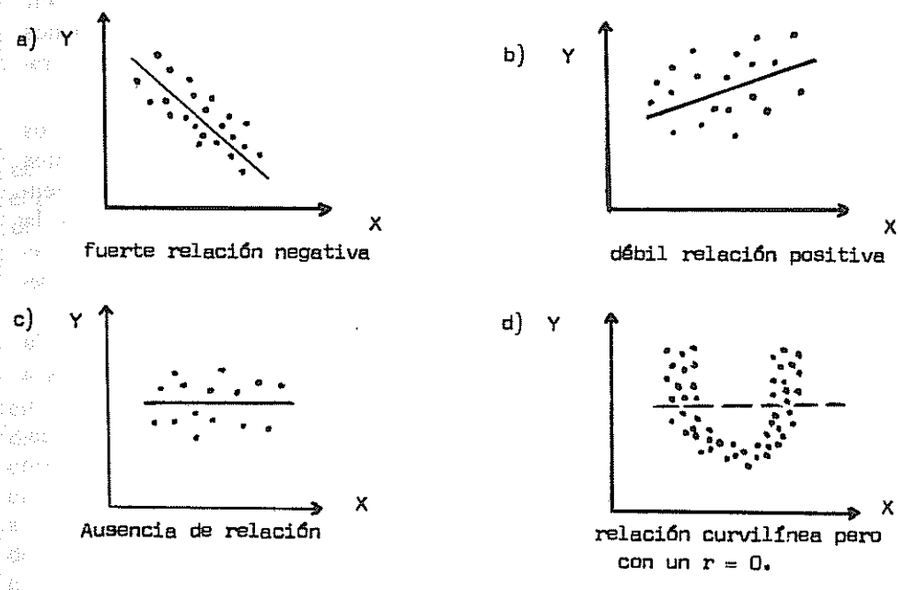
9.3. Correlación. Coeficiente r de correlación de Pearson

En realidad, en el estado actual del desarrollo de la investigación empírica en sociología, los sociólogos están con frecuencia más interesados en el descubrimiento de las variables más íntimamente asociadas con una variable dependiente determinada que en predecir, mediante una ecuación de regresión, los valores de la variable dependiente a partir de los valores conocidos de las variables independientes. Dado el carácter exploratorio de una parte todavía importante de la investigación empírica sociológica, el análisis de regresión pasa a un segundo plano, quedando como objetivo prioritario el estudio del grado de asociación o correlación entre las variables. En las ciencias más precisas, como la física o la biología, el problema, por el contrario, se centra más en la predicción exacta que en el análisis del grado de asociación. El énfasis, como se ve, depende del grado de desarrollo de los procedimientos de medición de cada ciencia. La sociología, con un nivel ciertamente bajo de desarrollo de la medición de sus variables, tiene hoy en día que concentrarse más en el estudio de la correlación que en el de la regresión y predicción.

El coeficiente de correlación más ampliamente difundido para el análisis de la asociación entre dos variables de intervalo fue desarrollado por Karl Pearson (1857-1936). Aunque fue el también británico Francis Galton el que desarrolló la idea de la correlación, Pearson generalizó los métodos y conclusiones de su compatriota y derivó la fórmula que actualmente se llama «momento-producto de Pearson», logrando una rutina de cálculo que ha alcanzado difusión universal. En la literatura estadística inglesa se habla del coeficiente de correlación del momento-producto de Pearson, r, aunque de una manera más simplificada se habla del coeficiente r de Pearson. Lo que mide en realidad este coeficiente es la cantidad de dispersión en relación a la ecuación lineal de mínimos cuadrados.

La dispersión en relación a dicha ecuación se podría igualmente medir mediante el cálculo de la desviación típica en relación a la recta, pero, como se ha dicho antes, el coeficiente r de Pearson ha logrado aceptación universal en el mundo de la ciencia. Se trata de un coeficiente fácilmente interpretable, ya que su recorrido oscila entre $-1{,}0$ (asociación perfecta negativa) hasta $+1{,}0$ (asociación perfecta positiva). Al tratarse de una medida de la relación lineal, que mide el grado de ajuste a la recta de mínimos cuadrados, no se puede interpretar el valor de $r=0$ como ausencia total de relación, ya que las variables X e Y pueden estar fuertemente asociadas de forma curvilínea y tener, sin embargo, un valor de r igual a cero o próximo a cero. De ahí que, antes de calcular el valor de r, resulta aconsejable representar en un sistema de coordenadas cartesiano los valores de X e Y, para observar si su distribución aproximada es lineal o curvilínea. En la actualidad, muchos programas estadísticos de ordenador incluyen entre sus rutinas de cálculo el diagrama de dispersión, lo que resulta muy conveniente para interpretar los resultados.

