Análisis de datos en los estudios epidemiológicos III
Correlación y regresión

Julia García Salinero. Departamento de Investigación Fuden

Introducción

En el capitulo anterior estudiamos lo que se denomina estadística descriptiva univariada (una sola variable), sin embargo la mayor parte de investigaciones llevadas a cabo no solo pretenden describir fenómenos en base a la distribución de sus variables principales, sino que intentan encontrar relación entre algunas de las variables estudiadas.
A la parte de la estadística que se encarga de estudiar este tipo de relaciones entre variables se le denomina estadística descriptiva bivariada o multivariada.
En los siguientes capítulos vamos a detenernos en el estudio de algunas de las pruebas estadísticas que debemos utilizar cuando queremos encontrar relación o asociación entre las diferentes variables del estudio.
Ya indicamos en el capitulo anterior que los datos de una distribución de frecuencias se ordenaban en tablas de distribución. En el caso de la estadística bivariada también organizamos los datos en tablas, que en este caso se denominan tablas de contingencia. Una tabla de contingencia es una distribución con dos o más dimensiones (bidimensional), en la cual las frecuencias de dos o más variables se tabulan de manera cruzada. A pesar de que se pueden construir tablas de contingencia con varias variables y categorías las más frecuentemente utilizadas son las tablas de contingencia de 2x2 (dos filas por dos columnas), es decir dos variables que presentan dos categorías cada una de ellas.
Volveremos más tarde sobre este tema.

Los conceptos de Correlación y Regresión

Correlación

Se utiliza para obtener una medida del grado o la fuerza de la asociación entre dos variables cuantitativas.

El método más comúnmente utilizado para describir la relación entre dos variables es el coeficiente de correlación. Este tipo de relaciones puede ilustrarse de forma grafica, o bien, como sucede casi siempre, calcularse a través de la realización de una prueba que defina la magnitud de esa relación. La representación gráfica de una correlación entre dos variables se denomina **grafica** o **diagrama de dispersión**, que no será estudiado en este capitulo, ya que no suele ser muy utilizado.
En la figura 1 observamos un ejemplo de un diagrama de dispersión que expresa la relación existente entre la dosis de un determinado fármaco hipotensor y los valores de la presión sanguínea.

A pesar de la utilización de diagramas, el índice más frecuentemente empleado, para determinar la intensidad de la relación entre dos variables X e Y, es el **coeficiente de correlación de Pearson**. Este coeficiente se calcula cuando las variables de estudio fueron medidas en escala de intervalos o de proporción. Cuando las variables fueron medidas en escala ordinal, se suele utilizar el coeficiente de correlación de **rho de Sperman.** Sus valores oscilan entre -1 y +1. Un valor de Pearson igual a 0 indica la ausencia de relación, es decir que las dos variables son independientes. Valores grandes de dicho coeficiente (r), ya sean positivos o negativos, indican una fuerte relación entre las dos variables. Un valor de r positivo indica que valores grandes de la variable X se asocian con valores grandes de la variable Y; y los valores bajos de la variable X se asocian con valores bajos de la variable Y. Por su parte, un valor de r negativo indica que los valores grandes de la variable X se asocian con valores bajos de la variable Y, y que valores bajos de la variable X se asocian con valores altos de la variable Y.
En la figura 2 observamos diferentes tipos de correlación, aunque nosotros solo nos detendremos en el análisis de la correlación lineal.

El coeficiente de correlación de Pearson depende fundamentalmente de:

- La variabilidad del grupo. A mayor variabilidad de la población , el r tiene mayor fuerza;
- El influjo de una tercera variable que pudiera enmascarar los resultados obtenidos.

Las correlaciones perfectas son muy poco frecuentes en investigación y resulta difícil indicar qué valor se considera razonable para determinar la magnitud de una correlación, dependiendo fundamentalmente del tipo de variable en estudio. Por ejemplo si intentamos buscar correlación entre la medida de la glucemia basal con diferentes métodos, una correlación entre las diferentes medidas de 0,70 puede considerarse baja; sin embargo este mismo valor para variables de tipo social o psicológico indicaría

una correlación muy alta. En realidad el único criterio que podemos seguir es remitirnos a estudios anteriores.
Por otra parte el coeficiente de correlación de Pearson indica únicamente que dos variables independientes, varían conjuntamente, pero hay que dejar claro **que esta variación conjunta no indica necesariamente que exista causalidad entre ambas**

