12

Estimación de la validez

La validez del tests está ligada a su capacidad para pronosticar otras variables relevantes distintas del test. El coeficiente de validez es la *correlación* del test con un criterio. Y el pronóstico de otras variables generalmente se efectúa mediante *ecuaciones de regresión*. En este capítulo nos ocuparemos de los principales procedimientos tradicionales para estimar la validez, mediante coeficientes de correlación, y de los principales procedimientos tradicionales para el pronóstico de otras variables, mediante las ecuaciones de regresión.

1. Estimación del coeficiente de validez: el coeficiente de correlación de Pearson

En la exposición de la Teoría Clásica de Tests hemos encontrado diversas situaciones en las que un concepto psicométrico se reduce en su cálculo al coeficiente de correlación de Pearson entre determinadas variables. Así por ejemplo, se definen como coeficientes de correlación de Pearson:

-la fiabilidad del test, definida en varias de sus modalidades como una correlación entre dos mediciones

-la homogeneidad del ítem, que es la correlación entre un ítem y el test total;

-el índice de validez del ítem que es la correlación del ítem con un criterio, variable ajena y distinta al test;

-la correlación entre cada par de ítems;

-el coeficiente de validez del test, que es la correlación del test con un criterio externo. Esta última aplicación es particularmente relevante dada la importancia de la validez en la teoría de los tests.

Si en cada una de las situaciones anteriores denominamos X a una variable e Y a la otra podremos aplicar la fórmula del coeficiente de correlación que resulte oportuna. El estadístico coeficiente de correlación de Pearson es el mismo, se trate de correlacionar un ítem con otro, un test con otro, un ítem con un test o cualquier otro caso, pero su significado psicométrico y psicológico puede variar en función de la naturaleza de aquello que se correlaciona.

El coeficiente de validez sintetiza la relación entre dos variables, el test y un criterio. Se estima mediante el coeficiente de correlación de Pearson de aquí que una comprensión adecuada de los diagramas de dispersión y del significado de este coeficiente sean requisitos para este punto esencial de la Teoría Clásica de Tests.

1.1. Distribuciones bidimensionales entre test y criterio

El concepto de distribución de una variable X hace referencia a la frecuencia (número de casos) con que ha aparecido cada uno de los valores posibles del rango de X. Una *distribución bidimensional* refleja la frecuencia con que ha aparecido cada punto del producto cartesiano de los valores posibles de ambas variables. Para dos variables, un test X y un criterio Y, su distribución bidimensional ó distribución conjunta refleja la frecuencia con que en una muestra dada han aparecido cada par de valores posibles de X e Y.

Diagramas de dispersión

Para estudiar la relación entre un test X y un criterio Y es necesario disponer de una muestra de N casos en los que se haya registrado tanto el valor de X como el de Y.

La representación gráfica de los pares de puntuaciones de ambas variables cuantitativas sobre unos ejes cartesianos se denomina *diagrama de correlación o diagrama de dispersión*.

Esta información puede resumirse en una tabla de doble entrada donde X e Y en que cada celdilla representa la frecuencia conjunta de los valores de X e Y que expresan los ejes de la tabla. Si X e Y son discretas de hecho es posible ubicar en los ejes todos los valores, si son continuas pueden utilizarse intervalos. Una tabla de correlación es una tabla de doble entrada en la que se representan las frecuencias conjuntas de dos variables métricas.

Supongamos que X es un pequeño test que toma el rango de 1 a 6 e Y un criterio que toma el rango de 1 a 9. Se ha tomado información para un conjunto de 25 casos obteniéndose las siguientes puntuaciones en X e Y, expresadas en una matriz de datos:

Caso:	X	Y
1	3	5
2	6	9
3	2	3
4	4	6
5	5	7
6	3	5
7	6	7
8	4	5
9	1	1
10	1	2
11	6	8
12	4	7
13	2	4
14	1	2
15	5	8
16	2	5
17	5	8
18	5	9
19	1	2
20	2	4
21	3	7
22	3	6
23	4	8
24	1	3
25	5	8

En la tabla anterior las puntuaciones de cada caso forman un par ordenado (X,Y). En la tabla puede observarse que, en general, las puntuaciones mayores de X van junto a puntuaciones mayores de Y, y viceversa, las menores de X junto a menores de Y. Pero la relación no puede ser exacta porque para el mismo valor de X encontramos varios de Y, y viceversa, para el mismo de Y también podemos encontrar en este ejemplo, varios de X.

La matriz de datos directos puede permitir algunas observaciones como las que acabamos de hacer, pero, no facilita excesivamente la comprensión de la relación entre los datos de los pares, especialmente si el tamaño de la muestra N es grande.

La matriz de datos de dos variables cuantitativas puede representarse adecuadamente mediante un diagrama de dispersión, que refleja la información de los pares en un plano cartesiano cuyos ejes son las variables. Si hay pares iguales, como es el caso en ejemplo, en el diagrama de dispersión pueden coincidir varios puntos en el mismo lugar, para que esto pueda apreciarse en el diagrama de dispersión en lugar de representar cada par como un punto, introducimos un número que señala cuantos casos hay en ese punto. Si en un punto correspondiente a un hipotético par no se escribe nada es que la frecuencia de ese punto es 0; es decir, que ningún caso ha presentado esa combinación de puntuaciones en el test X y el criterio Y.

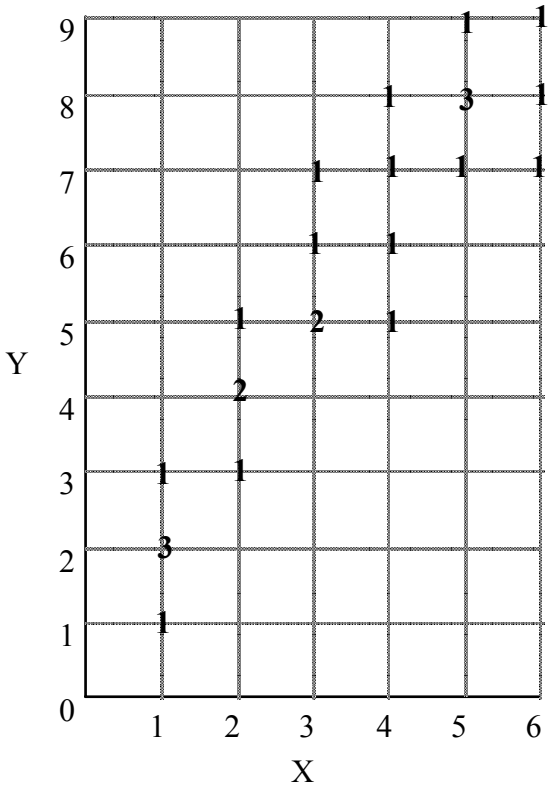


Diagrama de dispersión (también llamado diagrama de correlación) de los datos del ejemplo anterior, expresando en cada punto el número de casos que hay con esos dos valores.

En el diagrama de dispersión puede apreciarse con más claridad las características que señalábamos a partir de la matriz de datos: a) Presencia de una relación entre X e Y de modo que, en general, a puntuaciones mayores del test X corresponden puntuaciones mayores del criterio Y (y viceversa); b) Esa relación no es exacta: Para un mismo valor de X los casos pueden presentar diferentes valores de Y; y viceversa, para el mismo valor de Y pueden presentar diferentes valores de X.

Este tipo de relaciones podrán apreciarse también representando los datos mediante tablas de correlación.

Tablas de correlación

Una *tabla de correlación* es aquella que refleja una distribución bidimensional de variables cuantitativas presentando los valores de una variable X en el eje de abscisas y los valores de la otra Y en ordenadas, de modo que cada celdilla expresa la frecuencia de cada par.

Caso de ser X e Y variables continuas la tabla puede trazarse definiendo intervalos. Caso de ser X e Y variables cualitativas (de naturaleza nominal o atributos) entonces la tabla resultante se denomina *tabla de contingencia* en lugar de tabla de correlación.

Se denomina *distribución marginal* de X a la distribución de esa variable X sin tener en cuenta a la otra variable (es la distribución que hubiéramos obtenido para el test X sin necesidad de registrar el criterio).

Se denomina *distribución condicional* de una variable X para un valor dado de la variable Y, a la distribución que se obtiene de X considerando exclusivamente a aquellos que han obtenido ese valor dado de Y. Hay una distribución de X condicionada a cada valor de Y. La distribución condicional del criterio Y para un valor dado del test X ayuda a comprender que les pasa en el criterio Y alos sujetos que han sacado determinada puntuación en el test. Si a los sujetos que sacan diferentes puntuaciones en el test X les suceden cosas distintas en el criterio Y este hecho puede utilizarse para fundamentar pronósticos diferentes en Y a partir de las puntuaciones diferentes en el test.

Definiciones análogas pueden establecerse para la distribución marginal de Y y distribuciones condicionales de Y para cada valor de X.

La representación de las frecuencias conjuntas en modo de tabla de correlación es la siguiente:

	Y/X	1	2	3	4	5	6	Marg	M:X/Y
	9					1	1	2	5'5
	8				1	3	1	5	5
	7			1	1	1	1	4	4'5
	6			1	1			2	3'5
Y	**5**		1	2	1			4	3
	4		2					2	2
	3	1	1					2	1'5
	2	3						3	1
	1	1						1	1
	Marg	5	4	4	4	5	3	N=25	$\bar{X}=3'$
	M:Y/X	2	4	5'7	6'5	8	8	$\bar{Y}=5'$	

En la tabla de correlación anterior la primera fila (encabezados) expresa los valores del test X, desde 1 hasta 6. Es preferible ubicar este eje en la parte superior de la tabla para que en la parte inferior puedan expresarse algunos estadísticos.

La primera columna expresa los valores del criterio Y (desde 1 hasta 9). Los valores se han dispuesto en orden ascendente para respetar el orden cartesiano clásico y su analogía con un diagrama de correlación. La tabla de correlación dispuesta de este modo es un modo semi-gráfico de representar un diagrama de correlación.

Cada celdilla expresa el número de casos (ó frecuencia) que presenta simultáneamente los valores de X e Y correspondientes. Las celdillas vacías en el interior de la tabla se leen como ceros (es decir, que ninguno de los 25 casos de la muestra ha presentado esa combinación de valores de X e Y).

La penúltima fila, encabezada "Marg. X" expresa la distribución marginal de X. Es decir, la distribución de frecuencias de los 25 casos para los 6 valores posibles de X. (Equivale a la suma de frecuencias de las columnas).

Similarmente, la penúltima columna, encabezada "Marg. Y" expresa la distribución de frecuencias de los

25 casos para los 9 valores posibles de Y. (Equivale a la suma de frecuencias de las filas).

Cada columna de datos de la tabla expresa una distribución del criterio Y condicionada a un valor del test X determinado. Por ejemplo, la columna encabezada por 2 (X=2) expresa la distribución de Y (qué frecuencias adopta cada valor de Y) para el subconjunto de 4 casos que presentan X =2.

Similarmente, cada fila de datos de la tabla expresa una distribución del test X condicionada a determinado valor del criterio Y. Por ejemplo, la fila encabezada por 3 (Y=3) expresa de la distribución de X (qué frecuencias adopta cada valor del test X) para el subconjunto de 2 casos cuya puntuación el criterio Y es 3.

La última columna (encabezada "M:X/Y") expresa las medias de las distribuciones de X condicionadas a cada valor de Y. Por ejemplo, la distribución de X condicionada a Y=5 presenta una media de 3. Generalmente ese valor se considera la mejor representación de lo que han puntuado en el test aquellos sujetos que luego sacan un 5 en el criterio.

Del mismo modo, la última fila (encabezada "M:Y/X") expresa las medias de las distribuciones de Y condicionadas a cada valor de X. Por ejemplo, la distribución de Y condicionada a X=6 presenta una media de 8. Estas medias se consideran generalmente el mejor pronóstico posible en el criterio para un sujeto que ha sacado esa puntuación en el test. Si el sujeto obtiene un 6 en el test nuestro mejor pronóstico para el criterio es que obtendrá un 8.

La tabla permite expresar también, en las últimas celdillas, la media de X (= media de la distribución marginal de X), en este caso 3'36; y la media de Y (= media de la distribución marginal de Y) en este ejemplo 5´56.

Si nuestro objetivo es explicar o pronosticar los valores del criterio Y, la consecuencia más importante de un test X que actúa como predictor es que en lugar de disponer de una sola distribución global del criterio Y (la distribución marginal del criterio Y), ahora podemos descomponer ésta en un conjunto de distribuciones condicionales (las distribuciones del criterio Y condicionadas a cada valor del test X que deseemos considerar separadamente). De ese modo podemos estudiar si el criterio Y distribuye igual o no en cada nivel del test X. De ese modo podemos observar si el criterio Y va variando o no según en que puntuación del test X estemos. De ese modo las tablas de correlación nos facilitan el estudio de la relación entre test y criterio.

1.2. El coeficiente de validez como coeficiente de correlación de Pearson

El coeficiente de validez se estima como la correlación entre test y criterio. A su vez, el coeficiente de correlación de Pearson puede ser deducido por diferentes caminos y presentarse bajo una diversidad de fórmulas. Dada la importancia básica de este estadístico conviene familiarizarse con algunas de estas posibles expresiones equivalentes entre sí.

Primera expresión general del coeficiente de correlación de Pearson. El coeficiente de correlación de Pearson es igual a la media del producto de las puntuaciones típicas de dos variables.

$$r_{xy} = \frac{\sum z_x z_y}{N}$$

Es una fórmula poco indicada para el cálculo, pero que ayuda a comprender el significado de la correlación. El coeficiente de correlación de Pearson es igual al promedio de los productos de las puntuaciones típicas caso a caso.

La fórmula se puede escribir más brevemente:

$$r_{xy} = \overline{z_x z_y}$$

En las fórmulas anteriores tenemos que:

$$z_x = \frac{X - \bar{X}}{s_x}$$

$$z_y = \frac{Y - \bar{Y}}{s_y}$$

que es la definición de puntuación típica de X y puntuación típica de Y, respectivamente. Si hay que calcular el coeficiente de validez en un caso práctico esta fórmula no es adecuada debido a que implica obtener las puntuaciones típicas, lo que es tedioso y comporta dificultades de redondeo decimal.

Segunda expresión general del coeficiente de correlación de Pearson. *El Coeficiente de Correlación de Pearson puede definirse como la covarianza entre X e Y partida por el producto de sus desviaciones típicas.*

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$$

Es quizás la forma más popular de presentar r_{xy} :

En esta expresión el término del denominador es la covarianza entre X e Y. La covarianza, a su vez, es el sumatorio de los productos de las puntuaciones diferenciales partido por el número de casos N:

$$s_{xy} = \frac{\sum x\,y}{N}$$

siendo $x = X - \bar{X}$ e $y = Y - \bar{Y}$ definición de las puntuaciones diferenciales de X y de Y, respectivamente.

Tercera expresión general del coeficiente de correlación de Pearson.

Esta es mi expresión favorita para el coeficiente. Deduje esta forma buscando una expresión que fuera máximamente eficiente para las calculadoras de sobremesa que sólo daban medias y desviaciones típicas y que fueron durante mucho tiempo las más populares entre los estudiantes de psicología. Pero no sólo es práctica para esas calculadoras, es una forma elegante y compacta, y fácil de recordar. El Coeficiente de correlación de Pearson es igual a la media de los productos cruzados menos el producto de las medias, dividido por el producto de las desviaciones típicas.

$$r_{xy} = \frac{\overline{XY} - \bar{X} \cdot \bar{Y}}{s_x \cdot s_y}$$

En esta fórmula $\overline{XY}$,representa la media del producto entre X e Y en puntuaciones directas. Es decir:

$$\overline{XY} = \frac{\sum X_i Y_i}{N}$$

Al utilizar sólo medias y desviaciones típicas evita casi totalmente el manejo de decimales y signos negativos

propio de las fórmulas que utilizan puntuaciones diferenciales o típicas.

Comparando esta expresión con la segunda resulta obvió que la covarianza también puede definirse como *la media de los productos menos el producto de las medias,* y escribirse así:

$$s_{xy} = \overline{XY} - \bar{X} \cdot \bar{Y}$$

fórmula esta última muy adecuada para el cálculo manual de la covarianza.

Todas las fórmulas anteriores del coeficiente de correlación de Pearson son equivalentes entre sí. Si la psicóloga o psicólogo tiene que estimar el coeficiente de validez de un test mediante el coeficiente de correlación de Pearson puede utilizar cualquier programa informático o calculadora que lo provea, si todavía ha de utilizar una calculadora que no provea este coeficiente la tercera fórmula es sin duda la más indicada. Una vez que el cálculo de este coeficiente está disponible de modo automatizado cobran más interés estas tres fórmulas debido a que aportan comprensión sobre los significados del coeficiente.

Interpretación del coeficiente de validez como coeficiente de correlación de Pearson

¿Qué indica el coeficiente de validez?

Este coeficiente indica el *grado de asociación lineal entre dos variables,, un text y un criterio*.

Un test X y un criterio Y están asociados linealmente si el comportamiento del criterio Y puede describirse adecuadamente a partir de las puntuaciones en el test X utilizando una ecuación lineal, es decir, una ecuación de la recta del tipo:

$$Y' = a + bX$$

En esta ecuación los valores del criterio Y son pronosticados a partir de los valores del test X al multiplicar estos últimos por una constante b y sumarles otra constante a .

El análisis que permite determinar estas constantes de forma que obtengamos la mejor predicción posible de Y a partir de X se conoce como *análisis de regresión lineal simple.*

Dos variables están asociadas linealmente en la medida en que pueda encontrarse una función lineal (es decir un par de valores para a y b) que permita predecir una variable a partir de la otra con el mínimo error posible.

¿Cuál es el rango de valores que puede adoptar?

El coeficiente de correlación de Pearson *toma valores entre -1 y +1*. Por tanto el coeficiente de validez puede tomar valores entre –1 y +1.

¿Qué significan esos valores?

-1 expresa una *relación lineal negativa perfecta*: cuanto más puntúa el sujeto en el test X menos puntúa en el criterio Y.

Si el coeficiente da -1 entonces la relación negativa es perfecta: Esto supone que se puede pronosticar perfectamente lo que sucede en le criterio Y a partir del test X mediante una función lineal en la que b tiene signo negativo.

A medida que el coeficiente se aproxima a -1 la correlación expresa un mayor grado de asociación lineal entre las variables.

Mientras el coeficiente presente signo negativo indica que, en general, las dos variables presentan una relación de signo negativo más o menos intensa. Una relación de signo negativo significa que en general valores relativamente mayores del test X están asociados con valores relativamente menores del criterio Y. Qué a medida que se avanza en el eje de X hacia valores mayores tendemos a encontrar valores de Y más bajos.

0 expresa una *ausencia de relación lineal*.

Un coeficiente de correlación de Pearson de 0 no significa necesariamente que el test X y el criterio Y no están relacionados, pero si que *no están relacionadas linealmente*, es decir, que su relación si existe no puede describirse con la ecuación de la recta. Un coeficiente de correlación de Pearson de 0 significa que X e Y no están relacionadas en

absoluto o bien que su relación no puede describirse adecuadamente con una linea recta, aunque quizás exista entre ellas una relación no lineal, es decir, que no pueda describirse con una recta. De hecho algunas relaciones curvilíneas perfectas suponen necesariamente que el coeficiente de correlación de Pearson valga 0.

+1 expresa una *relación lineal positiva perfecta*: cuanto mayor el valor de la puntuación en el test X mayor el valor del criterio Y.

Una correlación de +1 indica que el criterio Y puede pronosticarse sin error a partir de las puntuaciones en el test utilizando una función lineal en la que b tiene signo positivo.

A medida que la correlación se aleja de 0 y se acerca a +1 tenemos un mayor grado de asociación lineal de signo positivo entre test y criterio.

Cuando test y criterio presentan una correlación de Pearson moderada a alta (positiva o negativa) se dice que las variables están correlacionadas o correlacionan, y entonces se podrá utilizar el test X para predecir o estimar la puntuación en el criterio Y mediante regresión lineal simple.

En la práctica con datos reales correlaciones perfectas de +1 ó -1 son bien poco usuales, podría decirse que en términos de tests y criterios no puede citarse un solo ejemplo.

Para interpretar el significado de una correlación distinta de 0 situada entre -1 y +1 es necesario considerar una serie de factores que incluyen el tamaño de la muestra, la naturaleza del test X y del criterio Y y los resultados previos que se hayan obtenido para la relación entre ese tipo de variables en estudios de validación anteriores. Cualquier indicación general sobre lo que es una correlación baja, moderada o alta es sólo una primera aproximación que tiene que tener en cuenta estos factores para tomar sentido.

La interpretación sustantiva, en términos de su significado psicológico, de un coeficiente de correlación depende fuertemente de los resultados que se conozcan para esas variables en ese campo de investigación y de aplicación profesional.

Coeficiente de correlación y relaciones curvilíneas

Tanto el coeficiente de correlación de Pearson como la covarianza o la regresión lineal, sólo recogen el grado de asociación o covariación *lineal.* La covarianza y el coeficiente de correlación de Pearson expresan el grado de asociación (dependencia) *lineal* entre las dos variables. Sólo reflejarán parcialmente relaciones no lineales en la medida en que éstas se puedan describir de un modo aproximado por una función lineal, lo que sucede para ciertas curvas, o para ciertos tramos de curvas. Pero hipotéticamente un test X y un criterio Y podrían presentar una dependencia perfecta no-lineal y que ésta no sea reflejada en absoluto por la covarianza, ni por el coeficiente de correlación de Pearson, ni por la regresión lineal simple.

Por ejemplo, en los datos siguientes:

Caso:	X	Y
1	5	25
2	4	16
3	3	9
4	2	4
5	1	1
6	-1	1
7	-2	4
8	-3	9
9	-4	16
10	-5	25

existe una correlación curvilínea perfecta entre X e Y dado que $Y=X^2$, pero el coeficiente de correlación de Pearson vale 0.

Sin embargo, si calculamos el coeficiente de correlación de Pearson solo para X>0 entonces el coeficiente de correlación de Pearson vale 0'98, indicando que un brazo de la parábola sí puede ser descrito aproximadamente por el coeficiente, en la medida en que puede ser aproximado por una recta. En el brazo de la parábola para X<0 el coeficiente de correlación es el mismo pero con signo opuesto y la aproximación lineal a la curva igualmente buena.

Evidentemente una relación del tipo $Y=X^2$,definida para X>0 no invita a calcular el coeficiente de correlación de Pearson -que está orientado a relaciones lineales- pero deberíamos

reconocer que es una buena aproximación. Muy raramente nuestros datos empíricos en las relaciones entre tests y criterios se comportan tan bien, así que la inadecuación en este caso es relativa.