Veamos ahora, antes de pasar a analizar la fórmula del coeficiente de correlación de Pearson, distintos ejemplos de diagramas de dispersión para valores de X e Y:



Hemos señalado anteriormente que los límites superiores de r son $+1{,}0$ y $-1{,}0$. Si todos los puntos se encuentran en la línea recta, el coe-

ficiente r valdrá la unidad, dependiendo el signo de que la relación sea positiva o negativa. Cuando la distribución de los puntos se aproxima a la línea recta, el valor de r se encontrará próximo a la unidad. Ese sería el caso de la distribución que se presenta en los ejes de coordenadas *a)*, en el que si se calculara r se obtendría un valor próximo a 0,90, aunque afectado de signo negativo, ya que es negativa la relación entre las variables. En los ejes de coordenadas *b)*, la distribución de los puntos pone de manifiesto una débil relación positiva, que daría lugar a un coeficiente r que no sería superior a 0,40. La ausencia de relación lineal, es decir, el valor de $r=0$, se representa en los gráficos *c)* y *d)*, pero existe una diferencia importante entre ambas distribuciones de puntos. Mientras que en el gráfico *c)* el valor de $r=0$ se corresponde con una ausencia de relación entre X e Y, en el gráfico *d)* el valor de $r=0$ se refiere únicamente a la ausencia de relación lineal, pero no de relación curvilínea, ya que los puntos se distribuyen perfectamente en forma de U, pero naturalmente la relación lineal es nula.

Por ello, cuando el investigador encuentra una $r=0$, no puede concluir de inmediato que las variables no se encuentran relacionadas. Por eso resulta conveniente la inspección del diagrama de dispersión de los puntos para poder saber si se trata, de hecho, de una ausencia de relación o si la relación es lo suficientemente curvilínea como para producir un coeficiente de correlación igual a cero. Afortunadamente, en muchos estudios sociológicos, las relaciones entre variables pueden estudiarse razonablemente bien mediante aproximaciones lineales.

Veamos ahora cómo se define el coeficiente r de Pearson. Hemos visto anteriormente que mediante la ecuación de regresión por mínimos cuadrados, se pueden predecir las puntuaciones en la variable dependiente Y con mayor precisión que la que se lograría con la utilización de la media global de Y. Por esta razón, se puede afirmar que la recta de regresión nos ayuda a «explicar» parte de la variación en la variable dependiente, quedando sin explicar el resto de la variación de Y. Naturalmente, la *variación total* de Y en relación a la media será igual a la suma de la *variación explicada más la variación no explicada*. Vamos a ilustrar estos conceptos mediante el desarrollo de un ejemplo práctico, utilizando los datos que venimos manejando en este capítulo que relacionan el nivel de escolaridad con el nivel de ingresos, y como ecuación de regresión utilizaremos $Y=1{,}1+1{,}3X$:

Puntuaciones reales		*Puntuaciones de y obtenidas mediante ec. regresión*	*Variación no explicada*	*Variación explicada*	*Variación total*
x	y	y'	$(y-y')^2$	$(y'-\bar{y})^2$	$(y-\bar{y})^2$
1	2	2,40	0,16	6,76	9
2	5	3,70	1,69	1,69	0
3	4	5,00	1,00	0	1
4	6	5,30	0,09	1,69	1
5	8	7,60	0,16	6,76	9
15	25	25,00	3,10	16,90	20

$\bar{y}=5$

$20=3{,}10+16{,}90$

$(y-\bar{y})^2 = (y-y')^2 + (y'-\bar{y})^2$

Variación total = Variación no explicada + Variación explicada

De lo que se trata es de explicar el máximo posible de variación, y el cuadrado del coeficiente de correlación de Pearson, r^2, expresa precisamente el grado en que la ecuación de regresión lineal explica la variación en la variable dependiente, tal como sigue:

$$r^2=\frac{\text{Variación explicada}}{\text{Variación total}}=\frac{\Sigma\,(Y'-\bar{Y})^2}{\Sigma\,(Y-\bar{Y})^2}$$