A pesar de que su calculo es laborioso y que como otras pruebas estadísticas (descriptivas o inferenciales) rara vez se realizan en la actualidad de forma manual, sino a través de paquetes estadísticos como el SPSS, vamos a intentar explicar su cálculo a través de un ejemplo
Imaginemos que estamos realizando un estudio para encontrar la relación entre el consumo de un determinado fármaco y la presión sanguínea en un grupo de cinco individuos. A la variable consumo del fármaco la denominaremos X y la categorizamos en diferentes niveles en función de la dosis en mg. A la variable presión sanguínea la denominaremos Y, y la expresaremos en mm de Hg.
Las dosis de fármacos serían: 1, 2, 3, 4 y 5 mg. Los valores de presión arterial serían: 278, 260, 198,160 y 154 mm Hg. Como podemos observar estos datos no tienen sentido en la forma en que están presentados, por lo que procederemos a organizarlos (tabla 1)

Para estudiar la existencia o no de relación entre estos dos factores o variables de estudio se calculará el coeficiente de correlación de Pearson (Figura 3). La correlación nos permitirá medir el grado o la fuerza de relación entre estas dos variables.

Para facilitar la aplicación de la formula organizamos pues los datos de la tabla anterior (tabla 2).

En nuestro ejemplo hemos obtenido un coeficiente de correlación negativo, lo cual interpretamos como que ambas variables no son independientes, es decir que a mayor valor de la variable X menor valor obtenemos en la variable Y. Cuanto mayor es la dosis del fármaco menor es el valor de la Presión sanguínea.

Regresión

Consiste en obtener una ecuación que se pueda usar para predecir el valor de una variable, teniendo en cuenta un valor asignado a otra variable.

Regresión lineal simple: Una de las condiciones que deben cumplirse para calcular el coeficiente de correlación de Pearson es que los puntos del diagrama de dispersión tiendan a la linealidad.
Como indicamos anteriormente utilizamos la regresión para obtener una ecuación que nos permita predecir los valores de una variable en función de los datos observados en la otra. Por lo tanto, la ecuación de regresión será la ecuación de la recta que mejor represente a todos los puntos del diagrama y que nos permitirá pronosticar el valor de una variable en función de otra con la que esta relacionada (regresión). La ecuación de la recta viene dada por la formula siguiente:

$$Y = a+bx$$

Donde **"b"** es la pendiente de la recta y mide algo así como la velocidad de ascenso; y **"a"** es el punto donde la recta corta al eje Y, denominándose ordenada en el origen.
En esta ecuación los valores de la variable Y dependerán de los valores de la variable X. La fórmula para el cálculo de "a" y "b" se recoge en el gráfico 4
Como sabemos, una recta queda definida por dos puntos, por tanto, si asignamos dos valores cualesquiera a la variable independiente X obtenemos los valores correspondientes de la variable Y.
Así para $x = 1$ $y = 3+2.1= 5$ $y = 5$
$X = 3$ $y = 3+ 2.3 = 9$ $y = 9$

En la figura 5 se muestran varios ejemplos de rectas de regresión.

A pesar de que como indicamos para el cálculo del coeficiente de correlación, actualmente se utilizan paquetes estadísticos, vamos a tratar de explicar el procedimiento del cálculo de la regresión utilizando el mismo ejemplo de la dosis de un fármaco hipotensor y los valores de presión sanguínea (tabla 3).

Si realizásemos un grafico con su diagrama de dispersión observaríamos que los puntos representados tienden a la linealidad (condición para cálculo del coeficiente de correlación). Sin embargo somos conscientes de que es imposible encontrar una línea recta que pase por todos los puntos de forma simultánea. La solución pasaría por encontrar una línea recta que se aproxime lo más posible a estos puntos.
Así pues para cada valor de la variable X tenemos dos valores de la variable Y. Por un lado, el valor Y obtenido, y por otro el valor y´ calculado mediante la ecuación de regresión $y = a+ bx$.
La diferencia entre la puntuación Y obtenida y el valor de y´ calculada se denomina **error de preedición.** Así pues podemos definir la ecuación de regresión como la ecuación de la recta que hace mínimos los errores de predicción, tal como observamos en la figura 6.

Seguirnos con nuestro ejemplo y procedemos a organizar la tabla 4, que nos permitirá aplicar la formula con mayor facilidad. De acuerdo a las fórmulas recogidas en la figura 4, los valores de a y b son 314,4 y -34,8 respectivamente. Por tanto nuestra ecuación de regresión seria Y = 314,4 - 34,8X

Aplicando la ecuación de regresión, podemos predecir la presión sanguínea de cualquier paciente sometido a un tratamiento con una determinada dosis de fármaco. No debemos olvidar que nuestra predicción no es exacta, puesto que observamos que cuando la dosis era de 1 mg (X= 1) el valor de la presión era de 278 mm en HG (Y= 278) para el primer paciente. Sin embrago en nuestra predicción el valor era de 279,6, la mejor que podíamos hacer.
Así pues nuestro error de preedición será la diferencia entre el valor real obtenido Y = 278 y el valor de preedición Y = 279,6. Observémoslo en la figura 7.