Coeficiente de correlación y distribuciones asimétricas

El coeficiente de correlación de Pearson detecta el grado de asociación incluso en distribuciones asimétricas, pero debe observarse el efecto de los puntos más alejados del centro de su distribución en el coeficiente .

Por ejemplo, en la distribución (X,Y)=(1,2) (1,2) (1,2) (1,2) (1,2) (2,1) (2,1) (2,1) (3,0) (4,-1) el coeficiente de correlación de Pearson vale -1.

Caso:	X	Y
1	1	2
2	1	2
3	1	2
4	1	2
5	1	2
6	2	1
7	2	1
8	2	1
9	3	0
10	4	-1

Sin embargo, en estas distribuciones hay que tener muy en cuenta el comportamiento de los datos en las colas, especialmente si son prolongadas. Por ejemplo si en la distribución anterior cambiamos en el primer par el valor de Y sumándole 2 puntos, tal que el primer par sea (1,4), entonces el coeficiente de correlación aumenta aproximadamente +0'11 y resulta igual a -0'89. Pero si introducimos un cambio de la misma magnitud (sumar 2 puntos) en el valor de Y de la otra cola tal que el último par sea (4,1) entonces el coeficiente de correlación de Pearson cambia +0'2 y resulta ser -0'8.

Coeficiente de correlación y valores extremos

La presencia de valores extremos por sí misma no distorsiona el coeficiente, siempre que estos se comporten conforme a la relación lineal entre X e Y. Pero hay que estar muy atentos a su presencia y su comportamiento pues un solo par de valores extremos puede llevar a conclusiones muy erróneas: puede hacer ver que hay correlación donde no la hay o puede hacer ver que no hay correlación donde la hay. Cuando estudiamos relaciones de validez la presencia de algún o algunos casos atípicos, (por ejemplo con aptitudes X o rendimientos Y extraordinariamente bajos o altos, o puntuaciones muy bajas en el test X por razones circunstanciales como la falta de cuidado y atención al realizarlo) pueden producir distorsiones importantes en el coeficiente de validez.

Por ejemplo, en los datos siguientes:

Caso:	X	Y
1	2	3
2	2	3
3	2	2
4	2	2
5	2,5	2,5
6	3	3
7	3	3
8	3	2
9	3	2
10	2,5	2,5

no hay ninguna relación lineal entre X e Y , y, el coeficiente de correlación de Pearson, que vale 0 para estos datos, lo refleja perfectamente.

Pero basta que cambiemos el par del último caso por (5,25) para que el coeficiente de correlación de Pearson cambie a 0'857. Un cambio espectacular producido por el valor extremo 25 definido en Y. De todas formas no es necesario que el valor sea tan extremo para inducir a error, si el último par es simplemente (5,5) el coeficiente todavía valdría 0'738 indicando una relación lineal "producida" por un solo caso. Por eso los valores extremos pueden hacer "aparecer" una

correlación importante entre test y criterio por sí solos, independientemente de que el resto de la distribución no presente relación lineal alguna.

Veamos el fenómeno inverso. Sean los datos:

Caso:	X	Y
1	0	1,5
2	0,5	2
3	1	2,5
4	1,3	2,8
5	2,5	4
6	3	4,5
7	0,2	1,7
8	0,7	2,2
9	1	2,5
10	5	6,5

La correlación de Pearson es perfecta, igual a 1, dado que hay una relación lineal perfecta Y=1'5+X. Aunque el último par presenta unas puntuaciones alejadas del centro de la distribución tanto en el test X como en el criterio Y, obsérvese que la correlación no depende especialmente de ese último par. Si se suprime el último par la correlación sigue siendo 1. Aunque se trate de valores fuertemente extremos siempre que mantengan la misma relación entre ellos que el resto de los pares estos valores extremos no distorsionan el coeficiente de correlación de Pearson.

Pero, si sustituimos el último par por (5,0) la correlación da un vuelco espectacular y se sitúa en -0'06, por ese solo caso, indicando una ausencia de relación. Conviene por tanto *siempre representar gráficamente mediante diagramas de dispersión* y estar atento al impacto de los datos alejados de las respectivas medias en una o en ambas variables, así como a aquellos que sin adoptar valores necesariamente extremos se comportan "contra" la función que relaciona a los demás.

Sugiriendo un estadístico de influencia de cada par en el coeficiente de validez

Los análisis anteriores suponen que hay que estar atentos a la influencia que ejerce cada par en el coeficiente de validez. Especialmente si el par o pares ocupa una posición particular en el diagrama de dispersión, alejado del resto de la nube de puntos.

De modo sencillo para evaluar la influencia de un par sobre el coeficiente de validez puede calcularse la correlación excluido ese par y comparar el resultado con la correlación

original incluido el par. La diferencia entre ambas indica la influencia del par.

Si simbolizamos el par concreto (X, Y) del caso i simplemente por i, podemos llamar influencia de en la correlación r_{xy} al cambio de esa correlación, que simbolizaremos Δr_{xy} al excluir el par i del cálculo de la misma $r_{xy(-i)}$:

$$\Delta r_{xy} = r_{xy} - r_{xy(-i)}$$

La diferencia entre la correlación de Pearson con todos los datos y la correlación excluido un par, es un sencillo y fácilmente inteligible indicador de la influencia que ejerce el par en el coeficiente de validez entre el test X y el criterio Y. Si no hubiera estado ese par la correlación habría cambiado tanto como expresa la diferencia.

Esta influencia así medida depende de N, en el sentido de que es inversamente proporcional al tamaño de la muestra. Pero esta es una característica aceptable y realista en este sencillo estadístico de influencia.

Por otra parte este procedimiento es fácilmente generalizable a la exclusión de grupos de casos.

Coeficiente de validez y heterogeneidad de la muestra

Si los datos son muy homogéneos desde el punto de vista de la varianza de X y de Y las relaciones lineales que puedan existir es posible que no se muestren con demasiada claridad.

El rango de la variable delimita cual es la máxima varianza posible, de modo que decisiones de rango sobre el test y el criterio pueden afectar los resultados en términos de la validez estimada.

Por ejemplo, sea Y una conducta medida en escala de 0 a 10, y X la opinión sobre cierto tema supuestamente relacionado con la conducta Y. Decidimos evaluar X en una escala de 7 puntos como la siguiente:

3 = Muy de acuerdo

2 = Bastante de acuerdo

1 = Algo de acuerdo

0 = Indiferente, ni de acuerdo ni en desacuerdo.

-1 = Algo en desacuerdo

-2 = Bastante en desacuerdo

-3 = Muy en desacuerdo

Supongamos que el grupo es muy heterogéneo, formado tanto por personas con una actitud positiva o muy positiva, como negativa o muy negativa ante esa cuestión.

Los resultados podrían ser estos:

Caso:	X	Y
1	2	9
2	3	9
3	2	8
4	3	8
5	2	9
6	-3	2
7	-2	1
8	-1	1
9	-3	0
10	-3	2

El coeficiente de correlación de Pearson muestra una fuerte relación lineal entre opinión y conducta. alcanzando un valor de 0'942. Como puede observarse en la tabla de datos las personas con opinión favorable o muy favorable en X presentan puntuaciones altas en la conducta Y, mientras que las personas con opinión desfavorable o muy desfavorable presentan bajas puntuaciones en la conducta Y. La dependencia de Y respecto de X no es perfecta, pero la relación es fuerte y clara.

Si el lector observa la tabla comprobará que los 5 primeros sujetos tienen opiniones favorables y los 5 últimos opiniones desfavorables ¿Qué hubiera pasado

si en lugar de una muestra heterogénea hubiéramos obtenido una muestra homogénea en X con solo los 5 casos que puntúan X>0? La correlación en ese caso es -0'16. El efecto general de relación entre X e Y se hubiera esfumado y en su lugar tendríamos una irrelevante relación de signo negativo.

¿Qué hubiera pasado si en lugar de una muestra heterogénea hubiéramos obtenido una muestra homogénea en X con solo los 5 casos que puntúan X<0? La correlación en ese caso es -0'2. Como en el caso anterior obtenemos una relación negativa (de poca magnitud e importancia dado que solo estamos trabajando con 5 casos).

Por añadidura el lector puede comprobar que la correlación en la muestra total no es precisamente el promedio de las correlaciones de las submuestras. De hecho puede haber virtualmente cualquier relación entre la correlación en una submuestra y la correlación en la muestra total, con tal que se seleccionen los datos intencionadamente. (Esta situación es distinta a la de dos muestras de suficiente tamaño obtenidas aleatoriamente de una población de la que pueden considerarse representativas).

Desde luego si una selección de casos afecta a X o a Y puede afectar a la correlación entre X e Y.

El ejemplo ilustra claramente como es necesario que se presente variabilidad en el test X para poder detectar una relación con un criterio Y. Podría ponerse un ejemplo análogo con el criterio Y fácilmente. Basta cambiar la etiqueta de las columnas del caso anterior y pensar que se ha seleccionado en base al criterio.

Coeficiente de validez y heterogeneidad de la muestra debida a tercera variable

Pero la relación entre homogeneidad - heterogeneidad de la muestra y coeficiente de validez no es siempre la del ejemplo anterior. Es decir, no siempre una mayor heterogeneidad de la muestra conduce a poner de manifiesto relaciones lineales que no se encuentran en muestras homogéneas. También puede suceder todo lo contrario. Si la muestra contiene casos que pertenecen a subpoblaciones definidas por una tercera variable, tal que en cada subpoblación la relación entre el test X y el criterio Y es distinta, tomar una muestra heterogénea conteniendo casos de todas las subpoblaciones y tratarla conjuntamente llevará a no detectar adecuadamente esas relaciones.

Supongamos de nuevo 10 casos:

Caso:	X	Y
1	0	10
2	2	8
3	5	5
4	9	1
5	7	4
6	0	1
7	1	3
8	5	7
9	7	9
10	8	10

Para ilustrar el ejemplo supongamos que X es el nivel de activación medido por cierto test e Y un criterio de rendimiento asociado a una respuesta de una habilidad muy específica, de modo que a más alto X más nivel de activación y a más alto Y más rendimiento. La correlación entre ambos es nula: el coeficiente de validez resulta ser 0'036. Pero esto puede ser un efecto de la heterogeneidad de la muestra. Es decir, la heterogeneidad de la muestra respecto a una tercera variable que afecte a la relación entre el test X y el criterio Y puede enmascarar la relación.

En efecto, supongamos que los 5 primeros casos son sujetos inexpertos y que los 5 últimos son expertos. ¿Qué relación hay entre X e Y para los inexpertos? La correlación es -0'99 mostrando una fuerte relación de signo negativo. Es decir, los inexpertos presentan mejores rendimientos cuanto mas tranquilos están (se trata sólo ejemplo con datos figurados). ¿Qué relación hay para los expertos? Tomando los últimos 5 casos por separado la relación es 0'99. También se trata de una relación lineal muy fuerte, pero ahora de signo positivo Es decir, los expertos mejoran su rendimiento a medida que aumenta su activación (insisto en que solo es un ejemplo figurado). Lo importante es advertir como una tercera variable Z, en este ejemplo el grado de experiencia de los sujetos, puede modular la relación entre test y criterio de tal modo que una muestra heterogénea en esa tercera variable Z puede enmascarar la relación entre test y criterio y distorsionarla fuertemente. En realidad esa distorsión puede ser de diversas formas, según sea la forma de la relación de las puntuaciones del test X y del criterio Y en cada grupo de la tercera variable Z.

Si la heterogeneidad en X e Y puede ser necesaria para poder observar relaciones que de otro modo no aparecen, la heterogeneidad de la muestra debida a una tercera

variable Z puede afectar también considerablemente, pero no en el sentido de revelar la correlación sino de enmascararla. El problema es más complejo si, a su vez Z está correlacionada con X, con Y o con ambas.

Coeficiente de validez y restricción del rango

En general, si test X y criterio Y están correlacionados y se reduce el rango de una de ellas (o de ambas) entonces la correlación puede tender a disminuir. La correlación muestra el grado en que las variables *varían* juntas. La correlación se basa pues en la variabilidad de las variables, en aquella variabilidad de las variables que *co-varía* conjuntamente, si ésta se restringe entonces el coeficiente de validez tiende a disminuir.

Sucede una restricción de rango cuando se introduce una selección de casos que limita la aparición de la variabilidad de una o ambas variables. Por ejemplo, si se estudia la relación entre un test X y un criterio Y considerando solo los casos que superan cierta puntuación en el test, la correlación será en general más baja que la que se habría obtenido considerando todo el rango posible de puntuaciones. Esta situación es frecuente en selección de personal donde una parte de los sujetos no están cuando se puede medir el criterio Y precisamente porque sacaron ciertas puntuaciones (altas o bajas según el caso) en el test X. Aunque la selección sea indirecta, a través de una tercera variable Z, si Z está de algún modo correlacionada con X ó Y, una selección en Z puede inducir una restricción del rango en X ó Y que afectará al coeficiente de correlación, en general, tendiendo a disminuirlo.

También se reduce la variabilidad si una o ambas variables se policotomizan (se agrupan sus puntuaciones en un número h de intervalos separados por un número h-1 de puntos de corte) y todavía más cuando se dicotomiza (se agrupan todas las puntuaciones en dos únicos intervalos separados por un punto de corte). Si la variabilidad que se reduce covariaba con la de la otra variable, entonces el coeficiente de correlación de Pearson será menor en las variables policotomizadas o dicotomizadas.

Pero no siempre ni necesariamente cuando se restringe el rango de una o ambas variables el coeficiente de correlación disminuye. Si en la zona o modo de rango que no se excluye existe una relación lineal más clara, o simplemente no contradictoria, con la que había en la zona o modo de rango excluidos, la correlación puede aumentar al restringir el rango.

Supongamos que X es una prueba de admisión puntuada de 0 a 10 e Y la nota al final del curso, también en escala de 0 a 10. Con los siguientes datos:

Caso:	X	Y
1	2	4
2	6	5
3	10	7
4	9	7
5	7	8
6	0	3
7	1	0
8	5	4
9	7	9
10	4	5

Aquí la prueba de admisión se ha usado para conocer su funcionamiento, sin seleccionar a los sujetos, dejando que todos acaben el curso para ver si podemos confiar en la prueba de admisión. La prueba de admisión funciona bastante razonablemente aunque es imperfecta. La correlación es 0'809.

¿Que pasaría si dicotomizamos la prueba de admisión en aprobados 1 y suspensos 0 por el punto de corte habitual de 5?

Los datos quedarían así:

Caso:	X'	Y
1	0	4
2	1	5
3	1	7
4	1	7
5	1	8
6	0	3
7	0	0
8	1	4
9	1	9
10	0	5

La correlación calculada sobre esos datos es 0'712. Se ha producido una perdida sensible en la capacidad de expresar la relación entre las variables. Al dicotomizar X por X=5 cualquier variación en X entre 0 y 4 se pierde y cualquier variación en X entre 5 y 10 se pierde también. Si esa variación en el test X estaba asociada a variación en el criterio Y entonces la perdida repercute reduciendo -como ha sucedido en el ejemplo- el coeficiente de validez. Por decirlo de un modo didáctico, se ha perdido variabilidad dentro de cada intervalo de X que covariaba con la variabilidad de Y. Así para X=6 teníamos Y=5 y para X=10 Y=7. La relación que muestran estos dos pares se ha perdido

al dicotomizar X; ahora solo tenemos X=1 con Y=5 y X=1 con Y=7.

Una nota sobre la escala en que se expresa la variable dicotomizada

Quizás el lector pueda creer que esto es debido a que hemos utilizado como valores para la parte superior de la escala 1 y para la inferior 0, y que las cosas serían mejores si en lugar de 0 utilizáramos para el segmento inferior la marca de clase del intervalo 0-4, que es 2, y para el tramo superior la marca de clase del intervalo 5-10 que es 7'5.

En este caso los datos quedarían así:

Caso:	X'	Y
1	2	4
2	7,5	5
3	7,5	7
4	7,5	7
5	7,5	8
6	2	3
7	2	0
8	7,5	4
9	7,5	9
10	2	5

Sin embargo, la correlación también es 0'71. Como antes con 0-1. ¿Por qué da igual utilizar 0-1 ó 2-7'5 en la variable dicotomizada X'? La razón es que, tratándose de una dicotomizada, donde sólo hay dos valores, cualquier par de valores siempre se puede expresar como función lineal de los primeros. De modo que los valores 2-7'5 sólo son los valores 0-1 transformados mediante la transformación lineal X"=2+7,5X'. La transformada lineal X" de una variable X' correlaciona con cualquier variable Y lo mismo que correlacionara X'. (Si se tratara de una tricotomización y necesitáramos por tanto tres puntuaciones puede que ya no daría igual cualquier trío de valores dado que tres puntos pueden o no estar todos sobre una recta aunque dos necesariamente lo están siempre).

Pero, volvamos a la cuestión central de la dicotomización y su efecto en el coeficiente de validez. Si también dicotomizamos Y tomando el 5 como punto de corte, los datos quedarían del siguiente modo:

Caso:	X	Y
1	0	0
2	1	1
3	1	1
4	1	1
5	1	1
6	0	0
7	0	0
8	1	0
9	1	1
10	0	0

En este caso la correlación vuelve a 0'816. Ello es debido a que se ha producido en Y una reducción del "ruido" o variación interna al intervalo que no varía sistemáticamente con la de X. Este fenómeno, tratándose de dicotomizaciones puede entenderse que depende parcialmente de la proporción de falsos pronósticos que refleje la tabla después de dicotomizada.

Pero en otras circunstancias las dicotomizaciones pueden producir efectos más drásticos en el sentido de reducir el coeficiente de validez.

Por ejemplo, las siguientes puntuaciones en un test X y un criterio Y correlacionan 0'909

Caso:	X	Y
1	2	0
2	6	7
3	10	9
4	0	1
5	7	7
6	4	5
7	4	5
8	5	4
9	5	4
10	4	5

Pero después de dicotomizar test y criterio en "Apto"=1 (para puntuaciones iguales o mayores que 5) ó "No-apto"=0, sólo correlacionan 0'167.

Caso:	X'	Y'
1	0	0
2	1	1
3	1	1
4	1	1
5	1	1
6	0	0
7	0	1
8	1	0
9	1	0
10	0	1

El efecto de las dicotomizaciones y policotomizaciones depende de la forma concreta que adopten los datos. En general, si reducen variabilidad que covaría con la otra variable tienden a disminuir el coeficiente de correlación; pero, si reducen variabilidad que no covaría con la otra variable pueden mantenerlo o incluso mejorarlo.

No es cierto que toda dicotomización o policotomización disminuya necesariamente el coeficiente de correlación de Pearson, pero ese es el efecto esperable más frecuentemente si los datos mostraban una buena relación. Por tanto no es cierto que toda dicotomización o policotomización reduzca el coeficiente de validez de un test, aunque este es un efecto frecuente si los datos mostraban razonablemente una relación lineal antes de la dicotomización o policotomización.

Coeficiente de validez y transformaciones lineales

Sea la transformación lineal X'=a+bX. Una variable X y su transformada lineal X' siempre presentan una correlación perfecta entre sí de 1 si b en la ecuación es positivo, ó de -1 si b es negativo. Si b fuera igual a 0 la transformada se convierte en una constante igual a "a", y por tanto no puede correlacionar con ninguna variable. Las constantes no pueden correlacionarse porque no presentan variación; por definición la correlación es de las variables.

Por tanto, en general, si X' es una transformación lineal de X tal que b no es igual a 0, entonces:

$$|r_{X\,X'}| = 1$$

Por ello, las transformaciones lineales donde b tiene signo positivo tienen la importante propiedad de no alterar las correlaciones lineales con otras variables. Es decir, el coeficiente de correlación de Pearson de X con cualquier variable Y no se ve alterado por sustituir X por ninguna transformación lineal de la variable X (lo mismo puede decirse de las transformadas lineales de Y, ó de ambas simultáneamente) donde b tenga signo positivo. Si b tiene

signo negativo entonces la correlación de la variable transformada X' con otra Y adoptará el mismo valor pero con signo contrario que el de X con Y.

Por ejemplo, si la correlación del test X con el criterio Y es c, entonces la correlación de X' con Y es también c siempre que b>0.

Si transformamos Y en Y' mediante cualquier transformación lineal, entonces, la correlación entre X e Y' seguirá siendo c, siempre que b>0.

La correlación entre X' e Y' transformadas ambas también será c, siempre que ambos b>0 o que ambos b<0.

Aunque X' e Y' fueran una o ambas vueltas a transformar mediante transformaciones lineales su correlación siempre seguiría siendo c, siempre teniendo en cuenta que en la ecuación de transformación lineal si b es negativo la correlación invierte el signo original.

El coeficiente de determinación R^2

Un coeficiente de correlación de Pearson elevado al cuadrado se denomina *coeficiente de determinación* R^2 y expresa que proporción de la varianza de una variable es atribuible o está asociada a la otra.

El coeficiente de determinación ayuda a interpretar el significado de una correlación. Si por ejemplo el coeficiente de validez del test X con el criterio Y es 0'5, ello significa que el test X explica el 25% de la varianza del criterio Y.

R^2 va entre 0 (ausencia de relación lineal) y 1 (un ajuste perfecto a una relación lineal, sea ésta de signo positivo o negativo). R^2 también expresa el grado de aproximación de los puntos a la recta de regresión (Amón, 1979). A medida que los valores reales de Y son más cercanos a los valores pronosticados Y' por la ecuación de regresión lineal, el coeficiente de determinación se aproxima a 1.

Coeficiente de validez y causalidad

Es posible que le criterio Y dependa perfectamente del test X y no viceversa (en una relación no lineal). La regresión,

incluso lineal de Y sobre X no es lo mismo que la de X sobre Y (aunque estén relacionadas).

Sin embargo, si X correlaciona 0'75 con Y entonces necesariamente Y correlaciona 0'75 con X. A diferencia de la dependencia o la regresión, el coeficiente de correlación no está direccionado de una variable hacia otra variable. Es decir, es lo mismo la correlación de X con Y que la de Y con X. Sin embargo, la interpretación puede que si esté direccionada en función de la naturaleza de las variables. Generalmente en el caso de coeficientes de validez la interpretación razonable es la que considera al test como antecedente y al criterio cono consecuente, pero esa relación no está garantizada por un resultado estadístico.

La correlación solo muestra que las cosas varían juntas. Si hay pocos casos o la extracción de los mismos es dudosa o sesgada, esa variación conjunta puede ser accidental y llevar a falsas conclusiones. Pero aun si hay muchos casos disponibles formando una buena muestra hay que ser muy precavidos a la hora de extraer conclusiones a partir del hecho de que un test y un criterio varíen concomitantemente.