También se puede expresar el coeficiente de correlación de Pearson en términos de varianzas. En concreto, el coeficiente de correlación es el cociente entre la covariación de X e Y y la raíz cuadrada del producto de la variación en X y de la variación en Y:

$$r=\frac{\Sigma\,(X-\bar{X})\,(Y-\bar{Y})}{\sqrt{[\Sigma\,(X-\bar{X})^2]\,[\Sigma\,(Y-\bar{Y})^2]}}=\frac{s_{yx}}{\sqrt{(s_x^2)\,(s_y^2)}} \qquad [9.4]$$

En el último término de la fórmula [9.4], el coeficiente de correlación r aparece como el cociente entre la covarianza y el producto de las desviaciones típicas de X e Y. Ahora bien, la primera expresión que se contiene en [9.4] no sirve como fórmula operacional porque puede producir valores superiores a la unidad. En efecto, sabemos que la covarianza es una medida de la variación conjunta de X e Y, pero su magnitud depende de la cantidad global de variabilidad en ambas variables, pudiendo en algunos casos sobrepasar considerablemente de la unidad su valor numérico. Por ello resulta inconveniente utilizar la expresión [9.4] como medida de asociación. Pero si se divide esta expresión por el producto de las dos desviaciones típicas se obtiene una medida estandariza-

da que varía entre $-1{,}0$ y $+1{,}0$, siendo el valor cero consecuencia de la falta de la relación lineal entre X e Y.

Veamos con más detalle estos extremos. Ya hemos visto anteriormente que la covarianza será cero cuando X e Y no están relacionados linealmen; luego, cuando esto ocurra, el coeficiente $r=0$. Con la misma sencillez se puede demostrar que el límite superior de r es la unidad. Tomemos el caso de un valor positivo para b y en el que todos los puntos se concentran en la recta. Sabemos que para cada valor de Y se puede escribir $Y=a+bX$. Ahora bien, como las medias también se encuentran en la recta, $\overline{Y}=a+b\overline{X}$. Por tanto, para todos los puntos de la recta:

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

de donde:

$$\Sigma\,(X-\overline{X})\,(Y-\overline{Y})=b\Sigma\,(X-\overline{X})^2 \qquad [9.5]$$

multiplicando por b los dos términos de la expresión queda:

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

y sustituyendo en [9.4] queda:

$$r=\frac{b\Sigma\,(X-\overline{X})^2}{\sqrt{[\Sigma\,(X-\overline{X})^2]\;[b^2\,\Sigma\,(X-\overline{X})^2]}}=1{,}0$$

E igualmente se puede demostrar que, para el caso en que todos los puntos se distribuyeran a lo largo de una recta de pendiente negativa, el valor resultante de r sería $-1{,}0$.

Resulta conveniente destacar también la relación existente entre el coeficiente de correlación r y la pendiente de la ecuación de regresión b. De la expresión [9.5] podemos despejar b, con lo que tenemos:

$$b=\frac{\Sigma\,(X-\overline{X})\,(Y-\overline{Y})}{\Sigma\,(X-\overline{X})^2} \qquad [9.6]$$

Vemos, pues, que la fórmula [9.4] de r y la fórmula [9.6] de b contienen idéntico numerador. Por tanto, cuando b sea cero, también valdrá cero r, y viceversa. Si consideramos tanto la regresión de X en Y como su opuesta, la regresión de Y en X, la comparación en [9.5] y [9.6] nos conduce a la conclusión de que:

$$r^2=b_{yx}b_{xy}=\frac{(s_{xy})^2}{s_x^2\,:\,s_y^2} \qquad [9.7]$$

Es decir, que el cuadrado del coeficiente de correlación de Pearson, r^2, entre dos variables X e Y es igual al producto del coeficiente angular o pendiente de la recta de regresión de Y en X, b_{yx}, por el coeficiente angular o pendiente de la recta de regresión de X en Y, b_{xy}. De [9.7] se deduce que cuando $r=1{,}0$, $b_{yx}=1/b_{xy}$, lo que significa que ambas ecuaciones de regresión coinciden. Por el contrario, cuando r se aproxima a cero, el ángulo entre las dos rectas se va haciendo más grande, y, finalmente, cuando $r=0$, las dos rectas son perpendiculares.