Existen también otros procedimientos estadísticos multivariados de regresión más complejos como son la regresión múltiple y la regresión logística que no serán abordados en este capitulo, debido a la complejidad de su cálculo.

Bibliografía

Carrasco JL. El método estadístico en la investigación médica. Editorial Ciencia 3. 6ª Edición. 1995

Rodríguez Miñon P. Estadística Aplicada a la Biología. Editorial UNED. 3º Edición. 1984.

Polit Denise y Hungler Bernadette. Investigación científica en ciencias de la salud. Editorial McGraw-Hill Interamericana. 6ª edición. 2000.

Tablas y gráficos

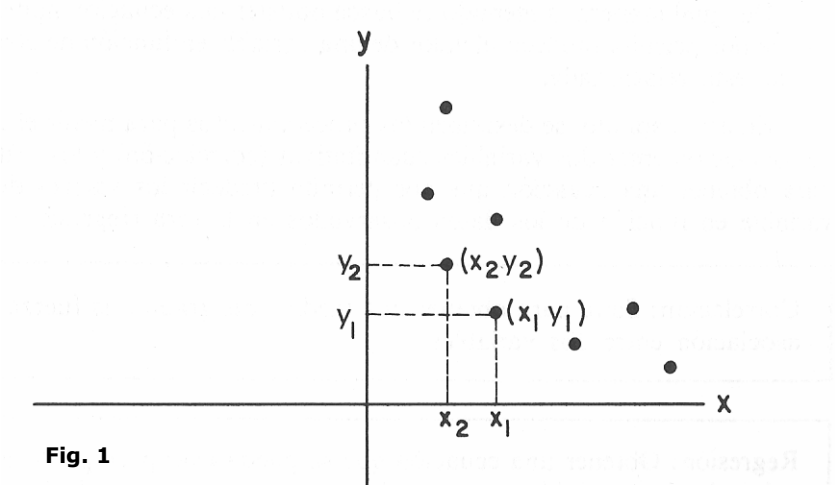


Fig. 1

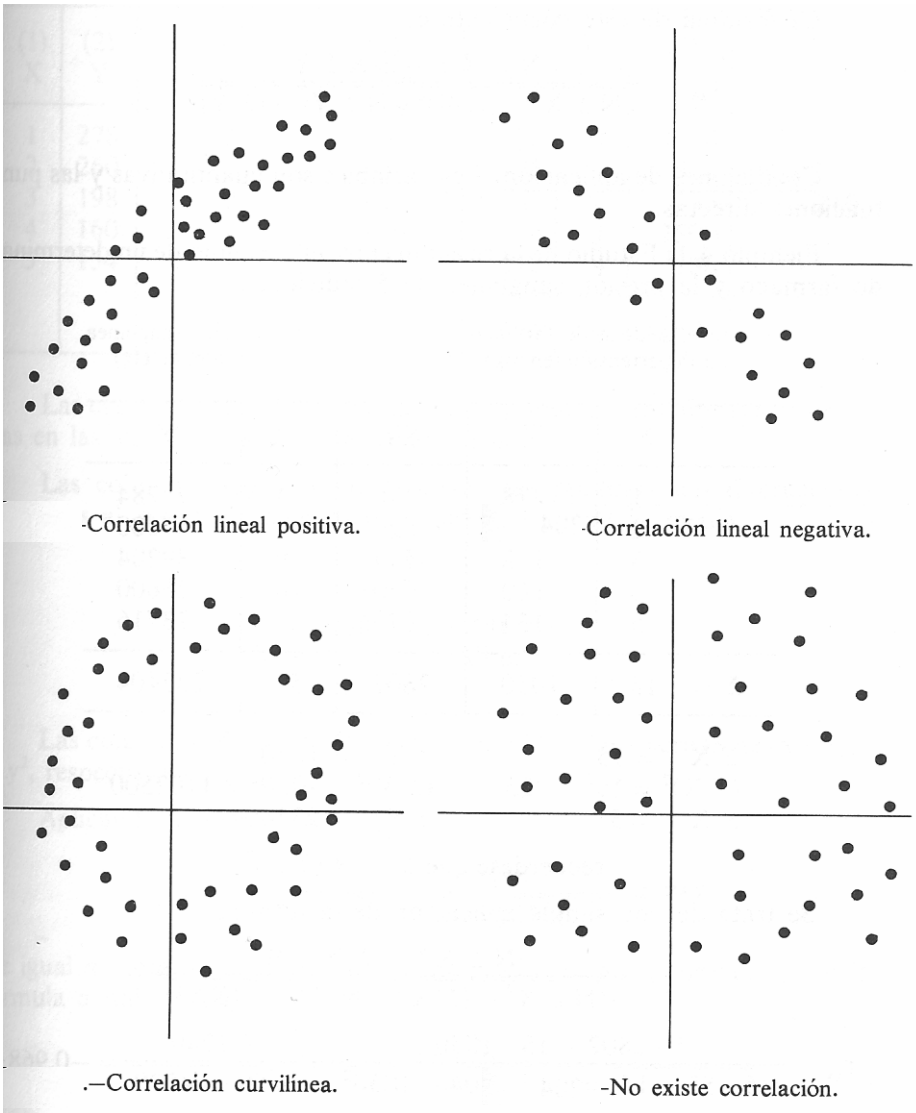


Fig. 2