Un coeficiente de validez moderado o importante no implica que, necesariamente, X es causa de Y. Si X e Y correlacionan es posible que:

1. X sea causa de Y.

2. Y sea causa de X.

3. Ambas sean causas de una tercera variable.

Adicionalmente, es posible que ambas se causen mutuamente, y

también es posible que X sea causa de Y o viceversa pero solo de modo indirecto, a través del efecto de terceras variables.

Incluso en los ejemplos en los que parece no haber duda sobre la dirección de la causalidad, si se analiza con más profundidad surgen dificultades. Tomemos el ejemplo de la correlación entre edad y una aptitud. Es obvio que la aptitud no causa la edad, pero ¿es lícito pensar que la edad es causa de la aptitud? Esto también parece discutible por muchas razones. Seguramente es más acertado pensar que hay determinados aspectos orgánicos que necesitan

tiempo para desarrollarse y que el tiempo da oportunidad a que se den las circunstancias que favorecen ciertos aprendizajes y que, quizá, ayuden al desarrollo de la aptitud. Parece más razonable considerar que la edad (o si se prefiere el tiempo transcurrido) no es propiamente causa, sino una condición concomitante que favorece o permite el desarrollo de las causas. Además, la relación variará sustancialmente según se acote el rango de la muestra en edad y en aptitud.

La relación entre un test y un criterio puede ser tan compleja que el simple coeficiente de correlación, por sí, no podría interpretarse más que como una descripción de cierto grado de covariación lineal observada. Una covariación que puede ser fruto de múltiples relaciones causales y concomitantes, débiles y fuertes, no necesariamente del mismo signo, intrincadamente relacionadas entre sí.

Hay causas que actúan en serie y causas que actúan en paralelo, causas suficientes y causas necesarias, causas unidireccionales y causas bidireccionales, causas inmediatas y causas de efectos tardíos, causas de efectos constantes y causas de efectos puntuales, causas de resultados deterministas y causas de resultados probabilísticos, y, además, frecuentemente las causas interactúan en sistemas de causas más o menos complejos. El concepto de causalidad es extremadamente rico y complejo y el coeficiente de validez tan solo un estadístico muy simple.

Está claro pues que la presencia de correlación no indica necesariamente causalidad, pero *¿la ausencia de correlación indica ausencia de causalidad?* El tratamiento clásico de la cuestión considera la presencia de covariación entre las variables (y, por tanto, en el caso lineal, necesariamente la presencia de correlación) como una de las condiciones necesarias para establecer causalidad. Desde ese punto de partida se ha venido considerando la ausencia de correlación como una indicación de ausencia de causalidad. En general se considera que si X e Y presentan una correlación nula entonces X no es causa de Y e Y no es causa de X. Se considera que la correlación no sirve para probar la causalidad, pero la ausencia de correlación puede utilizarse para descartar la presencia de causalidad. En mi opinión este punto de vista debe ser matizado cuidadosamente.

La ausencia de correlación observada, puede ser un indicio de ausencia de causalidad, pero no significa tampoco, *necesariamente,*

ausencia de causalidad. Esto sería cierto, bajo ciertas condiciones de observación, en un sistema simple donde sólo actuasen X e Y, bajo ciertas acepciones de causalidad. Pero en un contexto complejo, donde otras muchas causas pueden estar actuando a la vez, no necesariamente es así. X e Y podrían correlacionar 0 y, sin embargo, X ser una causa de Y, incluso una causa con una muy fuerte relación lineal, pero dentro de un sistema complejo de causas capaz de enmascarar el efecto cuando la tercera o terceras variables Z responsables del enmascaramiento no han sido identificadas y consideradas en el análisis.

Si el lector vuelve a leer el ejemplo del rendimiento de expertos y no expertos desde esta óptica puede observar una de las situaciones en que puede darse esta problemática. Allí aparentemente X e Y no están relacionadas, hasta que se considera Z, y se analizan las relaciones de X e Y separadamente para cada nivel de Z. Si X pudiera leerse como una causa de Y queda con ello ejemplificado como *la ausencia de correlación observada no implica una ausencia de causalidad.* Esta conclusión es importante y contradice la doctrina oficial que emanan algunos textos.

El problema de la posible tercera variable Z no identificada que afecta a la relación es prácticamente insoluble fuera de las condiciones experimentales, a pesar de que se han desarrollado técnicas muy sofisticadas para tratar con *modelos causales* que tratan de aproximar una solución al mismo. Por otra parte esta cuestión lleva al problema de las llamadas *variables moduladoras*: aquellas que alteran la relación entre otras dos sin ser un predictor de la dependiente. Tanto los modelos causales como las variables moduladoras están fuera de nuestros objetivos en este texto.

Interpretación de diagramas de dispersión

El **diagrama de dispersión** es un gráfico esencial para estudiar la relación entre un test X y un criterio Y Consiste en disponer en ejes cartesianos las puntuaciones en el test X (generalmente en abscisas, eje horizontal) y los valores del criterio Y (generalmente en ordenadas, eje vertical) y representar los casos como

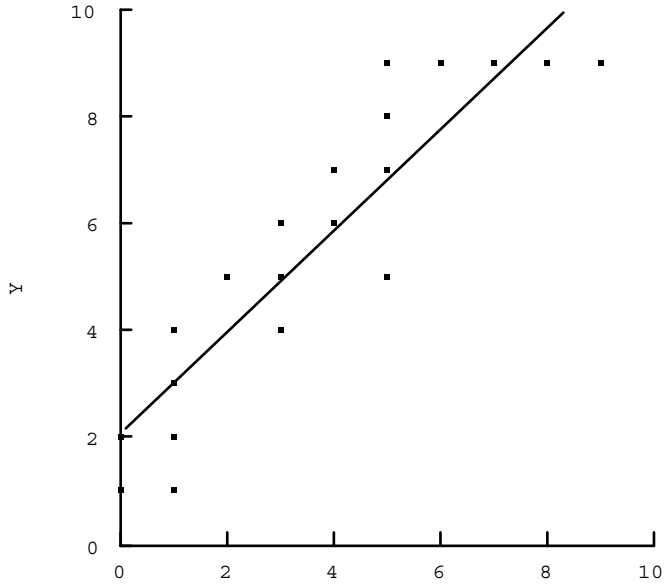
puntos en el plano (dado que cada caso puede expresarse como un par ordenado de puntuaciones). Si dos casos o más caen en el mismo punto, para no perder información, se puede sustituir éste por el número de casos que ocupan la misma posición.

Siempre que se analiza la relación entre dos variables cuantitativas un primer paso imprescindible consiste en trazar el diagrama de dispersión. Por eso antes de calcular el coeficiente de validez es preceptivo trazar el diagrama de dispersión y analizarlo Algunos efectos de la relación entre variables que no pueden ser descritos de modo claro por estadísticos son patentes si se mira al diagrama de dispersión correspondiente.

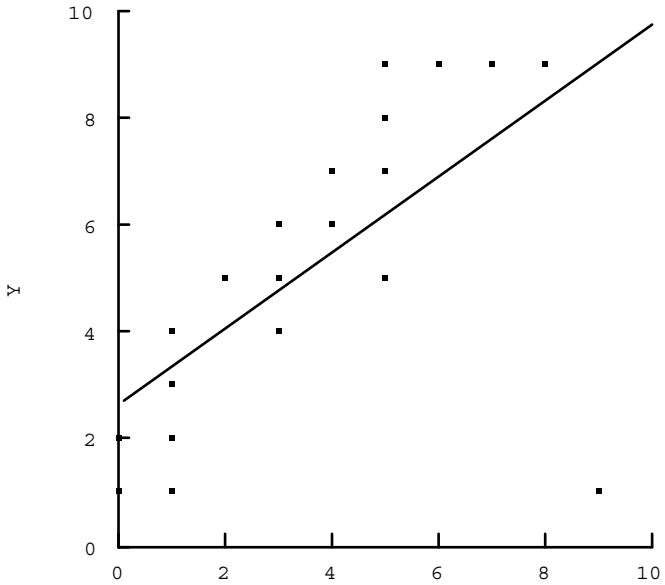
A continuación se presentan los diagramas de dispersión, con la recta de regresión correspondiente representada, para un conjunto de 25 casos cuyos datos han sido manipulados progresivamente para mostrar que relación hay entre diversos valores del coeficiente de validez y diferentes estados del diagrama de dispersión. Los siguientes gráficos ayudarán a entender visualmente la pregunta *¿Qué significa un coeficiente de validez de determinada cuantía?* ¿Cómo puede interpretarse?

No obstante debe tenerse en cuenta que, aunque un mismo diagrama de dispersión solo está asociado a un único valor del coeficiente de correlación de Pearson, sin embargo, un mismo coeficiente de correlación de Pearson puede deberse a innumerables situaciones posibles (incluso cualitativamente muy diferentes) del diagrama de dispersión. Por ello los gráficos siguientes ofrecen una imagen intuitiva pero no debe creerse que dado un valor del coeficiente de correlación de Pearson este es el único diagrama de dispersión posible. Los gráficos siguientes ilustran un ejemplo de una situación posible del diagrama de dispersión para un coeficiente de correlación determinado, aunque para ese mismo coeficiente de correlación los puntos podrían disponerse de otras muchas formas.

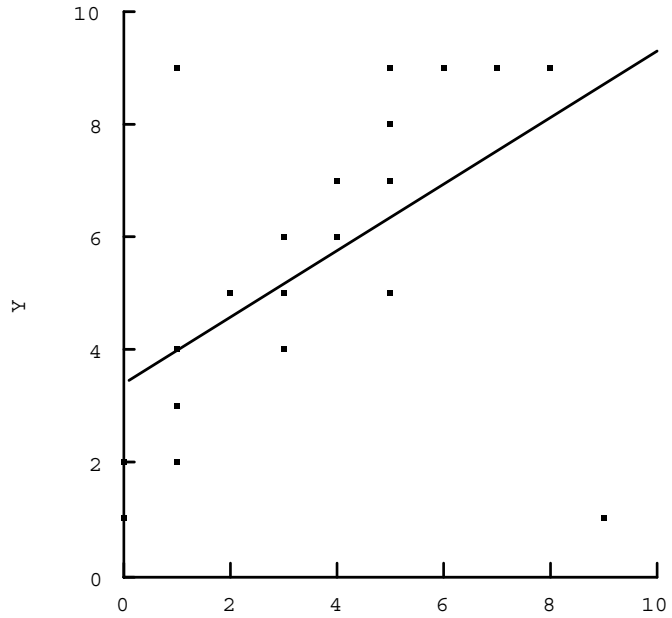
Ejemplo 1: r = 0,922.
(Coef. de Corr. de Pearson = 0'922).
Relación positiva fuerte. Inclinación positiva.
El ángulo de la recta es casi de 45º



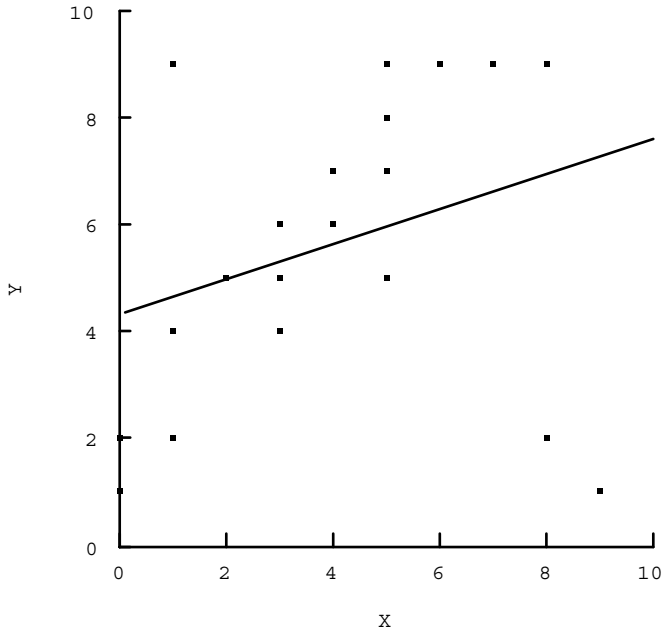
Ejemplo 2. r = 0,678.
Un punto se ha separado fuertemente de la recta.



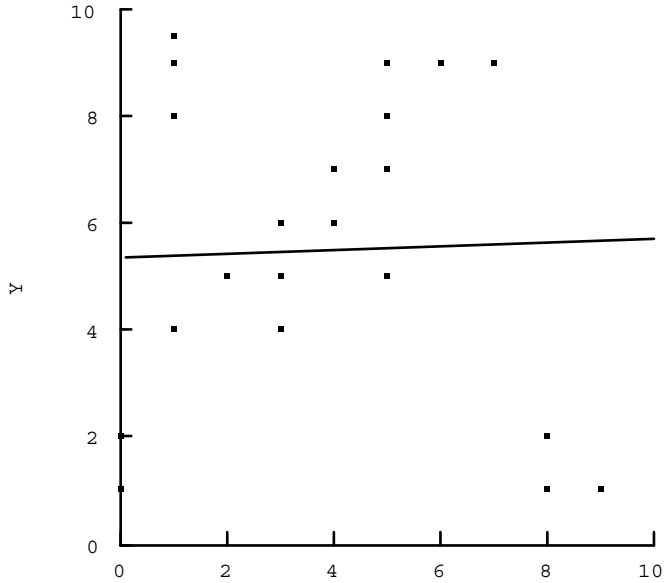
Ejemplo 3: r=0,571
Algunos puntos se han separado mucho de la recta.



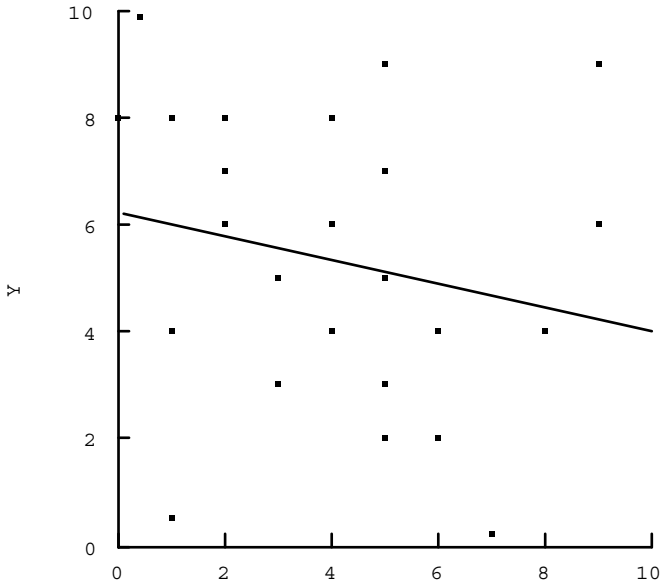
Ejemplo 4:
r=0'312
Muchos puntos lejos de la recta. Relación débil.



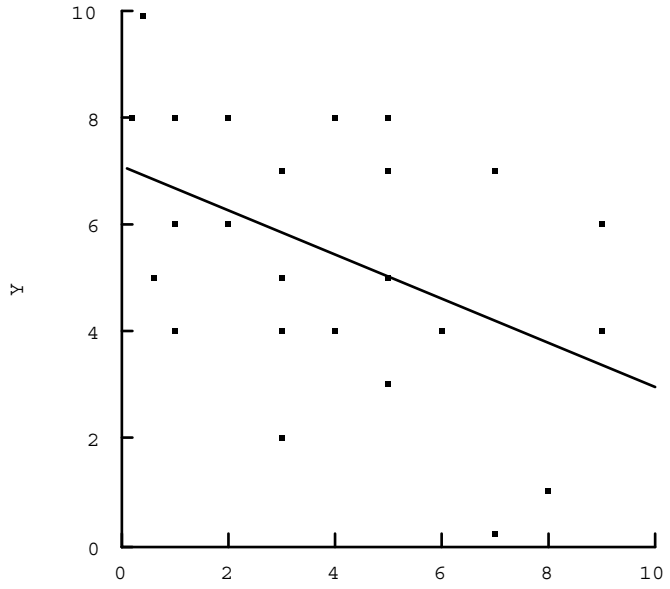
Ejemplo 5: r=0,032
Ausencia de relación. Linea casi horizontal.
Y no crece, ni decrece (prácticamente) al crecer X. No se pueden representar los datos adecuadamente con una recta. Es decir, la recta representa muy mal los puntos del diagrama de dispersión.



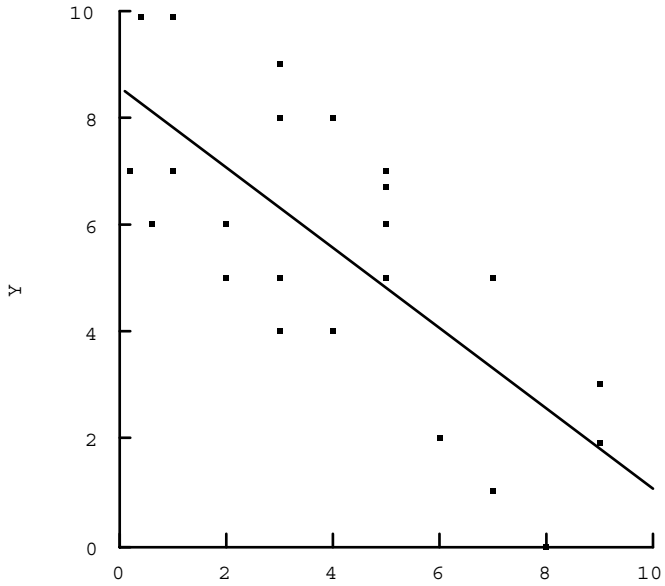
Ejemplo 6: r=-0,212
Comienza a definirse una relación negativa.



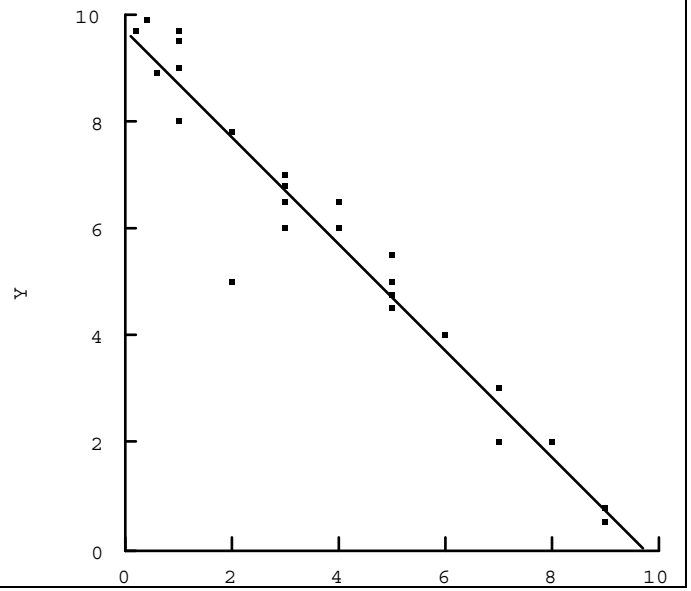
Ejemplo 7: r = -0.463.
Los datos se acercan un poco más a la recta.
Crece la inclinación negativa.



Ejemplo 8: r = -0,744
Claramente Y decrece al crecer X.



Ejemplo 9: r = -0,966
Y decrece al crecer X.
Todos los puntos casi sobre la recta.
La inclinación negativa es muy fuerte.



2. Modalidades del coeficiente de correlación de Pearson

No siempre test y criterio son dos variables cuantitativas adecuadas para utilizar alguna de las fórmulas generales del coeficiente de correlación de Pearson. En función de la naturaleza de las variables implicadas, el coeficiente de correlación de Pearson puede adoptar varias modalidades .

Cuando hablamos aquí de naturaleza de las variables nos referimos a si una o ambas son:

-*Cuantitativas*: Medidas de intervalo o de razón que se caracterizan porque hay una unidad de medida constante. A veces las llamamos también *métricas*.

-*Ordinales*: Los números representan una posición de orden, sin que exista una distancia constante entre números consecutivos (no hay unidad de medida).

-*Dicotómicas*: Una variable que se presenta en la realidad sólo bajo dos modalidades (como varón-mujer o vivo-muerto).

-*Dicotomizadas*: Una variable que o bien la medimos o bien la tratamos como si

tuviera solo dos categorías, pero que, en realidad es de naturaleza cuantitativa.

A continuación expondremos las principales modalidades del coeficiente de correlación de Pearson en función de la naturaleza de las variables implicadas:

Caso 1: Cuantitativa-cuantitativa: Expresiones generales de Pearson. Si ambas variables a correlacionar son de naturaleza cuantitativa, entonces corresponde aplicar el coeficiente de correlación de Pearson en alguna de sus fórmulas generales de las que ya nos hemos ocupado.

Caso 2. Dicotómica-cuantitativa. Si una variable es cuantitativa y la otra es dicotómica entonces puede utilizarse la fórmula denominada *coeficiente de correlación biserial puntual*, que también puede adoptar varios aspectos además del que se presenta aquí.

Este caso es frecuente al correlacionar un ítem con respuesta verdadera (dicotómico) con el total del test (cuantitativo), o con un criterio (cuantitativo). Una de sus fórmulas es:

$$r_{xy\ bis-pun} = \frac{\bar{Y}_1 - \bar{Y}}{s_y}\sqrt{\frac{p}{q}}$$

Se denomina *variable dicotómica* a aquella variable que se presenta en la realidad bajo solo dos modalidades. Para mayor sencillez supongamos que la variable dicotómica ha sido puntuada 1 ó 0, como suele hacerse.

En la fórmula anterior $\bar{Y}_1$ significa la media en la variable cuantitativa de aquellos que han obtenido un 1 en la dicotómica.

La media de las puntuaciones de la variable cuantitativa para toda la muestra se simboliza por $\bar{Y}$.

Los valores p y q tienen el significado usual en análisis de la dificultad: p es la proporción en la variable dicotómica de los que puntúan un 1 y q la proporción en la variable dicotómica de los que puntúan 0.

Esta fórmula da el mismo resultado que si aplicamos la fórmula general de Pearson sobre esos datos. Actualmente esta fórmula es prácticamente innecesaria, no obstante

prefiero mencionarla aquí para evitar confusión cuando el lector estudie otros manuales.

Caso 3. Dicotómica-dicotómica. Si una variable es dicotómica y la otra también es dicotómica puede utilizarse la fórmula denominada *coeficiente de correlación phi*, que puede adoptar varios aspectos además del que aquí se presenta:

$$r_{xy\ phi} = \frac{bc - ad}{\sqrt{(a+c)(b+d)(a+b)(c+d)}}$$

En la fórmula anterior a, b, c y d son las frecuencias en las celdillas de la tabla de contingencia 2 por 2 que representa la relación entre ambas variables dicotómicas:

Tabla de frecuencias conjuntas
a dos variables dicotómicas.