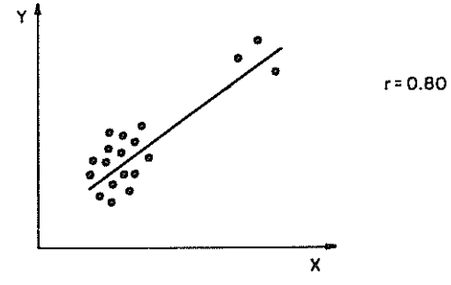
Ahora bien, ninguna de las expresiones empleadas hasta ahora para definir r resulta de interés a efectos operacionales. Se puede demostrar que r se puede expresar en términos de las mismas expresiones utilizadas para calcular a y b, del modo siguiente:

$$r=\frac{N\Sigma XY-(\Sigma X)\,(\Sigma Y)}{\sqrt{[N\Sigma X^2-(\Sigma X)^2]\;[N\Sigma Y^2-(\Sigma Y)^2]}} \qquad [9.8]$$

Con los cálculos previos realizados para determinar los parámetros a y b (ver ejemplo de la sección 9.2.2) resulta muy fácil conocer el valor r, que para los datos utilizados anteriormente es:

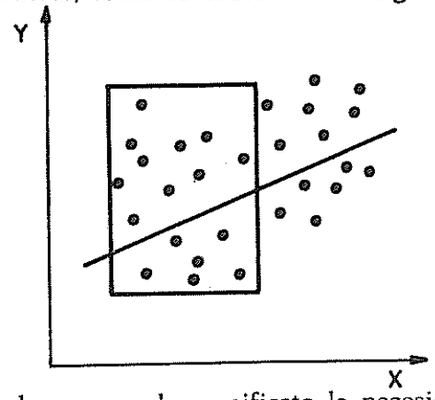
$$r=\frac{5\cdot 88-15\cdot 25}{\sqrt{(5\cdot 55-15^2)\;(5\cdot 145-25^2)}}=\frac{65}{70{,}6}=0{,}92$$

Sabemos ya, pues, que el coeficiente r de Pearson es una medida de correlación entre dos variables de intervalo y que sus valores extremos son $-1{,}0$ y $+1{,}0$. Los valores de r indican tanto la dirección como el grado (fuerza) de la asociación. Ahora bien, conviene notar que al tratarse de una medida que implica la noción y cálculo de varianzas y covarianzas, resulta ser muy sensible a la presencia de unos pocos valores extremos en una o en las dos variables. Observemos, como ejemplo, el siguiente diagrama de dispersión:



La presencia de tres puntos extremos da lugar a un valor de r próximo al 0,80, lo que representa, ciertamente, una fuerte correlación. Sin embargo, si hubiéramos calculado un coeficiente de correlación para cada uno de los dos grupos de puntos, los dos valores obtenidos hubieran sido notablemente inferiores, indicando sendas correlaciones débiles.

E, inversamente, puede ocurrir que dentro de un limitado recorrido de variabilidad de los valores de X e Y la correlación sea débil, pero considerado el conjunto de la distribución de los valores de X e Y la correlación sea fuerte, como se observa en el siguiente gráfico:



Ambos ejemplos ponen de manifiesto la necesidad de considerar la variabilidad total de X e Y antes de realizar una afirmación acerca de su grado de correlación. En el primero de los casos quizá pueda resultar aconsejable excluir a los casos extremos del cómputo global, mientras que en el segundo de los casos el investigador ha de esforzarse por lograr disponer del recorrido total de la variabilidad de los valores de ambas variables.

9.3.1. *Interpretación del coeficiente de correlación*

El coeficiente pearsoniano de correlación r es una medida de asociación del tipo que hemos denominado aquí «reducción proporcional del error», *RPE*. Elevado al cuadrado, r^2, el coeficiente expresa la reducción proporcional en el error cometido al predecir valores para la variable dependiente a partir de la ecuación de regresión, ajustada por mínimos cuadrados, en lugar de utilizar la media global como criterio predictivo. Dado que la regresión de Y en X y la regresión de X en Y tienen ambas la misma cantidad de dispersión alrededor de sus respectivas rectas de regresión, resultará el mismo coeficiente de correlación de ambas ecuaciones. Por tanto, r es una medida simétrica del grado de correlación. Dicho en otros términos, r^2 representa la proporción de la variación en una variable que queda explicada por su asociación lineal con otra variable.

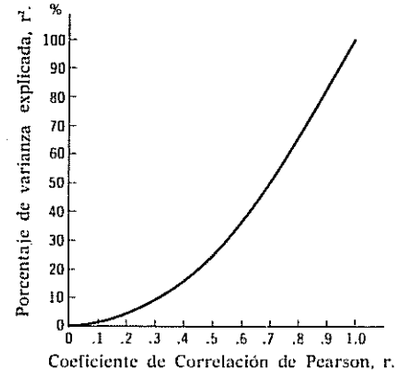
El tipo de relación existente entre r y r^2 se pone de manifiesto en la figura 1, en la que se puede observar la proporción de variación que queda explicada para diferentes valores de r.

Venimos interpretando el coeficiente r^2 en función de la cantidad de variación explicada. Ahora bien, conviene insistir en que cuando hablamos de explicación no nos estamos refiriendo a una explicación causal, sino simplemente a una asociación entre dos variables.

Como se trata de una medida simétrica, r^2 se puede interpretar tanto como el cociente *(ratio)* entre la variación explicada en Y y la variación total en Y como el cociente entre la variación explicada en X y la variación total en X. Es decir, que el cuadrado del coeficiente de correlación se puede interpretar como la proporción de la variación total en una variable que queda explicada por la otra. La cantidad $\sqrt{1-r^2}$, que se denomina *coeficiente de alienación*, representa la raíz cuadrada de la proporción de variación no explicada por la variable independiente.

FIGURA 1

Porcentaje de variación explicada por coeficientes de correlación de diferentes tamaños



FUENTE: LOETHER y MCTAVISH: *op. cit.* pág. 246.

Nótese que no existe una interpretación sencilla y directa para el propio coeficiente r. Como destaca Blalock (*op. cit.*, pág. 409), en la práctica los valores de r nos pueden desorientar porque, a excepción de los valores 0 y 1,0, serán superiores a los de r^2. Así, nos puede parecer que un coeficiente r de valor 0,55 representa una buena correlación, cuando de hecho sólo estamos explicando $(0{,}55)^2 = 0{,}3025$, es decir, el 30 por 100 de la varianza. Es por ello por lo que las correlaciones que valen 0,3 o menos explican una pequeña proporción de la variación. En la siguiente

tabla aparecen las relaciones numéricas existentes entre r, r^2, $1-r^2$ y $\sqrt{1-r^2}$:

Relaciones numéricas entre r, r^2, $1-r^2$ y $\sqrt{1-r^2}$

r	r^2	$1-r^2$	$\sqrt{1-r^2}$
0,90	0,81	0,19	0,44
0,80	0,64	0,36	0,60
0,70	0,49	0,51	0,71
0,60	0,36	0,64	0,80
0,50	0,25	0,75	0,87
0,40	0,16	0,84	0,92
0,30	0,09	0,91	0,95
0,20	0,04	0,96	0,98
0,10	0,01	0,01	0,995

FUENTE: BLALOCK, 1979, pág. 409.

Aquí se ve con toda claridad que, para que se produzca una reducción importante del porcentaje de variación explicada, el valor de r ha de ser superior a 0,70.

9.3.2. *Correlación y regresión con valores típicos, z*

Al estudiar la distribución normal vimos el interés que tienen los valores típicos o puntuaciones z, que representan el número de unidades de desviación típica que separa a cada puntuación de la media. Las puntuaciones típicas, al gozar de las propiedades de que la media de su distribución es cero y la desviación típica de dicha distribución es la unidad, son de gran utilidad en la estadística inferencial. Pero también resultan útiles para expresar, de forma más simplificada y directa, la correlación y la regresión.

En efecto, cuando los datos vienen expresados en términos de puntuaciones o valores z, es decir, cuando las puntuaciones se expresan en términos de z_x y z_y en lugar de X e Y, el coeficiente de correlación es simplemente un promedio de la suma del producto de los valores z:

$$r=\frac{\Sigma z_x \cdot z_y}{N} \qquad [9.9]$$

Esta expresión revela, una vez más, que un simple cociente *(ratio)* expresa el grado de asociación o correlación entre dos variables de igual manera que, por medio de otro tipo de cocientes, hemos expresado anteriormente la media aritmética, la varianza y las medidas ordinales de asociación.