		Variable X	
		0	1
Variable Y	0	a	b
	1	c	d

¿Cómo se lee la tabla? Supongamos que X e Y son dos items que deseamos correlacionar. Por ejemplo, si 20 sujetos puntúan a la vez 0 en el ítem X y 0 en el ítem Y entonces a=20.

La fórmula anterior también es una adaptación de la fórmula de Pearson, es decir, da lo mismo aplicar la fórmula general de Pearson que Phi a dos variables dicotómicas.

La probabilidad de que el lector tenga que aplicar la fórmula de Phi en su vida prácticamente se reduce al caso de que un hipotético profesor más interesado en lo académico que en lo útil decidiera utilizarla en un ítem de examen. Otra

aplicación más útil aunque también poco frecuente es permitir obtener la correlación a partir de datos publicados en tablas de este estilo (bastante del gusto de los profesionales de encuestas). Por lo demás si se dispone de los datos completos, caso a caso, actualmente es más sencillo utilizar la fórmula general, aun con una calculadora. (De hecho, existe otra fórmula de Phi que es exactamente la fórmula general de Pearson que he propuesto pero escrita sustituyendo media por valor p, dado que, como hemos visto, la media de un ítem dicotómico valorado 1 ó 0 es p)

Caso 4. Ordinal-ordinal. Si los datos de ambas variables fueran considerados ordinales, es decir, expresando posiciones de orden, entonces podría aplicarse la fórmula del llamado *coeficiente de correlación de Spearman.*

Este caso no será muy frecuente porque usualmente se hace el supuesto de que cualquier punto en cualquier ítem es igual a cualquier otro punto en cualquier otro ítem, y, por tanto, se hace el supuesto de que tratamos con escalas de intervalo.

$$r_{xy\ Spe} = 1 - \frac{6\sum(X-Y)^2}{N(N^2-1)}$$

En la fórmula anterior N es el número de casos.

Esta fórmula también es una reexpresión del coeficiente de correlación de Pearson que habrá de dar igual resultado que si aplicamos la fórmula general de Pearson sobre esos datos, con la particularidad de que cuando haya "empates" (igual X e Y para un caso) no equivale a Pearson. Por supuesto la conclusión de nuevo es que lo mejor es aplicar la fórmula general de Pearson. Se menciona, como las anteriores, a efectos informativos más que prácticos.

Es decir, en general, no necesitamos el coeficiente de correlación biserial-puntual ni el coeficiente phi, ni el de Spearman porque donde éstos pudieran aplicarse puede aplicarse con idéntico resultado la fórmula general de Pearson de la que éstos otros coeficientes son sólo reexpresiones particulares.

Caso 5. Cuantitativa-dicotomizada. Si una variable es cuantitativa y la otra dicotomizada puede aplicarse el *coeficiente de correlación biserial*.

$$r_{bis} = \frac{\bar{Y}_1 - \bar{Y}}{s_y} \cdot \frac{p}{y}$$

En la fórmula anterior y es la ordenada, en una curva normal estandarizada, asociada al valor z' correspondiente a la proporción p. -En la columna E de la tabla B del manual de Amón (1979) puede encontrarse el valor de p/y conocido el valor p (es decir la proporción de unos en la variable dicotomizada).-.Los demás términos significan lo mismo que en la fórmula de la correlación biserial-puntual.

La fórmula es muy semejante a la del coeficiente de correlación biserial-puntual, aunque conceptualmente las dos fórmulas son muy diferentes. Para un mismo conjunto de datos la correlación biserial siempre es mayor, en valores absolutos, que la biserial-puntual.

$$r_{bis} = \frac{\sqrt{pq}}{y} \cdot r_{xy\ bis-pun}$$

(Puede verse Amón, 1979, pag. 302, ó Glass y Stanley. 1974, pag. 171, para los detalles de la deducción de fórmulas).

La correlación biserial no es la fórmula de Pearson, sin más, como la biserial-puntual. La correlación biserial es una estimación de que hubiera dado Pearson si la variable dicotomizada no hubiera sido dicotomizada, bajo el supuesto de que tenga una distribución normal. Al dicotomizar una variable se produce una fuerte restricción de su variabilidad que reduce su correlación con cualquier otra variable. La correlación biserial es una corrección de Pearson "al alza" que, por cierto, propuso el mismo Pearson. Si la variable que ha sido dicotomizada no presentaba realmente una distribución normal, entonces la corrección no será correcta. De hecho si la variable difería mucho de la normalidad podemos encontrar correlaciones biseriales cuyos límites no son -1 y +1 (pudiendo ser mayores o menores). Por ejemplo en ciertos casos, podemos encontrar coeficientes biseriales mayores que 1. Esto es una indicación de que la variable dicotomizada no puede suponerse razonablemente que distribuía normalmente.

¿Cuando utilizar correlación biserial en lugar de biserial-puntual? Cuando estamos correlacionando dos variables siendo una de ellas cuantitativa y la otra *dicotomizada* (y *no* dicotómica). La cuestión se traduce pues en esta otra ¿Cuando considerar una variable dicotomizada y no dicotómica? Hay casos en que está claro que hablamos de una variable dicotómica (por ejemplo, vivo o muerto) y casos en que no está claro si una variable es dicotomizada.

Una variable dicotomizada es aquella que por su naturaleza la podríamos considerar cuantitativa pero que, o bien (caso a) la medimos con un método que solo permite establecer

una dicotomía, o bien (caso b) la medimos reflejando su carácter cuantitativo pero después la dicotomizamos antes de operar con ella.

En general el caso b no debe producirse. Es absurdo, a efectos estadísticos y psicométricos, disponer de la información cuantitativa real de una variable, que refleja como es realmente la variable, y distorsionar esta información rompiéndola en sólo dos valores para relacionarla con otra. Dicotomizar casi siempre es una práctica desaconsejable. Primero se fuerzan las cosas distorsionando gravemente la información hasta hacerla aparecer como si solo tuviera dos casos, después se intenta estimar mediante el coeficiente biserial que hubiera pasado si no se hubiera cometido este atropello a la capacidad de la medición para representar (lo mejor que se pueda si es que se puede) el mundo real. En el mejor de los casos la corrección se aproximará a la verdadera correlación que se hubiera obtenido con la variable antes de dicotomizar; pero esto solo es así si en la medida en que esa variable presente una distribución próxima a la de una curva normal.

¿Es cuantitativa la variable que subyace a acertar o no acertar items? Como no hay contraste, en muchos casos se suele suponer razonablemente sin riesgo de equivocarse que la variable subyacente es cuantitativa. Por ejemplo, suele suponerse que la variable subyacente a un ítem de matemáticas, denominada “capacidad matemática”, es cuantitativa. En mi opinión cuantas menos suposiciones (especialmente si son no contrastables) mejor. O, dicho de otro modo, que de lo único que estamos seguros en estos casos es de que la medición concreta sólo arroja dos valores, es decir, *de hecho,* disponemos de una medida dicotómica. Pero he de admitir que puede argumentarse razonablemente que se trata de una variable dicotomizada.

En cualquier caso , si la variable subyacente es cuantitativa pero presenta una distribución distinta de la normal, aun en el hipotético caso de que estemos seguros de que es dicotomizada, la correlación biserial no producirá una corrección adecuada de Pearson para estimar que hubiera dado la correlación entre la variable sin dicotomizar y la otra también cuantitativa.

En realidad no hace falta que la distribución subyacente a la dicotomizada sea una distribución extraña para que el coeficiente biserial no estime bien Pearson, basta que sea pronunciadamente platicúrtica (aplastada) en cuyo caso los límites pasan de -1 y +1, o pronunciadamente leptocúrtica (afilada), en cuyo caso los límites no llegan a -1 y +1. ¿Qué hubiera dado exactamente Pearson en estos casos? Si no disponemos de una medición de la variable sin dicotomizar no es posible saberlo. Y, si disponemos desde el principio de ella ¿para qué dicotomizar?.

Caso 6. Dicotomizada-Dicotomizada. Si tenemos dos variables dicotomizadas, es decir, que provienen de variables cuantitativas, y si además sus distribuciones originales son normales, entonces puede utilizarse el *coeficiente de correlación tetracórico* para estimar que hubiera pasado de haber correlacionado las variables sin dicotomizarlas. El coeficiente de correlación tetracórico es una corrección al alza de phi (es decir, del coeficiente de correlación de Pearson) que, por cierto, también fue deducido por el mismo Pearson. Como es de formulación compleja el modo cabal de calcularlo es mediante un programa o un paquete estadístico. Si no hay más remedio que utilizarlo a mano la siguiente es una fórmula de aproximación a la correlación tetracórica:

$$r_{tetrac.} = \cos\left(\frac{180^{o}}{1+\sqrt{\frac{bc}{ad}}}\right)$$

En la fórmula "cos" significa calcular el coseno del paréntesis (en grados). Los demás términos tienen el mismo significado que en phi. La aproximación tiene el defecto adicional de que en la medida en que las proporciones se alejan de 0,5 la fórmula sobrestima. (Una sobreestimación para corregir la infraestimación que produce Pearson no suena a buena solución). Por tanto, en general, el modo razonable de utilizar la correlación tetracórica, si es que hay que utilizarla alguna vez, es mediante ordenador.

Espero que de la discusión anterior el lector haya sacado la conclusión de que el coeficiente de correlación de Pearson es el coeficiente esencial.

3. Predicción del criterio mediante regresión lineal simple

Una ecuación de regresión permite estimar un criterio Y a partir de uno o más tests X. Se dice entonces que tenemos una ecuación de regresión de Y sobre la o las variables X. Cuando existe una sola variable independiente o predictor X la regresión se denomina simple. Si hay varios predictores X se habla de represión múltiple.

Dados un test X y un criterio Y, ambos variables cuantitativas, si la variable dependiente criterio Y es una función lineal (al menos aproximadamente) de la variable independiente test X entonces podemos utilizar una ecuación de regresión lineal simple para pronosticar Y a

partir de X. En general, si X es un predictor lineal de Y, podemos escribir:

$$Y'_i = a + bX_i$$

donde Y'_i refleja la *puntuación estimada o pronosticada* en la variable Y para el caso i a partir de su puntuación X_i

Para unos datos determinados, a es una constante en la ecuación que indica el valor de Y'_i cuando X_i vale 0. Gráficamente el parámetro a viene expresado por el punto en que la recta cruza el eje de ordenadas Y. Por eso el parámetro a se denomina "punto de corte del eje Y", *"punto de intercepción"* o "intercepción".

Para unos datos determinados, b es una constante en la ecuación que indica cuanto crece Y'_i cuando X_i crece una unidad. Gráficamente viene expresado por la inclinación de la recta. Cuanto mayor es esta constante mayor es la inclinación de la recta. Por eso el parámetro b se denomina "inclinación" o *"pendiente" de la recta*. El signo de esta constante indica si la relación es positiva (Y crece cuando el test X crece) o negativa (Y decrece cuando el test X decrece).

3.1. ¿Cómo pronosticar una variable a partir de un test?

Para obtener los valores de los parámetros de la ecuación de regresión es imprescindible disponer de un conjunto de casos en los que se conozca tanto el test X como el criterio Y. Una vez establecida la ecuación de regresión (es decir, una vez que se ha averiguado el valor de los parámetros de la ecuación para esos datos) ésta podrá utilizarse para estimar Y en los casos conocidos y, en consecuencia, estimar el error en los pronósticos. Y también, una vez conocida la ecuación de regresión ésta puede utilizarse para pronosticar el valor en Y de otros casos de la misma población de los que solo conozcamos su puntuación en el test X. Para estos últimos casos, a no ser que después podamos realmente medirlos en Y, no podremos calcular los errores de estimación.

Los pasos o fases de trabajo en regresión son los siguientes:

Paso 1) Obtener los *datos* del test X y del criterio Y.

Obtener una muestra adecuada de la población a estudiar. En cada caso de esa muestra es necesario medir tanto el test X, que actúa como variable independiente o predictora

en la ecuación, como el criterio Y, que actúa como variable dependiente o pronosticada en la ecuación. De cada sujeto hacen falta ambas puntuaciones.

Paso 2) *Estimar la relación* en esos datos mediante gráficos y estadísticos.

Paso 2.1) Trazar el *diagrama de dispersión* y evaluar si la relación entre X e Y puede describirse como lineal.

Para esclarecer que tipo de relación hay entre el test X y el criterio Y puede resultar imprescindible disponer los datos en una *tabla de correlación* y estudiar el comportamiento de las *distribuciones condicionadas*, su dispersión y su media. Para que la relación puede calificarse de lineal las medias condicionadas han de poder representarse mediante una recta.

Paso 2.2) Si la relación puede describirse aceptablemente como lineal *calcular el coeficiente de correlación* de Pearson y su cuadrado el coeficiente de determinación. El coeficiente de correlación entre test y criterio es el coeficiente de validez del test para ese criterio.

Paso 2.3.) Calcular la *ecuación de regresión* sobre esos datos. Es decir, calcular el valor de los parámetros a y b de la ecuación de regresión lineal.

Paso 2.4.) Utilizando la ecuación de regresión calcular para cada caso el valor de la Y' *pronosticada* a partir de su valor en X. A continuación calcular para cada caso el *residual* $E = Y - Y'$. Estudiar el comportamiento de los residuales y calcular el error típico de estimación para contrastar la adecuación del modelo a los datos.

Paso 3) Utilizar la ecuación de regresión para pronosticar en nuevos casos de esa población.

Para ello se obtienen las puntuaciones de los sujetos en el test X y se obtienen valores en Y' aplicando la ecuación. La ecuación con los parámetros calculados actúa como una "máquina de pronosticar" Y a partir de X.

En general, para la que la ecuación pueda aplicarse a un nuevo caso en el que solo conocemos la puntuación en el test X, este caso ha de pertenecer a la misma población en la que se obtuvo la ecuación. En caso contrario se podría estar atribuyendo al caso un comportamiento que no le corresponde.

El grado en que el caso es asimilable al grupo con que se calculó la ecuación depende de las características del caso y de las características del grupo, y está relacionado con la *validez externa de población* del modelo calculado, es decir,

con el grado en que éste puede generalizarse razonablemente a través de casos.

También la puntuación en el test X que presenta el nuevo caso está relacionada con la aplicabilidad al mismo del modelo de regresión. En efecto, cuando se calculó el valor de los parámetros de la ecuación de regresión, se hizo para determinado rango de valores del test X, aquellos que estaban disponibles en la muestra original.

Si el nuevo caso presenta un valor de X dentro del rango de X en la muestra original, y para el que había casos en aquella muestra, el uso de la ecuación para pronóstico o estimación está claramente justificado, siempre, por supuesto, que el nuevo caso pueda decirse que pertenece a esa misma población.

Si el nuevo caso presenta un valor del test X dentro del rango de X de la muestra original pero para el que no había casos disponibles en aquella muestra, la estimación o pronóstico de su valor en Y a partir de la ecuación es una *interpolación*.

La interpolación está tanto más justificada cuanto más cercanos estén al nuevo valor X los valores del test X para los que se conocía empíricamente su relación con Y en la muestra original y cuanto más ajusten los datos al modelo de regresión en su conjunto, siempre suponiendo que, adicionalmente, el nuevo caso pertenece a la misma población de la muestra original.

Por el contrario si el nuevo valor en X se encuentra en una zona del rango de puntuaciones del test X dentro del rango conocido originalmente, pero en un "agujero" sin información por ausencia de datos en la muestra original en esa zona de X, la interpolación, aunque desde un punto de vista formal no es cuestionable si el modelo ajusta bien, desde un punto de vista práctico debe ser hecha con cierta reserva y tratar de confirmar el comportamiento en Y tan pronto sea posible. Estas apreciaciones generales se hacen pensando en numerosas variables psicológicas sujetas a investigación. Pero es obvio que también pueden encontrarse muchas variables en estos campos cuya relación bien establecida permita interpolar con confianza.

Si el nuevo caso presenta un valor de X fuera del rango de X en la muestra original, el pronóstico de Y utilizando la ecuación de regresión es una *extrapolación*. Las extrapolaciones pretenden pronosticar en zonas de puntuaciones del test X donde no se conoce exactamente que sucede, y por tanto, al carecer de datos empíricos en esa zona sobre X, sobre Y y sobre la relación de X e Y, están expuestas a serios errores debidos a una posible inadecuación del modelo y/o de sus parámetros, más allá del residual o error de pronóstico habitual que es esperable en cualquier estimación de Y.

Incluso si el modelo es cierto y funciona muy bien en la zona conocida de X esto no significa que necesariamente la extrapolación esté justificada en zonas adyacentes del rango de puntuaciones de X. Pueden ponerse muchos ejemplos, algunos divertidos, de error de extrapolación. Por ejemplo, un bebe mide al nacer 60 cm, y crece al ritmo de 1'5 cm por semana durante las primeras tres semanas. Por prudencia, los padres obtienen datos de las semanas 4 y 5, y la relación se confirma: crece a 1'5 por semana. Los padres quieren pronosticar si podrá jugar al baloncesto, así que estiman el modelo siguiente: 1'5 cm./semana a 52 semanas por año, equivale a 78 cm./año. El modelo es pues Y'=60+78X, siendo X el número de cumpleaños celebrados e Y' la estatura pronosticada. Según el modelo, está claro, a los 10 años el niño medirá 8 m. y 40 cm. ¡Algo más de lo necesario para la selección de baloncesto! ¿Qué sucede? Hay dos tipos de problemas básicos con esta extrapolación. Primero, el modelo no se sostiene más allá de las primeras semanas de vida y es progresivamente peor a medida que el valor de X es mayor. Segundo, se conoce un rango de X de 0 a 5 (semanas) y se está pronosticando para X=520 (10 años por 52 semanas), muy lejos de la zona conocida de X e Y. Una dificultad que se aprecia en este ejemplo respecto a la inadecuación del modelo fuera del rango conocido de X, consiste en que el modelo es razonablemente lineal en la zona de X conocida, pero sería curvilíneo si dispusiéramos de más información en el rango de X. Esta dificultad tiene interés general pues es frecuente que fragmentos de curva puedan representarse muy bien como rectas, de modo que la interpolación en el trozo conocido puede ser razonable. Pero, dado que más allá de lo conocido la relación no describe una recta sino una curva, el modelo es totalmente inaceptable para extrapolación, produciendo estimaciones aberrantes como la del ejemplo.

La calidad de una extrapolación depende del valor en el test X que adopte el nuevo caso, de su posición respecto al rango conocido de X, y depende del grado en que el modelo sea generalizable a otros valores de X, lo que constituye una dimensión de generalizabilidad del modelo dentro de su *validez externa de situación*. La validez externa de situación del modelo se refiere al grado en que el modelo es generalizable a otras situaciones distintas de aquella en que fue calculado. Si se puede generalizar el modelo a otros valores del rango de puntuaciones del test X fuera del rango original, esto es una dimensión de la validez externa del modelo. Con la extrapolación toda prudencia es poca y en muchas variables psicológicas y sociales deja de ser confiable unas pocas unidades de X más allá del rango conocido.

Los modelos de regresión en validez sirven para dos *propósitos* básicamente. Primero, para **contrastar teorías** que afirman como se relacionan entre sí variables

cuantitativas. La relación de un test con un conjunto de criterios forma parte de lo que se conoce como red nomológica y la regresión es una de las técnicas estadísticas que puede utilizarse para su estudio, para establecer cuales son las relaciones entre las variables de un modo preciso. Y, en segundo lugar, para **pronosticar** el comportamiento en el criterio de nuevos casos de la misma población de los que sólo se conoce o sólo se puede conocer su valor en el test.

La confianza en el modelo depende de la *calidad de la muestra* en que se estimó (su tamaño y representatividad), del *grado de ajuste intrínseco* que el modelo muestra entre las variables, y del número de *replicaciones* (repetición de la investigación para comprobar si los resultados son ciertos en una nueva muestra) que lo avalan.

La *replicación* es un concepto muy importante en ciencia. Permite, si es una *replicación directa*, confirmar en una nueva muestra de las mismas características y bajo las mismas circunstancias, si el modelo calculado anteriormente se sostiene. Y si es una *replicación sistemática*, permite averiguar si el modelo puede extenderse a otras situaciones, entre ellas a otros valores del rango posible de X. Por tanto, estos aspectos han de ser considerados en cada caso concreto para plantearse el uso de una ecuación para pronosticar, interpolar o extrapolar información en un criterio.

3.2. ¿Cómo se calcula el valor de los parámetros de la ecuación para unos datos?

A partir de la ecuación de regresión, bajo el principio minimo-cuadrático a que ha de responder la estimación de los parámetros a y b, se deducen las fórmulas de cálculo de los mismos.

Expresada la ecuación de regresión lineal simple

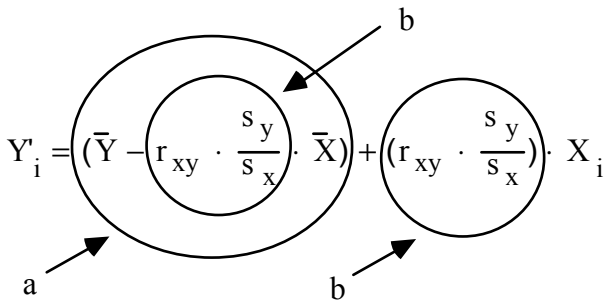
$$Y' = a + bX$$

en puntuaciones directas, los parámetros valen:

$$b = r_{xy} \cdot \frac{s_y}{s_x}$$

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

De este modo la ecuación de regresión lineal minimocuadrática Y'=a+bX puede escribirse, sustituyendo a y b por sus valores:



Si la ecuación de regresión se expresa en puntuaciones diferenciales o en típicas los valores de a y b varían consecuentemente. En el cuadro siguiente se resume el *valor de los parámetros* a y b en la ecuación de regresión según trabajemos con puntuaciones directas, diferenciales o típicas.

$$Y' = a + bX$$

	"a" en la ecuación es igual a:	"b" en la ecuación es igual a:
En puntuaciones Directas:	$\bar{Y} - b\bar{X}$	$r_{xy} \cdot \frac{s_y}{s_x}$
En Puntuaciones Diferenciales:	0	$r_{xy} \cdot \frac{s_y}{s_x}$
En puntuaciones Típicas:	0	r_{xy}

3.3. ¿Cómo obtener los residuales?