El valor de r expresado mediante [9.9] varía igualmente entre $-1{,}0$ y $+1{,}0$. En efecto, sabemos que la suma de los cuadrados de las puntuaciones z es igual al número total de casos, N. Pues bien, cuando un caso

tiene una puntuación que se encuentra en idéntica posición relativa tanto en la variable X como en la variable Y, el valor de z en ambas variables será también el mismo, y $\Sigma z_x \cdot z_y = N$, con lo que $r=N/N=1$. Pero en la medida en que las posiciones de los valores individuales sean diferentes en cada variable, los valores z también serán diferentes, con lo que $\Sigma z_x z_y < N$ y, por tanto, el valor de r será menor que 1, expresando su valor el grado de correlación entre X e Y.

La forma de la ecuación de regresión utilizando valores z, denominada *ecuación de regresión tipificada*, tiene también una expresión sencilla y directa:

$$z'_y = r\,(z_x)$$

Esto es, el valor estimado de la puntuación z en la variable Y, z'_y, se calcula a partir del producto del valor de z_x por una constante, que no es otra cosa que el coeficiente r de correlación de Pearson.

9.4. LA MATRIZ DE CORRELACIONES

De igual forma que vimos en el capítulo anterior la construcción de una matriz de medidas ordinales de asociación, se puede construir una matriz de correlaciones en base a los coeficientes r de correlación de Pearson obtenidos al calcular la correlación entre pares de variables de un conjunto de ellas.

En la siguiente tabla aparece una matriz de correlación entre ocho variables empleadas en un estudio sobre el significado del voto político en España:

Matriz de correlaciones entre ocho variables políticas, demográficas e históricas

	Históricas				*Orientación política*		*Demográficas*	
	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*
A. Porcentaje de votos de izquierda ...	—							
B. Porcentaje de votos de derecha ...	.608	—						
C. Porcentaje de votos del PSOE	.400	−.187	—					
D. Porcentaje de votos de la CEDA ...	−.331	.721	.046	—				
E. Puntuación media (izda.-dcha)	−.107	.210	.162	.396	—			
F. Proporción favorable al centralismo.	.015	.171	.355	.352	.596	—		
G. Población autóctona.	−.152	.026	.213	.265	.507	.573	—	
H. Incremento de población intercensal.	.256	−.142	−.049	−.327	−.525	−.626	−.804	—

Correlación entre las variable *E* y *D*.

FUENTE: *D.* VILA, F. A. Oriza y M. GÓMEZ REINO: «Sociología del actual cambio político en España», *FOESSA*, 1978, pág. 720.

Tomando como unidad de análisis la provincia, los autores de este estudio calcularon las correlaciones existentes entre los resultados de las elecciones legislativas de junio de 1977 en España y diversas variables demográficas, políticas e históricas. Las variables dependientes utilizadas fueron el voto a los cuatro grandes partidos, operativizadas como votos de izquierda (PSOE y PCE) y votos de derecha (AP y UCD).

La tabla está organizada de forma que cada fila se refiere a una variable, al igual que ocurre con las columnas. El número que aparece en la intersección de cada fila con cada columna es un coeficiente de correlación que pone de manifiesto la correlación existente entre las variables referidas en la cabecera de cada fila y cada columna. Así, y tal como se señala en la propia tabla, el número 0,396 representa el coeficiente de correlación entre la variable «puntuación media (izda.-dcha.)» y la variable «porcentaje de votos de la CEDA».

Obsérvese que, dado que el coeficiente de correlación empleado en la tabla es el r de Pearson, que es una medida simétrica, sólo se han presentado los coeficientes para la mitad de la matriz, ya que la otra mitad de la matriz es idéntica (esto es, la correlación entre las variables E y D es idéntica que la correlación entre las variables D y E).

9.5. Consideraciones finales sobre la selección e interpretación de las medidas de asociación

En el capítulo anterior y en el presente hemos podido estudiar las medidas de asociación que con mayor frecuencia utilizan los sociólogos en sus análisis de datos empíricos. Tal como se ha visto, la selección de la medida más apropiada para resolver un problema concreto se realiza en base a considerar el nivel de medición de las variables, el tipo de relación —simétrica o asimétrica— que las caracteriza y los rasgos de la asociación que se desean destacar.

La consideración del nivel de medición de las variables es determinante a la hora de seleccionar una medida de asociación apropiada. Si se utiliza una medida de bajo nivel de medición con datos definidos a un nivel más alto de medición se perderá una información apreciable, mientras que si se hace lo contrario, esto es, utilizar una medida de alto nivel, por ejemplo r, con datos de bajo nivel, por ejemplo ordinales, cometeremos un error estadístico. Por eso es preciso adecuar la selección de una medida de asociación apropiada al nivel de medición de los datos de que disponemos.