La diferencia entre el valor real de Y para el caso i, representado por Y_i, y el valor de Y pronosticado por la ecuación para ese mismo caso, representado por Y'_i, es un *error de estimación* que representaremos por E .

$$E_i = Y_i - Y'_i$$

El **error de estimación** E también se denomina error de predicción, error de pronostico o residual.

Los pronósticos Y' pueden (y así ocurre normalmente) no ser exactamente iguales a las puntuaciones reales en Y. Denominamos *error de predicción* E a la diferencia entre el valor que realmente adopta la variable Y y el valor Y' que pronostica la ecuación para ese caso. Para cada caso i existe una diferencia E que expresa el error de pronóstico. Pueden darse tres situaciones:

a) $Y' = Y$ de donde se desprende que $E = 0$. La predicción para ese caso es perfecta.

b) $Y' > Y$ de donde $E < 0$. El pronóstico Y' sobrestima la puntuación real Y.

c) $Y' < Y$ de donde $E > 0$. El pronóstico Y' infraestima la puntuación real Y.

Para obtener el error E es necesario (1) disponer de la puntuación del sujeto en el test X, (2) disponer del valor estimado de los parámetros a y b de la ecuación de regresión, (3) aplicar la ecuación de regresión sobre X para estimar el criterio Y' , y, (4) además, poder medir al sujeto en el criterio Y para (5) poder averiguar la diferencia entre el criterio pronostica Y' y el valor real en el criterio Y. Todo esto puede hacerse con todos los casos de la muestra original en que se estima la ecuación de regresión y para la que es necesario disponer tanto de sus puntuaciones en el test X como de sus puntuaciones en el criterio Y. Muchos paquetes y programas estadísticos pueden efectuar todo este proceso automáticamente.

3.4 ¿Cómo obtener el error típico de estimación?

Cuanto más error de estimación E cometa una ecuación tanto peor es capaz de predecir el criterio Y a partir del test predictor X. Como el objetivo es reducir ese error para el conjunto total de los casos, necesitamos un indicador de cuanto error de estimación comete la ecuación en el conjunto.

Dado que algunos errores son de signo positivo (las infraestimaciones) y otros de signo negativo (las sobrestimaciones), no podemos simplemente sumar E.

Para evitar esta dificultad se opta por considerar los errores elevados al cuadrado.

La suma:

$$\sum E^2$$

es un indicador adecuado de cuanto error (al cuadrado) se ha cometido en conjunto.

Para que este valor no dependa del número N de casos considerados en el análisis se procede a calcular una media de error cuadrático:

$$\frac{\sum E^2}{N}$$

Este promedio de error cuadrático tiene la estructura de una varianza:

$$\frac{\sum (Y - Y')^2}{N}$$

y, de hecho, expresa la distancia cuadrática media entre los valores del criterio Y y sus pronósticos Y'.

Evidentemente esa distancia cuadrática media está expresada en las unidades de Y elevadas al cuadrado.

Para devolver la expresión a las unidades de Y se procede a calcular la raíz cuadrada de esa expresión, al resultado se le denomina *error típico de estimación*.

$$s_{y.x} = \sqrt{\frac{\sum (Y - Y')^2}{N}}$$

El **error típico de estimación** es un estadístico que indica cuanto error de estimación comete globalmente la ecuación de regresión.

Obsérvese la similitud con la fórmula de la desviación típica. Si en la fórmula anterior sustituyéramos Y' por $\bar{Y}$ la fórmula se convertiría en la desviación típica de Y.

De esta analogía puede deducirse que del mismo modo que la desviación típica indica el error de estimación que cometemos al pronosticar la media $\bar{Y}$ en lugar de cada valor real de Y, el error típico de estimación indica el error de estimación que cometemos al pronosticar Y' en lugar de cada valor real de Y,.

Si se calculan los parámetros a y b de la ecuación de regresión lineal de tal modo que la suma de los errores de estimación al cuadrado sea lo menor posible se dice que efectuamos regresión lineal *minimocuadrática.* El propósito de un análisis de regresión es precisamente estimar estos dos parámetros mediante sus fórmulas, obtenidas de modo que garantizan que estas constantes hacen mínima la suma de los errores cuadráticos respecto a los valores estimados de Y. Calcular los parámetros a y b de la ecuación de regresión lineal de tal modo que la suma de los errores de estimación al cuadrado sea lo menor posible equivale a calcular los parámetros a y b de la ecuación de regresión lineal de tal modo que el error típico de estimación sea lo menor posible.

Error típico de estimación insesgado

Al igual que sucede con la varianza, el error típico de estimación, tal como lo hemos definido, es un estimador sesgado del error típico de estimación de la población. En la varianza, para corregir su sesgo, se modifica la fórmula de la misma sustituyendo el valor N del denominador por el valor N-1. Paralelamente, en el error típico de estimación de la regresión lineal simple, para corregir su sesgo, hay que sustituir el N del denominador por N-2. Algunos textos prefieren definir directamente al error típico de estimación con un denominador de N-2, que es el correcto cuando se pretende inferir el error típico de estimación de la población a partir de una muestra.

Una fórmula operativa para el error típico de estimación

La fórmula anterior del error típico de estimación es una fórmula conceptual. Puede usarse para calcular el valor del error típico de estimación, pero, sobre todo, expresa su significado.

Sustituyendo en la fórmula del error típico de estimación Y' por su valor en la ecuación, una vez deducidos los valores de los parámetros a y b, y posteriormente simplificando, se alcanza una fórmula operativa del error típico de estimación

$$s_{y.x} = s_y \sqrt{1 - r^2_{xy}}$$

En esta fórmula podemos obtener el valor de $s_{y.x}$ conociendo simplemente dos estadísticos: s_y la desviación típica de Y, y el coeficiente de determinación r^2_{xy} (es decir, el coeficiente de validez del test X con el criterio Y elevado al cuadrado).

3.5. Un ejemplo numérico de regresión lineal simple

Supongamos que el criterio Y es el rendimiento en una tarea en una escala de 0 a 10 puntos, donde 10 significa el máximo rendimiento. Supongamos que X es un test que estamos interesados en poder utilizar para pronosticar ese rendimiento. X mide cierta capacidad en una escala de 0 a 20.

Nuestro propósito es poder pronosticar Y a partir de X, de tal modo que sólo conociendo el resultado de una persona en X podamos anticipar cual sería su rendimiento en la tarea que mide Y. Sin embargo, para ello es necesario establecer cual es la relación que hay entre X e Y, y para establecer esa relación necesitamos conocer las puntuaciones en el test X y en el criterio Y de una muestra de sujetos representativa de aquellos para los que después queremos utilizar la ecuación con propósitos de predicción.

Supongamos que tenemos una muestra representativa y suficiente en la que hemos podido medir X y también medir Y. Por simplicidad supongamos que esa muestra está formada sólo por los 10 casos siguientes (Un tamaño de muestra de cómo este sólo tiene justificación por razones didácticas, para poder seguir adecuadamente la exposición).

Caso:	X	Y
1	5	13
2	2	5
3	4	9
4	9	19
5	5	9
6	7	15
7	1	3
8	6	13
9	3	7
10	8	17

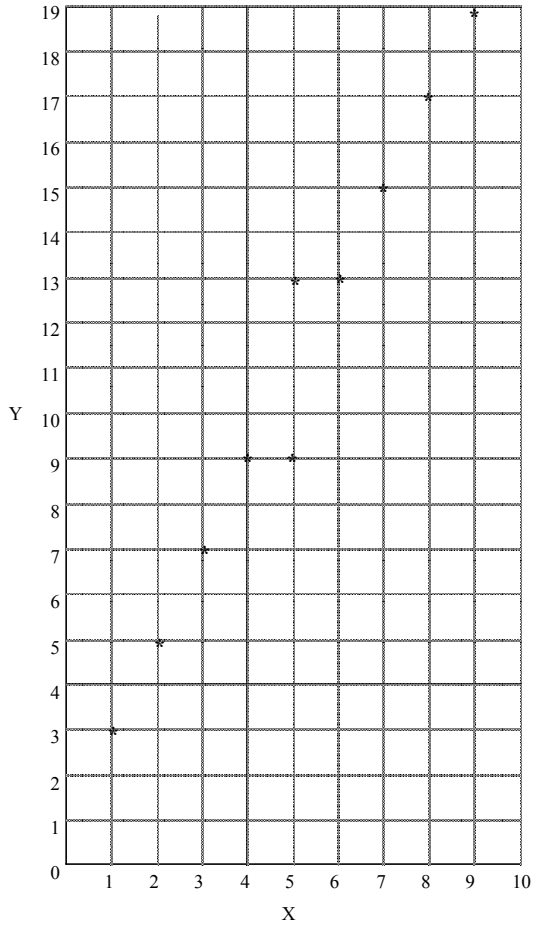
Para estudiar la relación entre dos variables el primer paso es trazar el diagrama de dispersión. En este diagrama se señala el par de puntuaciones en X e Y de cada caso como un punto. La variable independiente test X se sitúa en abscisas. La variable dependiente criterio Y se sitúa en ordenadas.

El propósito principal de trazar el *diagrama de dispersión* inicialmente es *observar visualmente el tipo de relación* que describen los datos y *detectar anomalías* (por ejemplo valores extremos, zonas sin observaciones, distribuciones marcadamente asimétricas...) en los mismos. Si por ejemplo los datos describen claramente una curva no estaría indicado utilizar las herramientas de regresión lineal ni el coeficiente de correlación de Pearson para cuantificar esa relación.

En el diagrama de dispersión de los datos anteriores para las variables X e Y podemos observar una relación lineal patente, dado que los puntos parece que pueden describirse adecuadamente mediante una linea recta, por lo que está indicado aplicar el coeficiente de correlación de Pearson (coeficiente de validez del test X para el criterio Y) y el análisis de regresión lineal para el pronostico del criterio Y a partir del test X.

Puede apreciarse en el diagrama de dispersión que esa relación es, además, de claro signo positivo pues, en general, a valores mayores de X corresponden valores mayores de Y y a valores menores de X corresponden valores menores de Y. Cuanto mayor es X mayor parece que tiende a ser Y.

Diagrama de dispersión



En el diagrama de dispersión y en el análisis de las distribuciones de Y condicionadas a los valores de X buscamos verificar que se cumplan las condiciones para poder aplicar regresión lineal simple mínimo-cuadrática. En este caso, a pesar de la escasez de puntos, esas condiciones parecen cumplirse suficientemente.

Dado el diagrama de dispersión, procedemos a calcular el coeficiente de correlación de Pearson y los parámetros de la ecuación de regresión lineal simple.

Para ello podemos organizar la información mediante una tabla de directas, diferenciales y típicas como la siguiente.

	Directas			Diferenciales			Típicas		
Caso:	X	Y	X*Y	x	y	x*y	Zx	Zy	Zx*Zy
1	5	13	65	0	2	0	0	0,402	0
2	2	5	10	-3	-6	18	-1,22	-1,2	1,476
3	4	9	36	-1	-2	2	-0,41	-0,4	0,164
4	9	19	171	4	8	32	1,633	1,606	2,623
5	5	9	45	0	-2	0	0	-0,4	0
6	7	15	105	2	4	8	0,816	0,803	0,656
7	1	3	3	-4	-8	32	-1,63	-1,61	2,623
8	6	13	78	1	2	2	0,408	0,402	0,164
9	3	7	21	-2	-4	8	-0,82	-0,8	0,656
10	8	17	136	3	6	18	1,225	1,205	1,476
Suma	50	110	670	0	0	120	0	0	9,837
Media	5	11	67	0	0	12	0	0	0,984
d.t.	2,45	4,98							

En la tabla puede verse que la media de X es 5, la de Y es 11. La desviación típica de X es 2'45 y la de Y es 4'98 (el proceso de cálculo de las deviaciones típicas paso a paso no se ha incluido en la tabla).

La covarianza (que es el promedio de los productos de las puntuaciones diferenciales) es igual a 12, indicando que existe una asociación lineal positiva.

El coeficiente de correlación de Pearson (que es el promedio de los productos de las puntuaciones típicas) es 0'98. Este valor es de signo positivo y de magnitud muy próxima a 1, lo que indica que existe una fuerte asociación lineal entre ambas variables. Como se trata de la correlación entre el test X y el criterio Y esa correlación es el *coeficiente de validez* del test X para con le criterio Y.

Si elevamos el coeficiente de correlación al cuadrado obtenemos el coeficiente de determinación, que en este caso es igual a 0'96. El test X pronostica un 96% de la varianza del criterio, lo que resulta una capacidad de pronostico muy alta.

Este resultado se interpreta diciendo que el 96% de la varianza del rendimiento Y es explicada o se debe a la aptitud medida por el test X.

En la tabla se ha ilustrado la obtención del coeficiente de correlación de Pearson por medio del producto de las puntuaciones típicas a efectos didácticos. Pero este método es muy costoso en tiempo y muy expuesto a errores debido a los cálculos con decimales y signos.

En realidad para obtener el coeficiente de correlación de Pearson hubiera bastado con las tres primeras columnas, referidas a puntuaciones directas, además de las desviaciones típicas. En términos prácticos calcularíamos el coeficiente de validez del siguiente modo:

$$r_{xy} = \frac{\overline{XY} - \bar{X}\bar{Y}}{s_x \, s_y} = \frac{67 - 5 \cdot 11}{2'45 \cdot 4'98} = 0'98$$

Establecido que existe una fuerte asociación lineal entre ambas variables vamos calcular una ecuación de regresión lineal simple tal que en el futuro podamos pronosticar los rendimientos Y conociendo sólo los resultados en el test X.

Podemos trabajar con las puntuaciones de las variables X e Y en forma de puntuaciones directas (tal como vienen de la medición) o en forma de puntuaciones diferenciales (cada puntuación menos la media) o en forma de típicas (cada diferencial dividida por su desviación típica). Lo más frecuente es que nos interesa la predicción en directas, pero, por razones didácticas, vamos a exponer el cálculo en las tres condiciones.

La ecuación de regresión en puntuaciones típicas es:

$$Z'_y = r_{xy} \cdot Z_x$$

Es decir, en típicas el parámetro a siempre vale 0 y el parámetro b estandarizado es el coeficiente de correlación de Pearson. Por tanto b=0'98 y la ecuación queda:

$$Z'_y = 0'98 \cdot Z_x$$

Esto significa que si convertimos en puntuación típica el resultado de un sujeto en el test X, basta multiplicar ese valor por 0'98 para obtener el rendimiento Y que esperamos expresado también en típicas.

La ecuación de regresión en puntuaciones diferenciales es:

$$y' = (\frac{s_y}{s_x} r_{xy}) \cdot x$$

Es decir, el parámetro a también vale 0 y el parámetro b vale el coeficiente de correlación multiplicado por el cociente entre las desviaciones típicas.

En este caso:

$$b = \frac{s_y}{s_x} r_{xy} = \frac{4'98}{2'45} \cdot 0'984 = 2$$

Por tanto, en puntuaciones diferenciales, bastará multiplicar la x diferencial por 2 para obtener la puntuación y' diferencial pronosticada:

$$y' = 2 \cdot x$$

Por último, en puntuaciones directas la ecuación

$$Y' = a + bX$$

necesita obtener el valor de b y el de a.

El de b ya lo tenemos, porque es el mismo que en diferenciales.

El de a es igual a la media de Y menos b por la media de X:

$$a = \bar{Y} - b\bar{X} = 11 - 2 \cdot 5 = 1$$

Es decir, b=2 y a=1, por tanto la ecuación de regresión lineal simple en puntuaciones directas es:

$$Y' = 1 + 2X$$

El valor de a=1 significa que la recta cruza el eje de ordenadas en el valor 1.

El valor de b=2 significa que la recta crece 2 unidades en Y cada vez que crece una unidad en X.

Conviene representar la recta de regresión sobre el diagrama de dispersión, como se hace en el gráfico siguiente.

Para trazar la recta basta con dos puntos. Se puede dar 2 valores cualquiera a X y trazar la recta por los puntos resultantes. Por ejemplo:

Para X=0 entonces Y=1

$$Y' = 1 + 2(0) = 1$$

Para X=5 entonces Y=11

$$Y' = 1 + 2(5) = 11$$

Por tanto trazamos una recta que pasa por los puntos (0;1) y (5;11).

Al trazar la recta de regresión sobre el diagrama de dispersión podemos observar gráficamente el significado de los parámetros a=1 y b=2.

El parámetro b es la tangente al ángulo α que forma la recta de regresión y el eje de abscisas: tg(α)=b.

Por tanto, la arcotangente de b es el ángulo α : Atan(b)= α.
.

En los datos del ejemplo b=2, por tanto Atan(2)=63'43º, por tanto α=63'43º. Recíprocamente tg(63'43)=2.

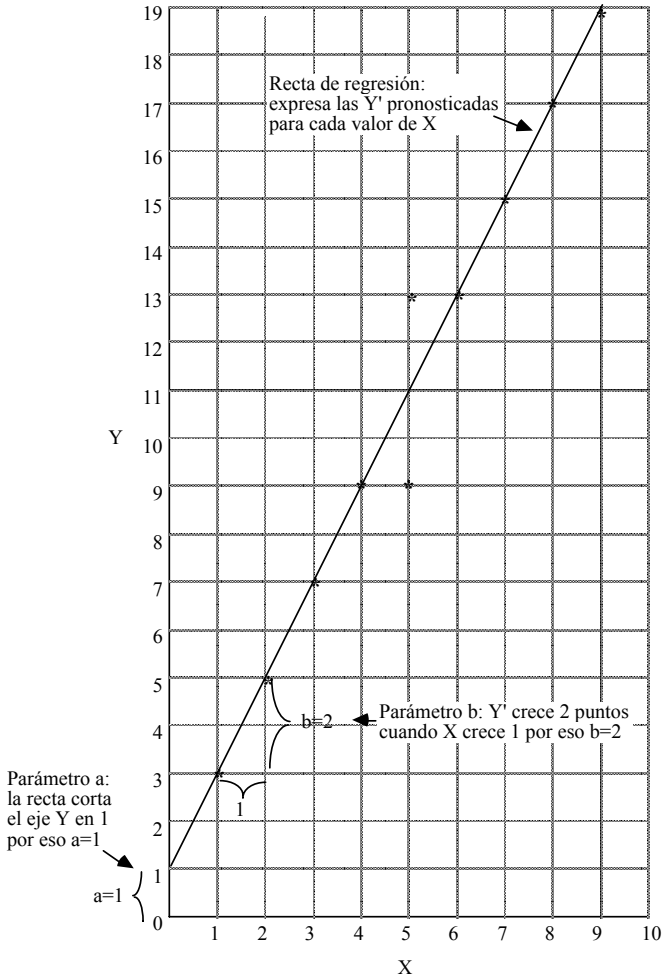
Evidentemente la recta también podría trazarse conociendo un solo punto de la misma (por ejemplo (0,a), en este caso (0,1) y el ángulo α .

También podemos observar en el diagrama de dispersión que, en este caso concreto, la recta describe muy bien los puntos reales.

De hecho la recta de regresión sirve para expresar cual es el pronóstico en Y para cada valor de X.

En el ejemplo, sólo 2 puntos de 10 no están encima de la recta, todos los demás se comportan perfectamente de acuerdo al modelo lineal que describe la ecuación de regresión.

Diagrama de dispersión con la recta de regresión



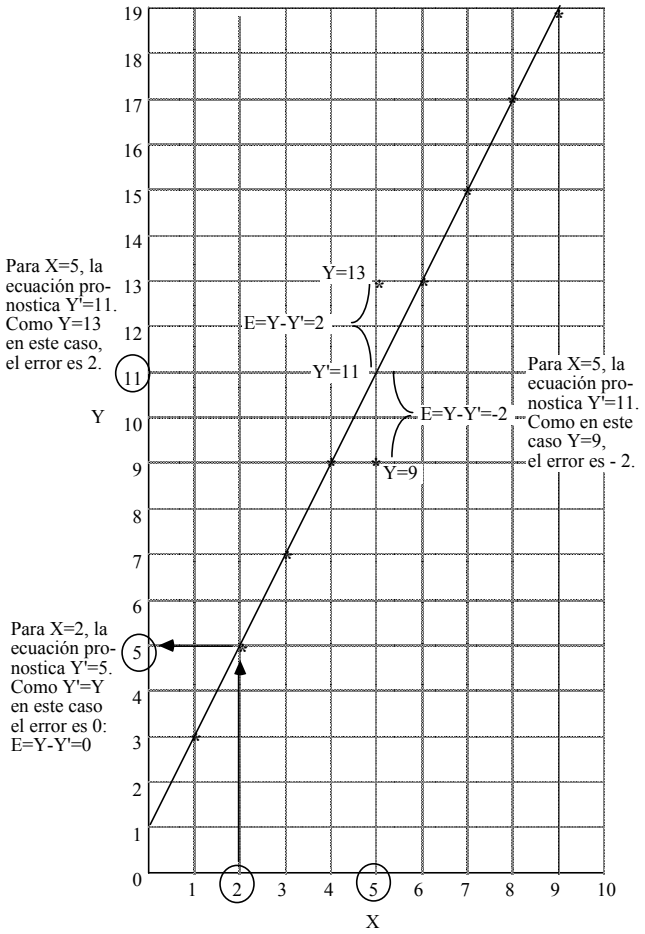
Una vez que disponemos de la ecuación podemos utilizar esta como una "máquina de predecir" que ofrece un pronóstico Y' para cada valor de X que queramos entrar en ella.

Así por ejemplo, si una persona puntúa en el test de aptitud 2 (X=2) la ecuación le pronostica 5.

$$Y' = 1 + 2(2) = 5$$

Esto puede verse gráficamente, y así se ha señalado con flechas en el diagrama siguiente.

Pronósticos y residuales sobre la resta.



Hay un caso en nuestros datos que precisamente había puntuado X=2 e Y=5. Como la ecuación le pronostica Y'=5 para este valor de X, en este caso la ecuación hace un pronóstico perfecto, y el error de estimación es 0

Observemos, sin embargo, que sucede con los dos casos que no caen encima de la recta. En primer lugar tomemos aquel caso que ha puntuado X=5 e Y=13.

Para X=5 la ecuación no pronostica 13, sino 11

$$Y' = 1 + 2(5) = 11$$

Por eso, en este caso la ecuación infraestima el verdadero valor del caso, cometiendo un error de pronóstico o estimación de 2:

$$E = Y - Y' = 13 - 11 = 2$$

Se trata de 2 puntos en la escala de Y. Si la escala de Y está expresada en puntos de un rendimiento académico son 2 puntos de infraestimación y si Y está expresada en millones de pesetas de ventas, son 2 millones de infraestimación de ventas.