También es importante tener en cuenta la manera en que se considera la relación entre la variable independiente y la dependiente. Cuando lo que se busca es la explicación y predicción de una variable dependiente se seleccionará una medida asimétrica. Pero si lo que realmente andamos buscando es la forma en que las dos variables covarían o se relacionan entre sí, entonces nos basta con elegir una medida simétrica.

Igualmente, hemos visto con anterioridad que las medidas de asociación difieren en los rasgos de la asociación a los que son más sensibles, con lo que seleccionaremos para resolver un problema concreto aquella medida que se adecúe mejor al rasgo que se pretende estudiar. Así, por ejemplo, algunas medidas como λ_{yx} y r están orientadas a la predicción de un valor central de una variable dependiente. Otras, como el coeficiente Tau, predicen la distribución de una variable; mientras que las hay, como r y G, que permiten el contraste entre un conjunto de datos observados y un modelo de asociación (o independencia) perfecta.

La selección de una medida concreta de asociación para resolver un problema determinado será, pues, el resultado de ponderar una serie de decisiones en relación a los diferentes aspectos que hemos analizado con anterioridad, alcanzando un óptimo por lo que se refiere a los fines de la investigación y al tipo de información que suministra el coeficiente elegido.

No quisiéramos finalizar este capítulo sin dejar de señalar un tipo de error que se ha cometido más de una vez al interpretar los resultados de un análisis de asociación entre variables. A veces se tiende a otorgar un significado a las medidas de asociación que no contienen. Nos referimos a la tendencia a atribuir a las variables independientes la capacidad de explicar el comportamiento de las variables dependientes. Así, por ejemplo, si el nivel de educación nos permite reducir el error al predecir la anomia, se puede estar tentado de afirmar que un bajo nivel de educación provoca o causa niveles altos de anomia, y viceversa. Pero esto no resulta en absoluto ser una interpretación correcta. Porque una cosa es la existencia de una fuerte asociación o correlación entre dos variables y una muy diferente es la existencia de una relación causal entre ambas.

En sociología se conocen muchas asociaciones entre variables, pero pocas relaciones causales. En puridad, sólo el experimento permite constatar la existencia o no de relaciones causales. Desgraciadamente, el sociólogo tiene pocas oportunidades de realizar experimentos sociales con los que contrastar sus teorías y poner a prueba las hipótesis sobre relaciones causales entre variables. En realidad, el sociólogo tiene que conformarse la mayor parte de las veces con ilustrar sus teorías con la obtención de datos empíricos por medios no experimentales, que suelen tener un alcance bastante limitado. Incluso si su teoría postula la existencia de una relación causal entre dos variables, y al realizar una encuesta encuentra que tales variables se encuentran fuertemente asociadas, no se puede concluir de ello que, en efecto, tales variables estén causalmente relacionadas. Porque la causalidad estará implícita en la teoría, pero no lo está en absoluto en la asociación o correlación. Esta hay que interpretarla, tal como se ha venido haciendo aquí, como una covariación o una influencia de una variable en otra. Sólo eso. Pero para inferir causalidad hace falta bastante más que la existencia de una fuerte correla-

ción. Por eso conviene tener siempre presente que *ni la asociación ni la correlación significan causación.*

9.6. Terminología

Se recomienda la memorización y comprensión del significado de cada uno de los términos y conceptos siguientes:

— Correlación.
— Regresión.
— Ecuación de regresión lineal.
— Ajuste por mínimos cuadrados.
— Ordenada en el origen.
— Coeficiente angular o pendiente de la recta.
— Covarianza; covariación.
— Coeficiente r de correlación de Pearson.
— Varianza explicada; varianza inexplicada.
— Coeficiente de alienación.
— Ecuación de regresión tipificada.
— Matriz de correlaciones.

EJERCICIOS

1. Ajustar una recta de mínimos cuadrados a los datos de la siguiente tabla, utilizando: a) x como variable independiente, y b) x como variable dependiente:

x	1	3	4	6	9	11	13
y	1	4	5	5	7	8	10

2. En una encuesta sobre ingresos familiares, se obtuvieron los siguientes resultados sobre los ingresos medios familiares para hogares de diferentes tamaños:

Número de miembros del hogar	*Ingresos medios (miles de pesetas)*
1	94
2	152
3	218
4	248
5	268
6	281