En el otro caso que no ajusta a la recta tenemos que la persona ha puntuado X=5 e Y=9. (El hecho de que ambos casos que no ajustan tengan el mismo valor en X es mera casualidad del ejemplo).

Para este caso la ecuación, dado que tiene la misma X=5, también pronostica Y'=11.

$$Y' = 1 + 2(5) = 11$$

Pero en este caso Y=9 por lo que la ecuación ahora sobrestima el comportamiento del caso en la variable Y.

Esto da lugar a un error de estimación de menos 2.

$$E = Y - Y' = 9 - 11 = -2$$

Podríamos seguir calculando el error de estimación para cada caso, elevar al cuadrado esos errores, sumarlos, dividirlos por el número de casos, y hacer la raíz para obtener el error típico de estimación.

	Directas				
Caso:	X	Y	Y'	E	E^2
1	5	13	11	2	4
2	2	5	5	0	0
3	4	9	9	0	0
4	9	19	19	0	0
5	5	9	11	-2	4
6	7	15	15	0	0
7	1	3	3	0	0
8	6	13	13	0	0
9	3	7	7	0	0
10	8	17	17	0	0
Suma	50	110	110	0	8
Media	5	11	11	0	0,8
Raíz					0,894

En el ejemplo solo hay dos casos con errores de estimación. Cada error es de magnitud 2. La suma de los errores cuadráticos es 8. La media de error cuadrático es 0'8. El error típico de estimación es la raíz cuadrada de ese 0'8, es decir 0'894.

Los errores deben ser analizados para observar que estos presentan un comportamiento aleatorio. Para ello pueden utilizarse diversos tipos de gráficos que muestran la distribución de los mismos para cada valor de X. Si la ecuación es adecuada a los datos los errores no pueden estar concentrados de modo desigual, en frecuencia, signo o magnitud para diferentes zonas de la variable.

Puede observarse que si todas estas condiciones se cumplen la determinación de la regresión no es función de la frecuencia de casos que adopten cada valor X. En realidad, si las medias condicionadas describen una recta, bastan dos medias de Y condicionadas a X bien determinadas para poder trazar la recta de regresión correcta.

La homocedasticidad en los modelos de regresión

El ejemplo anterior, con solo dos casos con error de estimación y un 80% de los puntos sobre la resta parece estar hecho a propósito (y en realidad así es) para explicar de un modo fácil los conceptos básicos de la regresión. En la práctica, lo usual en ciencias sociales es encontrar diagramas de dispersión con nubes de puntos que, en el mejor de los casos, describen elipsoides cuya densidad es mayor alrededor de la linea de regresión.

A pesar de su sencillez puede formularse una objeción al ejemplo anterior desde el punto de vista de la adecuación de la ecuación de regresión. Si se reflexiona sobre el comportamiento de los puntos puede verse que la predicción es perfecta para el intervalo de X que va de 0 a 4 y también para el que va de 6 a 9, dentro de la zona donde tenemos datos. Sin embargo, no es así para los valores X=5. En una situación así, con suficiente número de datos en cada distribución condicionada, podría optarse por establecer pronósticos para las dos zonas con diferentes grados de confiabilidad.

Si se observa la cuestión desde el punto de vista de las varianzas de las distribuciones Y condicionadas a los sucesivos valores de X puede observarse que las varianzas son homogéneas entre sí en todas las distribuciones condicionales (y en este caso además felizmente igual a 0 sobre el pronóstico) pero hay una varianza diferente para la distribución de Y condicionada a X=5. Podría decirse que las varianzas de Y no son constantes para los niveles de X. Esta situación se conoce como *heterocedasticidad*.

Para que un modelo de regresión pueda aplicarse adecuadamente para todo el rango de X, las varianzas de las distribuciones de Y condicionadas a los valores de X han de ser al menos semejantes y en el caso ideal iguales entre sí. Esto es lo que se conoce como *homocedasticidad*. Si la variabilidad en torno a la recta no es constante entonces la recta puede estar pronosticando bien en unas zonas y con mucho error en otras. Adicionalmente, como la recta busca el punto de equilibrio minimo-cuadrático será más sensible a los errores grandes producidos allí donde haya más varianza de modo que la estimación obtenida puede volverse inadecuada.

En la práctica la homocedasticidad es más bien una cuestión de grado. Las varianzas de las distribuciones condicionales muy rara vez son todas iguales entre sí. Si no son suficientemente distintas puede sostenerse la homocedasticidad y por tanto, aceptar calcular la ecuación de regresión.

Errores de medición en ecuaciones de regresión

Una parte importante de las mediciones en ciencias sociales están expuestas a errores de medición apreciables. Esos errores de medición significan que debido a defectos del instrumento o del acto de medición principalmente, en lugar del verdadero valor de una variable obtenemos otro más o menos aproximado a aquel. Los errores de medición en la mayoría de los casos se suponen de naturaleza aleatoria en torno a la valor verdadero, con igual probabilidad de producir infraestimaciones que sobrestimaciones de ese valor.

Aunque es frecuente que los investigadores tiendan a ignorarlos como inevitables, los errores de medición tienen importancia en regresión y consecuencias distintas según afecten a la variable dependiente Y o a la independiente X.

Si los errores, siendo de naturaleza aleatoria, suceden en la variable dependiente Y esto no tiene excesiva importancia para la determinación de la recta de regresión, dado que con una muestra adecuada los errores positivos y los negativos se contrabalancearán y las medias condicionadas seguirán ocupando el mismo lugar.

Sin embargo, si los errores se cometen en la variable independiente X los casos quedarán clasificados fuera de su lugar, en distinta distribución condicional de la que les correspondería, y podría obtenerse una linea de regresión equivocada.

Estrictamente, los modelos de regresión suponen variables independientes medidas sin error. En muchos campos éste es un problema de difícil solución pues, precisamente, muchos de los indicadores que se utilizan como predictores se caracterizan por ser mediciones particularmente expuestas a error de medida. En general su uso en validez presupone precisamente, que utilizamos tests (ya también criterios) expuestos a error de medida.

Ejemplos resueltos de estimación de un criterio a partir de un test mediante regresión simple

Ejemplo 1

Sea el test X y el criterio Y para un conjunto de 25 sujetos.

	X	Y
S_1	2.00	5.00
S_2	3.00	5.00
S_3	5.00	9.00
S_4	1.00	2.00
S_5	9.00	10.00
S_6	8.00	9.00
S_7	1.00	1.00
S_8	5.00	7.00
S_9	3.00	4.00
S_{10}	.00	1.00
S_{11}	.00	2.00
S_{12}	1.00	3.00
S_{13}	7.00	10.00
S_{14}	6.00	9.00
S_{15}	5.00	8.00
S_{16}	5.00	5.00
S_{17}	4.00	6.00
S_{18}	4.00	7.00

S_{19}	3.00	6.00
S_{20}	2.00	5.00
S_{21}	1.00	4.00
S_{22}	.00	2.00
S_{23}	7.00	9.00
S_{24}	8.00	10.00
S_{25}	3.00	4.00

Parte 1: Describir las variables: Tabulación, estadísticos descriptivos, dispersión. Dada la descripción de la variable X ¿Qué probabilidad hay de encontrar un sujeto con puntuación X menor o igual a 6? ¿Y con una puntuación en Y mayor a 6?

Tabulación de las variables.

FRECU	ACU FRECU	PCT	ACU PCT	TEST X
3	3	12.0	12.0	0.000
4	7	16.0	28.0	1.000
2	9	8.0	36.0	2.000
4	13	16.0	52.0	3.000
2	15	8.0	60.0	4.000
4	19	16.0	76.0	5.000
1	20	4.0	80.0	6.000
2	22	8.0	88.0	7.000
2	24	8.0	96.0	8.000
1	25	4.0	100.0	9.000

Hay una probabilidad de 0'8 de encontrar un sujeto con una puntuación menor o igual a 6.

FRECU	ACU FRECU	PCT	ACU PCT	CRITERIO Y
2	2	8.0	8.0	1.000
3	5	12.0	20.0	2.000
1	6	4.0	24.0	3.000
3	9	12.0	36.0	4.000
4	13	16.0	52.0	5.000
2	15	8.0	60.0	6.000
2	17	8.0	68.0	7.000
1	18	4.0	72.0	8.000
4	22	16.0	88.0	9.000
3	25	12.0	100.0	10.000

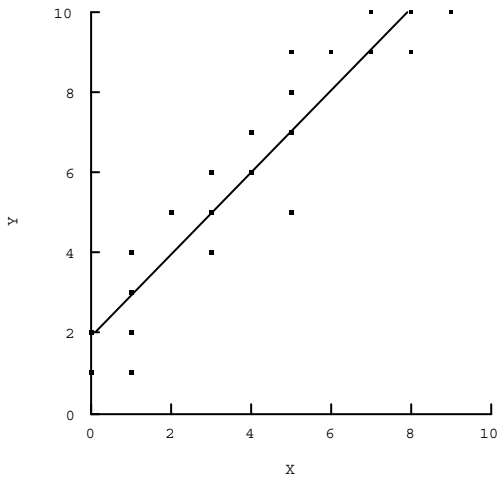
Hay una probabilidad de 0'4 de encontrar un sujeto con una puntuación mayor a 6.

Estadísticos descriptivos:

	X	Y
N	25	25
MINIMO	0.000	1.000
MAXIMO	9.000	10.000
RANGO	9.000	9.000
MEDIA	3.720	5.720
VARIANZA	7.377	8.793
STANDARD DESV	2.716	2.965

Parte 2: Analizar la relación entre ambas variables: diagrama de dispersión, correlación, y regresión simple de Y sobre X.

Diagrama de dispersión (con la recta de regresión trazada)



Análisis de Regresión lineal simple: (Tabla de Resultados característica ofrecida por un paquete estadístico)

DEP VAR: Y N:25 MULTIPLE R: .937 SQUARED MULTIPLE R: .877
ADJUSTED SQUARED MULTIPLE R: .872 STANDARD ERROR OF ESTIMATE: 1.061

VARIABLE	COEFFICIENT	STD ERROR	STD COEF	TOLERANCE	T	P(2 TAIL)
CONSTANT	1.916	0.365	0.000	.	5.251	0.000
X	1.023	0.080	0.937	1.	12.820	0.000

ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	185.130	1	185.130	164.340	0.000
RESIDUAL	25.910	23	1.127		

CORRELACIONES DE PEARSON ENTRE X, Y, Y' y el residual Y-Y'

	X	Y	Y'
X	1.000		
Y	0.937	1.000	
Y'	1.000	0.937	1.000
RESIDUAL	0.000	0.350	0.000

Se ofrece la tabla característica de resultados de un análisis de regresión ofrecida por un paquete estadístico (en este caso SYSTAT) para familiarizar al lector con este formato de presentación de la información que se mantiene relativamente constante entre los principales paquetes estadísticos. En ese "output" de ordenador están todos los estadísticos de interés que hemos analizado y algunos más, relativos a aspectos inferenciales, en los que no entraremos aquí.

Interpretación de los principales resultados

El diagrama de dispersión muestra una relación aproximadamente lineal entre X e Y.

En general a mayor puntuación de X mayor puntuación de Y.

El coeficiente de validez (correlación de Pearson) del test X respecto al criterio Y es 0'937.

El coeficiente de determinación es 0'877

La ecuación de regresión en puntuaciones directas es:

Y'= 1'916 + 1'023 X

La ecuación en puntuaciones diferenciales es:

y' = 1'023 x

La ecuación de regresión en puntuaciones típicas es:

$Z_{y'} = 0'973\ Z_x$

El error típico de estimación es 1'061.

El error típico de estimación que calculan los paquetes estadísticos es la estimación poblacional del mismo. Es decir, con denominador N-2 para regresión lineal simple.

La suma de cuadrados de la regresión es 185'130 y la suma de cuadrados del error es 25'910

Como puede observarse en el diagrama de dispersión y como señala el coeficiente de correlación y su cuadrado el coeficiente de determinación, existe una fuerte relación lineal entre ambas variables. De hecho X explica el 87'7% de la varianza de Y.

En la tabla de correlaciones de Pearson entre X, Y, Y' y el residual E=Y-Y' pueden observarse algunos hechos interesantes sobre regresión:

1º) La correlación entre X e Y' es 1. Esta correlación siempre es 1 (independientemente de los datos). Esto es así debido a que Y' solo es una transformación lineal de X y cualquier variable correlaciona siempre 1 con cualquier transformación lineal de la misma (excepto que la transformación lineal suponga una constante b con signo negativo en cuyo caso siempre correlacionará -1).

2º) La correlación entre X e Y es la misma que la correlación entre Y' e Y. Esto siempre es así debido a que Y' solo es una transformación lineal de X y, por tanto, correlacionará con cualquier otra variable lo mismo que X correlacionara.

3º) Los residuales correlacionan 0 con X y con su transformada Y'. esto siempre es así. Significa que los residuales son independientes de los valores de X, y por tanto de los valores de su transformada Y'. Es decir, los residuales no varían sistemáticamente según se varía X o la Y' pronosticada.

4º) Si elevamos al cuadrado la correlación entre Y e Y' tendremos que el 87'7% de Y está explicado por Y'.

Si elevamos al cuadrado la correlación entre el residual e Y, tendremos que el 12'3% de la varianza de Y es atribuible al residual, es decir, es varianza no explicada. El porcentaje de varianza explicada más el porcentaje de varianza no explicada siempre ha de sumar 100.

Ejemplo resuelto 2

Sean los siguientes datos un test X y un criterio Y para N=4.

X	Y
3	5
2	7
4	13
5	18

Obtener el coeficiente de validez, los parámetros de la regresión lineal simple, la suma de cuadrados de error, la suma de cuadrados de la regresión (= suma de cuadrados explicada por el modelo) y la suma de cuadrados total.

Obtener el coeficiente de determinación y verificar que efectivamente es el cuadrado del coeficiente de correlación de Pearson.

Solución:

El coeficiente de validez (correlación de Pearson) es:

$$r_{xy} = 0'8958$$

Los parámetros de la ecuación de regresión lineal simple son:

$$a = -3'6004$$

$$b = 4'1001$$

Por tanto la ecuación puede escribirse:

$$Y' = -3'6004 + 4'1001 \cdot X$$

En la tabla siguiente se recogen los valores de Y' pronosticados, los residuales al cuadrado $(Y - Y')^2$ y los valores pronosticados menos la media de Y al cuadrado $(Y' - \bar{Y})^2$:

X	Y	Y'	$(Y - Y')^2$	$(Y' - \bar{Y})^2$
3	5	8'6999	13'69	4'2025
2	7	4'6	5'76	37'8225
4	13	12'8	0'04	4'2025
5	18	16'9	1'21	37'8225
Suma			20'7	84'05
Media			3'5	10'75
D.T.			1'118	5'1174

Error típico de estimación $= \sqrt{\frac{20'7}{4}} = 2'2749$

Suma de Cuadrados de Error = 20'7

Suma de cuadrados de la Regresión=84'05

Suma de Cuadrados total = 20'7 + 84'05 = 104'7511

$$R^2 = \frac{84'05}{104'7511} = 0'8024$$

lo que equivale a

$$r_{xy}^2 = 0'8958^2 = 0'8024$$

4. Pronóstico de un criterio por dos o más tests: regresión lineal múltiple

La regresión lineal múltiple es probablemente la técnica multivariada más utilizada. Puede verse como una extensión del análisis de regresión lineal simple. En la regresión lineal múltiple tenemos un criterio Y que actúa como una variable dependiente y dos o más tests X que actúan como variables independientes, como predictores de Y.

Si el criterio Y es una función lineal (aproximadamente) de los tests X, entonces podremos utilizar una ecuación de regresión lineal múltiple para pronosticar el criterio Y a partir de los tests X.

$$Y'_i = a + b_1X_1 + b_2X_2 + ... + b_vX_v$$

donde:

Y'_i refleja la puntuación pronosticada para el sujeto i a partir de sus puntuaciones en las v variables independientes X

a es una constante en la ecuación que indica el valor de Y'_i cuando todas las variables X valen 0. (Gráficamente expresa el punto en que el plano (con 2 variables X) o el hiperplano (con tres o más variables X) cruza el eje de ordenadas.

b_1, b_2, ... b_v son los coeficientes de regresión. Para v variables independientes X hay v coeficientes de regresión b, de modo que en la ecuación de regresión

múltiple cada variable independiente X es multiplicada por su coeficiente de regresión b

Cualquier b_j que multiplica a su respectiva X_j, es una constante en la ecuación que indica cuanto crece Y'_i cuando esa X_j crece 1 *manteniéndose constantes* todas las demás variables X.

El propósito del análisis de regresión múltiple es calcular el valor del coeficiente a y de los v coeficientes b de modo que la suma de las diferencias elevadas al cuadrado entre los valores reales del criterio Y y los valores del criterio Y' (estimados o pronosticados) sea la mínima posible. Es decir, calcular los parámetros de la ecuación de modo que esta cometa el mínimo error de predicción cuadrático posible.

Para poder calcular el valor de los parámetros de la ecuación es necesario disponer de una muestra de casos en los que conozcamos tanto los valores de Y como los de cada X. Una vez calculada la ecuación (es decir, una vez calculados los valores del coeficiente a y de los v coeficientes b) puede utilizarse esta ecuación para pronosticar el valor de Y en nuevos casos de la misma población.

En ocasiones para expresar brevemente la relación entre las variables independientes y la variable dependiente se utiliza una notación funcional, en ese caso se escribe, por ejemplo, $Y' = f(X_1, X_2, \ldots)$

4.1. Residuales y error típico de estimación

La forma de obtención de los residuales y el error típico de estimación es una extensión de la explicada en regresión lineal simple. La diferencia entre el valor real de Y para el sujeto i, representado por Y_i, y el valor de Y pronosticado por la ecuación para ese mismo sujeto, representado por

Y'_i, es un error de estimación o residual que representamos por E_i.

$$E_i = Y_i - Y'_i$$

Cuanto más error de estimación cometa una ecuación tanto peor es capaz de predecir Y a partir de los tests X.

Los residuales forman una nueva variable que expresa la parte de la variable dependiente Y de la que no hemos dado cuenta mediante la ecuación de regresión. Significan la parte del criterio Y que no hemos podido explicar, aquella parte cuya variabilidad no podemos atribuir a los tests predictores.

Un estadístico que indica cuanto error de estimación comete globalmente la ecuación de regresión es el error típico de estimación:

$$s_{Y.X} = \sqrt{\frac{\sum (Y - Y')^2}{N}}$$

La fórmula es análoga a la del error típico de estimación de la regresión lineal simple. En realidad sólo varía como se ha calculado la Y' pronosticada a cada sujeto. En regresión lineal simple la Y' se calcula a partir de un solo predictor; en la regresión lineal múltiple la Y' se pronostica a partir de 2 o más variables independientes. Pero el cálculo de cada error de estimación o residual es el mismo, y la lógica de construcción de la fórmula del error típico de estimación como un indicador global de los errores de estimación, también.

Como en la regresión lineal simple, para el cálculo de los parámetros de la ecuación y para el cálculo del error típico de estimación, es imprescindible disponer de un conjunto de casos en los que se conozca tanto las X como Y. Una vez establecida la ecuación de regresión (es decir, una vez que se ha averiguado el valor de los parámetros de la ecuación para esos datos) ésta podrá utilizarse para estimar Y' en los casos conocidos y, en consecuencia, obtener el error típico de estimación. Conocida la ecuación de regresión, ésta puede utilizarse para pronosticar el valor en Y de otros casos de la misma población de los que solo conozcamos sus puntuaciones en los tests X.

Error típico de estimación expresado mediante coeficientes de correlación

Con sólo 2 variables independientes X_1 y X_2 el error típico de la estimación de Y puede expresarse en términos de las correlaciones de Pearson entre las variables:

$$s_{Y.12} = s_Y \cdot \sqrt{\frac{1 - r^2_{Y1} - r^2_{Y2} - r^2_{12} + 2\, r_{Y1} r_{Y2} r_{12}}{1 - r^2_{12}}}$$

En esta fórmula por simplicidad los subíndices 1 y 2 representan respectivamente a los tests X_1 y X_2.

Si el número de variables independientes es mayor de 2 este tipo de formulación del error típico de estimación, aunque posible, se vuelve excesivamente complicada para que sea práctico su uso.

Error típico de estimación insesgado

De modo análogo a lo que sucedía en regresión lineal simple, el error típico de estimación en la ecuación de regresión lineal múltiple, tal como lo hemos definido, es un estimador sesgado del error típico de estimación de la población. Para evitar este sesgo, a la N del denominador de la fórmula del error típico, se le sustrae un factor k igual al número de parámetros en la ecuación. El número de parámetros en una ecuación de regresión lineal múltiple es igual al número de variables independientes más 1. Por ejemplo, si tenemos dos variables independientes en la ecuación hay que calcular 3 parámetros (uno de intercepción y dos pendientes b asociadas a las variables), por tanto k=3 y el denominador a utilizar en la fórmula del error típico de estimación insesgado es N-3 en lugar de N.

4.2. Estimación de la regresión lineal múltiple

Consideremos el caso más sencillo de regresión lineal múltiple, el caso de dos variables predictoras:

$$Y'_i = a + b_1 X_1 + b_2 X_2$$

Si se expresan las variables en **puntuaciones típicas**, los parámetros estandarizados (señalados con un *) pueden calcularse mediante las siguientes fórmulas, expresadas en términos de los coeficientes de correlación de Pearson entre las tres variables: (Por simplicidad los subíndices 1 y 2 representan a la primera y a la segunda variable independiente X, respectivamente).

$$a^* = 0$$

$$b^*_1 = \frac{r_{Y1} - r_{Y2}r_{12}}{1 - r^2_{12}}$$

$$b^*_2 = \frac{r_{Y2} - r_{Y1}r_{12}}{1 - r^2_{12}}$$

Si se expresan las variables en **puntuaciones diferenciales**, los parámetros pueden calcularse mediante las siguientes fórmulas:

$$a' = 0$$

$$b_1 = \frac{r_{Y1} - r_{Y2}r_{12}}{1 - r^2_{12}} \cdot \frac{s_Y}{s_1}$$

$$b_2 = \frac{r_{Y2} - r_{Y1}r_{12}}{1 - r^2_{12}} \cdot \frac{s_Y}{s_2}$$

Es decir,

$$b_1 = b^*_1 \cdot \frac{s_Y}{s_1}$$

$$b_2 = b^*_2 \cdot \frac{s_Y}{s_2}$$

Si se expresan las variables en **puntuaciones directas**, entonces:

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$$

$$b_1 = b^*_1 \cdot \frac{s_Y}{s_1}$$

$$b_2 = b^*_2 \cdot \frac{s_Y}{s_2}$$

Para 3 o más variables independientes el modo más elegante de expresar el cálculo de coeficientes es mediante álgebra matricial. En la práctica, la ecuación de regresión lineal múltiple se calcula mediante programas o paquetes estadísticos que permiten resolver este tipo de ecuaciones para cualquier número de variables independientes a partir de la matriz de datos.

4.3. El coeficiente de correlación múltiple entre tests y criterio

El coeficiente de correlación de Pearson entre los valores reales de un criterio Y, y los valores pronosticados mediante la ecuación de regresión lineal múltiple (Y'), se conoce como coeficiente de correlación múltiple.

$$R = r_{YY'}$$

El coeficiente de correlación múltiple expresa la relación entre el criterio Y y los tests predictores X tomados conjuntamente en el compuesto lineal predictor que denominamos Y'.

El coeficiente puede hallarse de modo general por este procedimiento:

1) Calcular la ecuación de regresión,

2) Obtener los valores de Y';

3) Calcular la correlación entre Y e Y'.

Para el caso de sólo dos variables independientes, si se dispone de los coeficientes de correlación de Pearson entre las variables (también llamados en este contexto coeficientes de *correlación de orden cero*), puede obtenerse el coeficiente de correlación múltiple aplicando la siguiente fórmula:

$$R_{Y.12} = \sqrt{\frac{r_{Y1}^2 - 2\, r_{Y1} r_{Y2} r_{12} + r_{Y2}^2}{1 - r_{12}^2}}$$

Para más de dos variables independientes el manejo de este tipo de formulación se vuelve demasiado engorrosa.

El coeficiente de correlación múltiple puede tomar valores entre 0 y 1, cuanto más se acerca a 1 mayor es el grado de relación lineal entre Y y los valores pronosticados Y'. Un valor 0 indica ausencia de relación lineal (pero no necesariamente ausencia de relación) y un resultado de 1 indicaría una relación lineal perfecta.

El coeficiente de correlación lineal múltiple elevado al cuadrado se denomina **coeficiente de determinación múltiple** y expresa la proporción de varianza del criterio Y explicada por el compuesto lineal Y'. Es decir, expresa que proporción de Y son capaces de explicar los tests predictores X mediante la ecuación de regresión lineal múltiple. Es frecuente multiplicar por 100 los coeficientes de determinación y hablar de "porcentaje de varianza explicada".

El coeficiente de determinación lineal múltiple es un *estadístico de reducción proporcional del error*, y como tal puede definirse y calcularse como:

$$R^2 = \frac{\sum(Y' - \bar{Y})^2}{\sum(Y - \bar{Y})^2}$$

Esta fórmula, como en cualquier estadístico de este tipo, expresa la relación entre la varianza de la variable dependiente explicada por el modelo de pronóstico (en este caso por la ecuación de regresión lineal múltiple) y la varianza de la variable dependiente a explicar (reflejada, como muestra el denominador, por las diferencias entre las puntuaciones de la variable dependiente y su media).

$$\text{Estadístico de Reducción Proporcional del Error} = \frac{\text{Varianza explicada por el modelo}}{\text{Varianza total a explicar}}$$

Partiendo de la fórmula anterior de puede efectuarse una *definición general del coeficiente de correlación múltiple.* En efecto, tomando la raíz positiva, tenemos:

$$R = + \sqrt{\frac{\sum(Y' - \bar{Y})^2}{\sum(Y - \bar{Y})^2}}$$

Por otra parte, el coeficiente de determinación múltiple permite reexpresar el *error típico de estimación* de un modo más sencillo:

$$s_{y\ .12} = s_y \sqrt{1 - R^2_{y\ .12}}$$

Y, en general, sea cual sea el número de variables independientes X:

$$s_{y\ .x_v} = s_y \sqrt{1 - R^2_{y\ .x_v}}$$

En la fórmula anterior el subíndice x_v se emplea para referirse a todas las v variables independientes X incluidas en la ecuación de regresión lineal múltiple.

4.4. Coeficiente de correlación semiparcial

Al concebir los residuales E=Y-Y' como una nueva variable que expresa la parte de Y de la que no ha dado cuenta una ecuación de regresión determinada, es posible resolver fácilmente cuestiones acerca de la relación entre dos variables cuando de una o de ambas se quiere eliminar primero el influjo o parte explicada por una tercera.

Supongamos que estamos interesados en la relación entre el criterio Y y el test X. Pero sabemos que parte de la variabilidad de Y puede deberse a una tercera variable Z. ¿Podemos calcular la relación entre X y la variable dependiente Y pero "limpia" ésta última del influjo de Z? Efectivamente podemos hacerlo. Para ello podríamos seguir los siguientes pasos:

1) Calcular la regresión lineal simple de Y sobre Z. Obtener los pronósticos Y' y los residuales E=Y-Y'. Esos residuales significan la parte de Y que queda después de sustraer a esta variable aquello que Z es capaz de pronosticar linealmente.

2) Calcular la correlación de Pearson entre E y X.

A esa correlación se la conoce como **coeficiente de correlación semiparcial** $R_{(y.z)x}$. Es un coeficiente de correlación entre X y E=Y-Y', siendo Y'=a + bZ y expresa la relación de X con la parte de Y libre de relación lineal con Z.

El procedimiento de cálculo descrito es de naturaleza general y puede extenderse fácilmente a más de una tercera variable. Si por ejemplo deseáramos obtener la correlación entre X e Y, estando Y libre de Z y W bastaría con obtener Y' como función lineal múltiple de Z y W, para proceder a continuación del mismo modo correlacionando los residuales E=Y-Y' con X.

Para el caso en que en la correlación entre X e Y sólo se desea eliminar el efecto de una tercera variable Z sobre la variable Y, la correlación semi-parcial puede calcularse a partir de los coeficientes de correlación de orden cero entre

las variables (los "coeficientes de correlación de orden cero" es como se suele llamar en este contexto a los coeficientes de correlación de Pearson):

$$R_{(y\,.\,z)\,x} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{1 - r_{yz}^2}}$$

4.5. Coeficiente de correlación parcial

El concepto de correlación semiparcial se puede generalizar al caso en que deseamos la correlación de Pearson entre un test X y un criterio Y excluido *de cada una de ellos* el efecto de una (o más de una) tercera variable Z.

El procedimiento de cálculo de la correlación parcial es una extensión del descrito para la correlación semiparcial y puede aplicarse utilizando las combinaciones de terceras variables que se requiera.

Supongamos que deseamos obtener la correlación entre X e Y, habiendo eliminado de cada una de ellas el influjo lineal de una tercera variable Z. El método de cálculo de la correlación parcial correspondiente sería el siguiente:

1) Calcular la regresión lineal simple de Y sobre Z. Obtener los pronósticos Y' y los residuales Ey=Y-Y'. Esos residuales significan la parte de Y que queda después de sustraer a esta variable aquello que Z es capaz de pronosticar linealmente.

2) Calcular la regresión lineal simple de X sobre Z. Obtener los pronósticos X' y los residuales Ex=X-X'. Esos residuales significan la parte de X que queda después de sustraer a esta variable aquella parte que Z es capaz de pronosticar linealmente.

2) Calcular la correlación de Pearson entre Ey y Ex. Esa correlación de Pearson es la correlación parcial entre X e Y eliminado de ambas el influjo de Z.

Si hubiéramos deseado, por ejemplo, considerar la parte de Y "limpia" del influjo de Z y W hubiera bastado con calcular los residuales E_Y=Y-Y' utilizando para calcular Y' la correspondiente ecuación de regresión lineal múltiple. El procedimiento, utilizando este método general de cálculo no requiere la restricción de que de ambas variables a correlacionar se sustraiga previamente el efecto de la misma o las mismas variables.

Si se trata de la correlación entre X e Y eliminando previamente de ambas el influjo lineal de Z, la correlación parcial puede obtenerse en ese caso a partir de los coeficientes de correlación de orden cero entre las variables (coeficientes de correlación de Pearson):

$$R_{xy\,.\,z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{1 - r_{xz}^2}\sqrt{1 - r_{yz}^2}}$$

La correlación parcial $R_{xy\,.\,z}$ significa en ese caso la correlación entre X e Y eliminado de ambas variables el influjo de una tercera variable Z. Puede escribirse por tanto:

$$R_{xy\,.\,z} = R_{(X-X')(Y-Y')}$$

4.6. Un ejemplo numérico

En el siguiente ejemplo vamos a aplicar los conceptos adquiridos acerca de correlación, regresión simple, regresión múltiple, correlación múltiple, semi-parcial y parcial, y vamos a poner en conexión algunos de ellos mostrando diversas aproximaciones e interpretaciones de esos coeficientes. Para estructurar los análisis los hemos divido por partes.

Los siguientes datos representan los valores de 3 variables (X, Y, Z) para 5 casos.

	X	Y	Z
S_1	1	2	5
S_2	3	5	5
S_3	5	7	4
S_4	7	6	3
S_5	9	8	1

Parte Primera.

1. Media, Varianza y Desviación Típica de cada variable.

2. Las correlaciones de Pearson entre las tres variables:

   2.1. Efectuando el cálculo utilizando puntuaciones directas

   2.2. Utilizando puntuaciones diferenciales.

   2.3 Utilizando medias y desviaciones típicas.

   2.4 Correlación de Pearson entre X e Y como promedio de los productos de las puntuaciones típicas.

   2.5. Correlación de Pearson entre X e Y como diferencias de las puntuaciones típicas.

   2.6. Coeficientes de determinación.

   2.7. Interpretación.

3. Coeficientes de determinación múltiple y correlaciones múltiples entre las tres variables. Interpretación.

4. Correlaciones semiparciales entre las tres variables. Interpretación.

5. Correlaciones parciales entre las tres variables. Interpretación.

6. Regresión simple de X sobre Y y regresión simple de X sobre Z.

7. Regresión múltiple de X sobre Y y Z.

1. Media, varianza y desviación típica de cada variable

Variable X:

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{25}{5} = 5$$

$$s_x^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N} = \frac{\sum_{i=1}^{N} x_i^2}{N} = \frac{\sum_{i=1}^{N} X_i^2}{N} - \bar{X}^2 = \frac{165}{5} - 5^2 = 8$$

$$s_x = \sqrt{s_x^2} = \sqrt{8} = 2'8284$$

Variable Y:

$$\bar{Y} = \frac{\sum_{i=1}^{N} Y_i}{N} = \frac{28}{5} = 5'6$$

$$s_y^2 = \frac{\sum_{i=1}^{N} (Y_i - \bar{Y})^2}{N} = \frac{\sum_{i=1}^{N} y_i^2}{N} = \frac{\sum_{i=1}^{N} Y_i^2}{N} - \bar{Y}^2 = \frac{178}{5} - 5'6^2 = 4'24$$

$$s_y = \sqrt{s_y^2} = \sqrt{4'24} = 2'0591$$

Variable Z:

$$\bar{Z} = \frac{\sum_{i=1}^{N} Z_i}{N} = \frac{18}{5} = 3'6$$

$$s_z^2 = \frac{\sum_{i=1}^{N} (Z_i - \bar{Z})^2}{N} = \frac{\sum_{i=1}^{N} z_i^2}{N} = \frac{\sum_{i=1}^{N} Z_i^2}{N} - \bar{Z}^2 = \frac{76}{5} - 3'6^2 = 2'24$$

$$s_z = \sqrt{s_z^2} = \sqrt{2'24} = 1'4966$$

2. Correlaciones de Pearson entre las tres variables

$$r_{xy} = \frac{n\sum XY - \sum X \cdot \sum Y}{\sqrt{(n\sum X^2 - (\sum X)^2)\,(n\sum Y^2 - (\sum Y)^2)}}$$

2.1. Correlaciones de Pearson utilizando puntuaciones directas

	X	Y	Z		XY	XZ	YZ
S_1	1	2	5		2	5	10
S_2	3	5	5		15	15	25
S_3	5	7	4		35	20	28
S_4	7	6	3		42	21	18
S_5	9	8	1		72	9	8
	25	28	18		166	70	89

$$r_{xy} = \frac{5 \cdot 166 - 25 \cdot 28}{\sqrt{(5 \cdot 165 - 25^2)\,(5 \cdot 178 - 28^2)}} = 0'8928$$

$$r_{xz} = \frac{5 \cdot 70 - 25 \cdot 18}{\sqrt{(5 \cdot 165 - 25^2)\,(5 \cdot 76 - 18^2)}} = -0'9449$$

$$r_{yz} = \frac{5 \cdot 89 - 28 \cdot 18}{\sqrt{(5 \cdot 178 - 28^2)\,(5 \cdot 76 - 18^2)}} = -0'7658$$

2.2. Correlaciones de Pearson utilizando puntuaciones diferenciales

	X	Y	Z	x	y	z	xy	xz	yz
S_1	1	2	5	-4	-3'6	1'4	14'4	-5'6	-5'04
S_2	3	5	5	-2	-0'6	1'4	1'2	-2'8	-0'84
S_3	5	7	4	0	1'4	0'4	0	0	0'56
S_4	7	6	3	2	0'4	-0'6	0'8	-1'2	-0'24
S_5	9	8	1	4	2'4	-2'6	9'6	-10'4	-6'24
$\sum$	25	28	18	0	0	0	26	-20	-11'8
M:	5	5'6	3'6		Cov :	5'2	-4	-2'36	
d.t. :	2'828	2'059	1'497						

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{5'2}{2'8284 \cdot 2'0591} = 0'8928$$

$$r_{xz} = \frac{-4}{2'8284 \cdot 1'4966} = -0'9449$$

$$r_{yz} = \frac{-2'36}{2'0591 \cdot 1'4966} = -0'7658$$

2.3. Correlaciones de Pearson utilizando medias y desviaciones típicas

	X	Y	Z		XY	XZ	YZ
S_1	1	2	5		2	5	10
S_2	3	5	5		15	15	25
S_3	5	7	4		35	20	28
S_4	7	6	3		42	21	18
S_5	9	8	1		72	9	8
$\sum$	25	28	18		166	70	89
M:	5	5'6	3'6		33'2	14	17'8
d.t. :	2'8284	2'0591	1'4966				

$$r_{xy} = \frac{\overline{XY} - \bar{X}\bar{Y}}{s_x s_y} = \frac{33'2 - 5 \cdot 5'6}{2'8284 \cdot 2'0591} = 0'8928$$

$$r_{xy} = \frac{\overline{XZ} - \bar{X}\bar{Z}}{s_x s_z} = \frac{14 - 5 \cdot 3'6}{2'8284 \cdot 1'4966} = -0'9449$$

$$r_{xy} = \frac{\overline{YZ} - \bar{Y}\bar{Z}}{s_y s_z} = \frac{17'8 - 5'6 \cdot 3'6}{2'0591 \cdot 1'4966} = -0'7658$$

2.4. Correlación de Pearson entre X e Y utilizando la fórmula del promedio de los productos de las puntuaciones típicas

z_x	z_y	$z_x z_y$
-1'4142	-1'7483	2'4724
-0'7071	-0'2913	0'2059
0	0'6799	0
0'7071	0'1942	0'1373
1'4142	1'1655	1'6482

$$r_{xy} = \frac{\sum z_x z_y}{N} = \frac{4'4638}{5} = 0'8927$$

2.5. Correlación de Pearson entre X e Y utilizando la fórmula de las diferencias de las puntuaciones típicas

z_x	z_y	$z_x - z_y$	$(z_x - z_y)^2$
-1'4142	-1'7483	0'3341	0'1116
-0'7071	-0'2913	-0'4158	0'1728
0	0'6799	-0'6799	0'4622
0'7071	0'1942	0'5129	0'2630
1'4142	1'1655	0'2487	0'0618

$$r_{xy} = 1 - \frac{1}{2} \cdot \frac{\sum (z_x - z_y)^2}{N} = 1 - \frac{1}{2} \cdot \frac{1'0714}{5} = 0'8928$$

2.6. Coeficientes de determinación

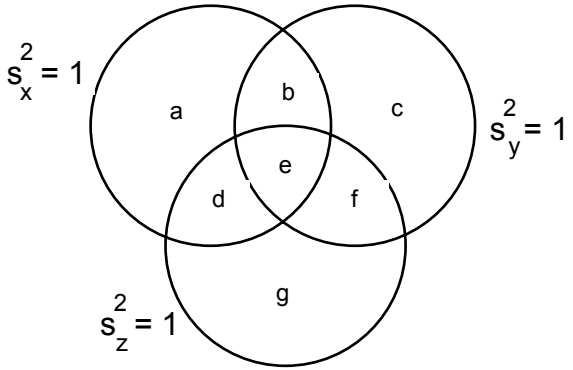
$$r^2_{x\,y} = 0'8928^{\,2} = 0'7970$$

$$r^2_{x\,z} = -\,0'9449^{\,2} = 0'8928$$

$$r^2_{y\,z} = -\,0'7658^{\,2} = 0'5864$$

2.7. Interpretación

Los diagramas de Euler-Venn ayudan a "ver" algunas de las relaciones entre las variables que expresan los coeficientes de correlación. Para mayor simplicidad, vamos a representar las varianzas de las variables por un círculo de área igual a la unidad: $s^2_x = 1;\quad s^2_y = 1;\quad s^2_z = 1$, lo que ocurre trabajando en puntuaciones típicas.



La intersección entre cada par de círculos representará r^2 que en este caso coincide con la covarianza al cuadrado ya que, trabajando en puntuaciones típicas:

$$r^2_{xy} = \frac{s^2_{xy}}{s^2_x\, s^2_y} = \frac{s^2_{xy}}{1\cdot 1} = s^2_{xy}$$

Si las variables están expresadas en típicas, la covarianza al cuadrado entre las puntuaciones típicas equivale al coeficiente de determinación.

Por tanto, sobre el diagrama:

$$r^2_{xy} = s^2_{xy} = s^2_x \cap s^2_y = b + e = 0'79$$

$$r^2_{xz} = s^2_{xz} = s^2_x \cap s^2_z = d + e = 0'89$$

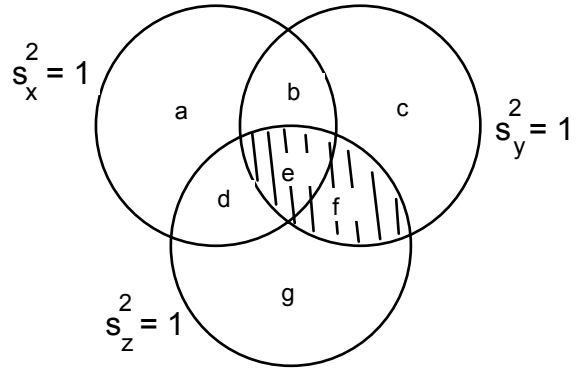
$$r^2_{yz} = s^2_{yz} = s^2_y \cap s^2_z = f + e = 0'58$$

donde las letras a, b, c, d, e, f y g expresan las áreas correspondientes de la figura anterior.

Es decir, el coeficiente de determinación se refiere a la proporción de varianza explicada por una variable sobre otra (o varianza compartida, o covarianza de las puntuaciones típicas).

Por ejemplo, $r^2_{yz} = 0'58$ significa que Y explica el 58% de la varianza de Z, o también que Z explica el 58% de la varianza de Y.

Gráficamente corresponde a la zona f + e.



3. Coeficientes de determinación múltiples y correlaciones múltiples entre las tres variables. Interpretación

Coeficientes de determinación múltiples:

Coeficiente de determinación múltiple de X con Las variables Y y Z tomadas conjuntamente:

$$R^2_{x.yz} = \frac{r^2_{xy} + r^2_{xz} - 2\, r_{xy}\, r_{xz}\, r_{yz}}{1 - r^2_{yz}}$$

$$R^2_{x.\,yz} = \frac{0'7970 + 0'8928 - 2 \cdot 0'8928 \cdot (-0'9449)(-0'7657)}{1 - 0'5862} = 0'9618$$

Coeficiente de determinación múltiple de Y con Las variables X y Z tomadas conjuntamente:

$$R^2_{y.\,xz} = 0'8526$$

Coeficiente de determinación múltiple de Z con Las variables X e Y tomadas conjuntamente:

$$R^2_{z.\,xy} = 0'9221$$

De donde obtenemos los siguientes **coeficientes de correlación múltiple**:

$$R_{x.\,yz} = \sqrt{0'9618} = 0'9807$$

$$R_{y.\,xz} = \sqrt{0'8526} = 0'9233$$

$$R_{z.\,xy} = \sqrt{0'9221} = 0'9602$$

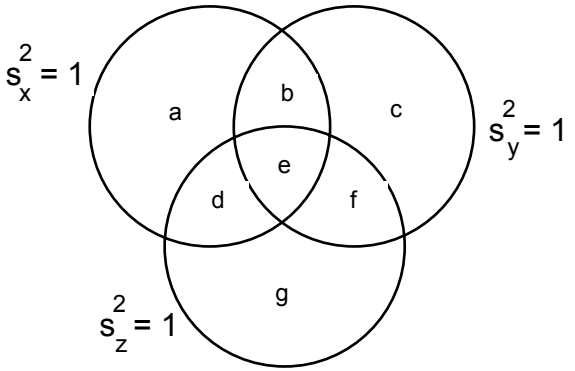
Interpretación:

El coeficiente de determinación múltiple indica que proporción de la varianza de una variable es explicada por otras (dos o más) tomadas conjuntamente. Por ejemplo,

$$R^2_{z.\,xy} = 0'9221$$

indica que el 92'21% de la variable Z es explicada por las variables X e Y.

Expresado en los diagramas de Euler-Venn:



$$R^2_{x.\,yz} = 0'9618 = b + e + d$$

$$R^2_{y.xz} = 0'8526 = b + e + f$$

$$R^2_{z.xy} = 0'9221 = d + e + f$$

Es decir, los coeficientes de determinación múltiple se refieren a la proporción de varianza explicada conjuntamente por las variables predictoras sobre la variable criterio.

7.4. Correlaciones semiparciales entre las tres variables

Dado que se trata de calcular correlaciones semiparciales en las que sólo participan tres variables, podemos obtenerlas utilizando la fórmula basada en los coeficientes de correlación de orden cero (coeficiente de correlación de Pearson entre las variables). De este modo tenemos;

$$R_{(x.z)y} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{1 - r^2_{xz}}} = \frac{0'8928 - (-0'9449)(-0'7657)}{\sqrt{1 - 0'8928}} = 0'5171$$

$$R_{(y.z)x} = 0'2632$$

$$R_{(x.y)z} = -0'5800$$

$$R_{(z.y)x} = -0'4062$$

$$R_{(z.x)y} = 0'2379$$

$$R_{(y.x)z} = 0'1729$$

Interpretación:

La relación entre las variable Y y Z está simbolizada por la zona e + f. Si estudiamos, por ejemplo, la correlación entre X e Y, eliminado el influjo de la variable Z sobre la variable Y, es decir, $R_{(y.z)x}$, desaparece la parte correspondiente a la zona e + f. La zona e+f puede considerarse la parte de Y que puede estimarse desde Z y simbolizarse como Y' (Y estimada desde Z).

La parte de Y restante, es decir, b + c, es la parte de Y que no puede estimarse desde Z, es decir es Y-Y'.

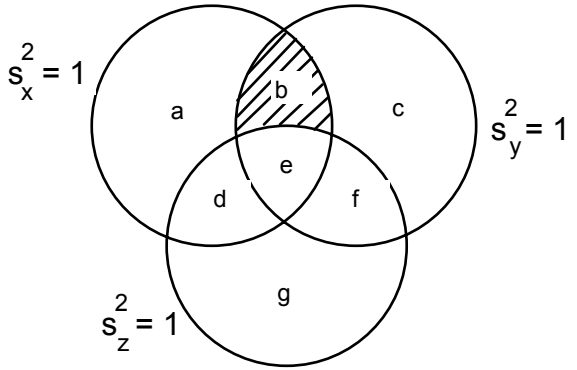
Así el coeficiente de correlación semiparcial $R_{(y.z)x}$ o coeficiente de correlación entre X e (Y-Y') es:

$$R_{(y-y')x} = R_{(y.z)x} = 0'2632$$

El coeficiente de correlación semiparcial elevado al cuadrado:

$$R^2_{(y.z)x} = 0'2632^2 = 0'0692$$

relativo a la correlación $R_{(y.z)x}$ entre X e (Y-Y'), corresponde gráficamente a la varianza representada por la zona b



El coeficiente $R^2_{(y.z)x}$ (zona b) puede entenderse como el coeficiente el coeficiente $R^2_{x.yz}$ (zona b + d + e) menos el coeficiente r^2_{xz} (zona d + e).

Efectivamente esta relación se sostiene entre los coeficientes:

$$R^2_{(y.z)x} = R^2_{x.yz} - r^2_{xz}$$

desarrollando $R^2_{x.\,yz}$ tenemos:

$$R^2_{(y.\,z)\,x} = \frac{r^2_{xy} + r^2_{xz} - 2r_{xy}\,r_{xz}\,r_{yz}}{1 - r^2_{yz}} - r^2_{xz} =$$

$$= \frac{r^2_{xy} + r^2_{xz} - 2\,r_{xy}\,r_{xz}\,r_{yz} - r^2_{xz} + r^2_{xz}\,r^2_{yz}}{1 - r^2_{yz}} =$$

$$= \frac{(r_{xy} - r_{xz}\,r_{yz})^2}{1 - r^2_{yz}}$$

De donde:

$$R_{(y.\,z)\,x} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{1 - r^2_{yz}}}$$

También podemos efectuar la comprobación para nuestros datos:

$$R^2_{(y.\,z)\,x} = 0'0693 \qquad R^2_{x.\,yz} = 0'9618 \qquad r^2_{xz} = 0'8928$$

$$R^2_{(y.\,z)\,x} = R^2_{x.\,yz} - r^2_{xz} = 0'9618 - 0'8928 = 0'069$$

(zona b del gráfico)

Del mismo modo puede procederse con las demás correlaciones semiparciales.

En suma, el *cuadrado de la correlación semiparcial* indica la varianza explicada en la variable dependiente o criterio (sea X) por otra variable (sea Y), sobre y además de la explicada por una tercera (sea Z). Dicho de otra forma, la contribución propia e independiente de Y a la varianza de X. Dicho de otra forma, la relación de X con Y cuando en Y se consideran sólo las puntuaciones residuales no relacionadas con una tercera variable Z. Estas relaciones podrán verse con más claridad después, al enfocar la cuestión desde las puntuaciones residuales.

La representación gráfica de las variables como diagramas de Venn *es sólo una metáfora o recurso didáctico que tiene limitaciones.* Ayuda a entender algunos conceptos básicos pero no hay una correspondencia perfecta entre los coeficientes de correlación y sus cuadrados los coeficientes de determinación y las áreas o zonas de unión e intersección entre los círculos.

Por ejemplo, gráficamente puede verse que $R^2_{(x.z)y}$ corresponde a la zona b, del mismo modo que $R^2_{(y.z)x}$ corresponde gráficamente a la zona b, lo que induce a pensar que ambos coeficientes de determinación son iguales, lo que es *falso* dado que:

$$R_{(x.z)y} = 0'5171$$

$$R_{(y.z)x} = 0'2632$$

como hemos calculado anteriormente.

Pueden obtenerse diversas aparentes paradojas de este tipo calculando el valor de la varianza correspondiente a cada una de las zonas del diagrama desde diferentes caminos. La metáfora gráfica ayuda a comprender e interpretar determinadas relaciones entre variables pero no debe olvidarse que es, tan solo, una metáfora gráfica.

5. Correlaciones parciales entre las tres variables

Calcularemos las correlaciones parciales a partir de los coeficientes de orden cero:

$$R_{xy.z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{1-r^2_{xz}}\sqrt{1-r^2_{yz}}} = \frac{0'8928 - (-0'9449)(-0'7657)}{\sqrt{(1-0'8928)(1-0'5862)}} = 0'8042$$

$$R_{xz.y} = -0'9019$$

$$R_{yz.x} = 0'5284$$

La correlación parcial $R_{xy.z}$ significa la correlación entre X e Y eliminado de ambas variables el influjo de una tercera variable Z. Por tanto,

$$R_{xy.z} = R_{(X-X')(Y-Y')}$$

y lo mismo puede expresarse en términos de proporciones de varianza:

$$R^2_{xy.z} = R^2_{(X-X')(Y-Y')}$$

Gráficamente puede observarse que esta expresión *corresponde también a la zona b.*

Por supuesto esto no significa que esta correlación parcial al cuadrado:

$$R^2_{xy\,.\,z} = 0'6467$$

sea igual a las semiparciales al cuadrado que también quedan representadas por la misma zona b:

$$R^2_{(y\,.\,z)\,x} = 0'0693$$

$$R^2_{(x.\,z)\,y} = 0'2674$$

La metáfora gráfica de los diagramas de Euler-Venn no resulta adecuada para expresar la diferencia entre estos coeficientes.

En la correlación semiparcial la varianza explicada $R^2_{(y\,.\,z)\,x}$ se refiere al total de s^2_x. Ahora, en la correlación parcial, la varianza explicada $R^2_{xy\,.\,z}$, es sobre una parte de s^2_x, precisamente sobre 1- r^2_{xz}.

Así, realmente:

$$R^2_{xy\,.\,z} = \frac{R^2_{(y\,.\,z)\,x}}{1 - r^2_{xz}}$$

Dado que:

$$R^2_{(y\,.\,z)\,x} = R^2_{x.\,yz} - r^2_{xz}$$

entonces:

$$R^2_{xy\,.\,z} = \frac{R^2_{x.\,yz} - r^2_{xz}}{1 - r^2_{xz}}$$

o también:

$$R_{xy\ .\ z} = \frac{R_{(y\ .\ z)\ x}}{\sqrt{1 - r^2_{xz}}} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{1 - r^2_{xz}}\sqrt{1 - r^2_{yz}}}$$

En suma, el *cuadrado del coeficiente de correlación parcial* indica la varianza explicada en la variable dependiente o criterio (sea, en este ejemplo, X) por otra variable (sea Y), cuando se elimina en ambas el influjo de una tercera variable (sea Z). Dicho de otra forma, cuando se considera en ambas variables (X e Y) solamente las puntuaciones residuales no relacionadas con una tercera variable (Z).

6. Regresión lineal simple de X sobre Y y regresión lineal simple de X sobre Z

La ecuación de la recta de regresión de X sobre Y es:

X'= a + bY

Expresada en puntuaciones típicas:

a = 0 $\quad$ b = r_{xy} =0'8928

por tanto:

$z_{x'} = 0'8928\ z_y$

Expresada en puntuaciones diferenciales:

a = 0

$$b = r_{xy}\frac{s_x}{s_y} = 0'8928\ \frac{2'8284}{2'0591} = 1'2263$$

por tanto:

x' = 1'2263 y

Expresada en puntuaciones directas:

$$a = \bar{X} - b\bar{Y} = 5 - 1'2263\ \cdot 5'6 = -1'8672$$

$$b = r_{xy}\frac{s_x}{s_y} = 0'8928\ \frac{2'8284}{2'0591} = 1'2263$$

por tanto:

X' = -1'8672 + 1'2263 Y

El valor del parámetro b es siempre el mismo en puntuaciones directas y en puntuaciones diferenciales. El parámetro a es siempre igual a 0 en puntuaciones diferenciales y típicas.

La ecuación de la recta de regresión de X sobre Z es:

X' = a + bZ

Expresadas las variables X y Z en puntuaciones típicas:

a = 0 $b = r_{xz}$ = -0'9449

por tanto:

$z_{x'}$ = -0'9449 z_z

En puntuaciones diferenciales:

a = 0

$$b = r_{xz}\frac{s_x}{s_z} = -0'9449\ \frac{2'8284}{1'4966} = -1'7857$$

por tanto:

x' = -1'7857 z

En puntuaciones directas:

$$a = \bar{X} - b\bar{Z} = 5 - (-1'7857)\cdot 3'6 = 11'4285$$

$$b = r_{xz}\frac{s_x}{s_z} = -1'7857$$

por tanto:

X' = 11'4285 - 1'7857 Z

7. Regresión múltiple de X sobre Y y Z

La ecuación del plano de regresión de X sobre Y y Z es:

$X' = a + b_1 Y + b_2 Z$

Expresadas las variables en puntuaciones típicas:

a = 0

$$b'_1 = \frac{r_{xy} - r_{yz}r_{xz}}{1 - r_{yz}^2} = \frac{0'8928 - (-0'7657)(-0'9449)}{1 - 0'5862} = 0'4091$$

$$b'_2 = \frac{r_{xz} - r_{yz}r_{yx}}{1 - r_{yz}^2} = \frac{(-0'9449) - (-0'7657)(0'8928)}{1 - 0'5862} = -0'6314$$

por tanto:

$z_{x'}$ = 0'4091 z_y - 0'6314 z_z

Expresadas las variables en puntuaciones diferenciales:

a = 0

$$b_1 = b'_1 \frac{s_x}{s_y} = 0'4091 \ \frac{2'8284}{2'0591} = 0'5619$$

$$b_2 = b'_2 \frac{s_x}{s_z} = -0'6314 \ \frac{2'8284}{1'4966} = -1'1932$$

por tanto:

x' = 0'5619 y - 1'1932 z

Expresadas en puntuaciones directas:

$$a = \bar{X} - b_1\bar{Y} - b_2\bar{Z} = 5 - (0'5619)(5'6) - (-1'1932)(3'6) = 6'1489$$

$$b_1 = b'_1 \frac{s_x}{s_y} = 0'5619$$

$$b_2 = b'_2 \frac{s_x}{s_z} = -1'1932$$

por tanto:

X' = 6'1489 + 0'5619 Y - 1'1932 Z

Parte Segunda.

En esta segunda parte vamos a enfocar los análisis utilizando las puntuaciones pronosticadas por las ecuaciones de regresión y sus residuales.

8. Puntuaciones pronosticadas X' obtenidas utilizando la regresión múltiple de X sobre Y y Z.

9. Puntuaciones residuales X - X'.

10. Error típico de estimación de X mediante la regresión múltiple de X sobre Y y Z.

11. Correlación múltiple de X con Y y Z utilizando los valores X'. Obtener el coeficiente de determinación correspondiente.

12. Coeficiente de correlación múltiple de X con Y y Z utilizando los coeficientes b' de la ecuación de regresión estandarizada.

13. Correlación múltiple como varianza de las puntuaciones pronosticadas en X (expresadas en típicas).

14. Calcular la correlación semi-parcial $R_{(x.z)y}$ utilizando los residuales de la regresión de X sobre Z.

15. Calcular la correlación parcial $R_{xy.z}$ utilizando los residuales de la regresión de X sobre Z y los residuales de la regresión de Y sobre Z.

8. Puntuaciones pronosticadas X' obtenidas utilizando la regresión múltiple de X sobre Y y Z.

La ecuación de regresión múltiple, obtenida anteriormente, es:

X' = 6'1489 + 0'5619 Y - 1'1932 Z

Esta ecuación nos permite calcular valores estimados o pronosticados en la variable X a partir de los valores en Y y Z. Aplicándola sobre los datos tenemos:

	Y	Z	Valor pronosticado: X'
s_1	2	5	X'_1 = 6'1489 + 0'5619 (2) - 1'1932 (5) = 1'3067
s_2	5	5	X'_2 = 6'1489 + 0'5619 (5) - 1'1932 (5) = 2'9924
s_3	7	4	X'_3 = 6'1489 + 0'5619 (7) - 1'1932 (4) = 5'3094
s_4	6	3	X'_4 = 6'1489 + 0'5619 (6) - 1'1932 (3) = 5'9407
s_5	8	1	X'_5 = 6'1489 + 0'5619 (8) - 1'1932 (1) = 9'4509

9. Puntuaciones residuales X - X'

Los residuales o errores de estimación E se calculan simplemente: E = X - X'

X	X'	E
1	1'3067	-0'3067
3	2'9924	0'0076
5	5'3094	-0'3094
7	5'9407	1'0593
9	9'4509	0'4509

10. Error típico de estimación de X mediante la regresión múltiple de X sobre Y y Z

Obtención del error típico de estimación a partir de los residuales:

X	X'	E	E^2
1	1'3067	-0'3067	0'0940
3	2'9924	0'0076	0'0000
5	5'3094	-0'3094	0'0957
7	5'9407	1'0593	1'1221
9	9'4509	0'4509	0'2033

$$E.T.E. = s_{x\,.\,yz} = \sqrt{\frac{\sum (X - X')^2}{N}} = \sqrt{\frac{\sum E^2}{N}} = \sqrt{\frac{1'5151}{5}} = 0'5504$$

Obtención del error típico de estimación a partir de la correlación múltiple:

Por otra parte sabemos que:

$$s_{x\,.\,yz} = s_x \sqrt{1 - r^2_{x\,.\,yz}}$$

por tanto, aplicando esta fórmula:

$$s_{x\,.\,yz} = 2'8284 \sqrt{1 - 0'9618} = 0'55$$

11. Correlación múltiple de X con Y y Z utilizando los valores X'. Coeficiente de determinación correspondiente.

La correlación múltiple $R_{x.yz}$ es igual a la correlación entre X y X', es decir:

$R_{x.yz} = r_{xx'}$

Podemos calcular pues $R_{x.yz}$ como la correlación producto-momento entre x y x'.

X	X'	X^2	X'^2	XX'
1	1'3067	1	1'7074	1'3067
3	2'9924	9	8'9544	8'9772
5	5'3094	25	28'1897	26'547
7	5'9407	49	35'2919	41'5849
9	9'4509	81	89'3195	85'0581
25	25'0001	165	163'4629	163'4739

$$r_{xx'} = \frac{n\sum XX' - \sum X \sum X'}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum X'^2 - (\sum X')^2)}}$$

$$r_{xx'} = \frac{5\,(163'4739) - 25\,(25'0001)}{\sqrt{(5 \cdot 165 - 625)(5 \cdot 163'4629 - 625'005)}} = 0'9809$$

$$R_{x.yz} = 0'9809$$

De donde obtenemos el coeficiente de determinación múltiple:

$$R^2_{x\,.\,yz} = 0'9619$$

12. Coeficiente de correlación múltiple de X con Y y Z utilizando los coeficientes b' de la ecuación de regresión estandarizada

El coeficiente de correlación múltiple de X con Y y Z también puede obtenerse utilizando los coeficientes b' de la ecuación de regresión estandarizada mediante la siguiente fórmula:

$$R_{x\,.\,yz} = \sqrt{b'_1 r_{xy} + b'_2 r_{xz}}$$

$$R_{x\,.\,yz} = \sqrt{0'4091 \cdot 0'8928 + (-0'6314)(-0'9449)} = 0'9807$$

13. Coeficiente de correlación múltiple como varianza de las puntuaciones pronosticadas en X (expresadas en típicas).

El coeficiente de correlación múltiple puede obtenerse como varianza de las puntuaciones estimadas en X (expresadas en típicas).

Primero necesitamos obtener las puntuaciones de las tres variables expresadas en típicas. (Para ello dividimos las puntuaciones diferenciales del apartado 2 por las desviaciones típicas respectivas).

z_x	z_y	z_z
-1'4142	-1'7483	0'9354
-0'7071	-0'2913	0'9354
0	0'6799	0'2672
0'7071	0'1942	-0'4009
1'4142	1'1655	-1'7372

Ahora utilizaremos la ecuación de regresión múltiple expresada en puntuaciones típicas (que obtuvimos en el apartado 7) para obtener las puntuaciones estimadas expresadas en típicas $z_{x'}$

$$z_{x'} = 0'4091\, z_y - 0'6314\, z_z$$

Puntuaciones estimadas:

$z_{x'_1}$ = 0'4091 (-1'7483) - 0'6314 (0'9354) = -1'3058

$z_{x'_2}$ = 0'4091 (-0'2913) - 0'6314 (0'9354) = -0'7097

$z_{x'_3}$ = 0'4091 (0'6799) - 0'6314 (0'2672) = 0'1094

$z_{x'_4}$ = 0'4091 (0'1942) - 0'6314 (-0'4009) = 0'3325

$z_{x'_5}$ = 0'4091 (1'1655) - 0'6314 (-1'7372) = 1'5736

$$\sum z_{x'} = 0 \qquad s^2_{Zx'} = 0'9805$$

Por tanto:

$$r_{x.yz} = 0'9805$$

La varianza de las puntuaciones típicas estimadas $z_{x'}$ no es igual a 1, por lo que a estas puntuaciones se las suele denominar *puntuaciones pseudo-típicas*. Esa varianza coincide con el coeficiente de determinación múltiple:

$$R^2_{x.yz} = \frac{s^2_{x'}}{s^2_x}$$

Pero expresadas las variables en típicas tenemos:

$$R^2_{x.yz} = \frac{s^2_{Zx'}}{s^2_{Zx}} = \frac{s^2_{Zx'}}{1} = s^2_{Zx'}$$

14. Correlación semiparcial $R_{(x.z)y}$ utilizando los residuales de la regresión de X sobre Z

Podemos calcular la correlación semiparcial $R_{(x.z)y}$, (es decir la correlación de la variable Y con la variable X una vez eliminados de esta última los 'efectos' de la variable Z,) como la correlación de Pearson entre los residuales de X (obtenidos de la regresión de X sobre Z) y la variable Y.

La ecuación de la recta de regresión de X sobre Z es:

X' = 11'4285 - 1'7857 Z

con ella podemos calcular los valores X' pronosticados o estimados, y a partir de estos los residuales o errores de estimación E = X - X'

X	Z	X'	E
1	5	2'5	-1'5
3	5	2'5	0'5
5	4	4'2857	0'7143
7	3	6'0714	0'9286
9	1	9'6428	-0'6428

Los residuales E representan la parte de X que Z no explica, o dicho de otro modo, la parte de X que puede considerarse independiente de Z.

Ahora, para calcular la correlación semi-parcial propuesta, procedemos a calcular la correlación de Pearson de esos residuales con Y

	E	Y	EY
	-1'5	2	-3
	0'5	5	2'5
	0'7143	7	5'0001
	0'9286	6	5'5716
	-0'6428	8	-5'1424
Suma V	0'0001	28	4'9293
Suma V^2	4'2857	178	97'738

$$r_{EY} = \frac{n\sum EY - \sum E \sum Y}{\sqrt{(n\sum E^2 - (\sum E)^2)(n\sum Y^2 - (\sum Y)^2)}}$$

$$r_{EY} = \frac{5 \cdot 4'9293 - 0'0001 \cdot 28}{\sqrt{(5 \cdot 4'2857 - 0)(5 \cdot 178 - 784)}} = 0'5170$$

Por tanto:

$$r_{(X.Z)Y} = 0'5170$$

15. Correlación parcial $R_{xy.z}$ utilizando los residuales de la regresión de X sobre Z y los residuales de la regresión de Y sobre Z

La correlación parcial $R_{xy.z}$ puede definirse y obtenerse como la correlación de Pearson entre los residuales E_x (obtenidos a través de la regresión de X sobre Z) y los residuales E_y (obtenidos a través de la regresión de Y sobre Z).

Los residuales E_x los hemos calculado en el apartado anterior, ahora necesitamos calcular los residuales E_y. Para obtenerlos calcularemos primero la ecuación de regresión de Y sobre Z.

Ecuación de regresión de Y sobre Z: $Y' = a + b\,Z$

En puntuaciones directas:

$$b = r_{yz}\,\frac{s_y}{s_z} = -0'7657\,\frac{2'0591}{1'4966} = -1'0534$$

$$a = \bar{Y} - b\,\bar{Z} = 5'6 + 1'0534 \cdot 3'6 = 9'3922$$

La ecuación por tanto es:

$Y' = 9'3922 - 1'0534\,Z$

Cálculo de los residuales $E_y = Y - Y'$

X	Y	Z	Y'	E_y
1	2	5	4'1252	-2'1252
3	5	5	4'1252	0'8748
5	7	4	5'1786	1'8214
7	6	3	6'232	-0'232
9	8	1	8'3388	-0'3388

Cálculo de la correlación: $r_{E_yE_x}$

	E_y	E_x	E_yE_x
	-2'1252	-1'5	3'1878
	0'8748	0'5	0'4374
	1'8214	0'7143	1'3010
	-0'232	0'9286	-0'2154
	-0'3388	-0'6428	0'2177
Suma V	0'0002	0'0001	4'9285
Suma V^2	8'7678	4'2857	12'1397

$$r_{Ex\,Ey} = \frac{n\sum E_x E_y - \sum E_x \sum E_y}{\sqrt{(n\sum E_x^2 - (\sum E_x)^2)(n\sum E_y^2 - (\sum E_y)^2)}}$$

$$r_{Ex\,Ey} = \frac{5 \cdot 4'9285 - 0'0002 \cdot 0'0001}{\sqrt{(5 \cdot 4'2857 - 0)(5 \cdot 8'7678 - 0)}} = 0'8040$$

Por tanto:

$$r_{xy\,.\,z} = 0'8040$$

Las pequeñas diferencias entre métodos de cálculo que aparecen en los últimos decimales se deben a la acumulación de errores de redondeo en problemas resueltos con calculadora u hoja de cálculo.