X	Y
1	278
2	260
3	198
4	160
5	154

Tabla 1

$$r_{xy} = \frac{N\sum XY - \sum x \sum Y}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}\sqrt{N\sum Y^2 - \left(\sum Y\right)^2}} \qquad \frac{5.2802 - 15.1050}{\sqrt{5.55 - 15^2}\sqrt{5.233404 - 1050^2}} = -0{,}968$$

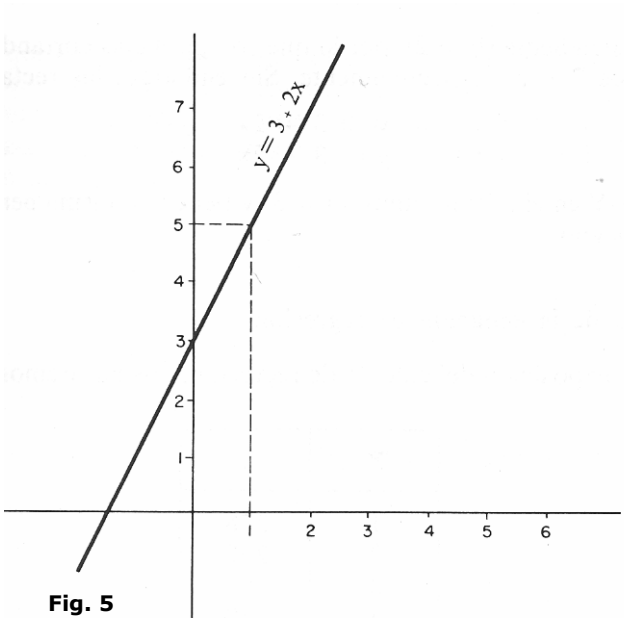
Fig. 3

X	**Y**	**XY**	$\mathbf{x}^2$	Y^2
1	278	278	1	**77284**
2	260	520	4	**67600**
3	198	594	9	**39204**
4	160	640	16	**25600**
5	154	770	25	**23716**
$\sum = 15$	**1050**	**2802**	**55**	**233404**

Tabla 2

$$a = \bar{Y} - b\bar{X} \qquad b = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - \left(\sum X\right)^2}$$

Fig. 4



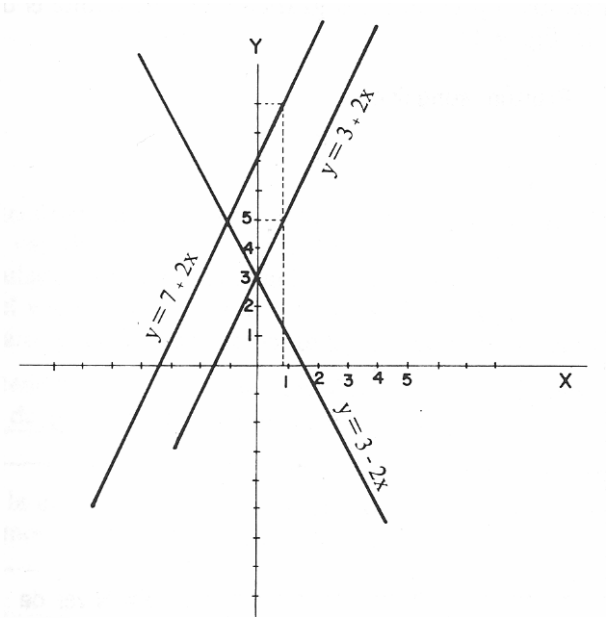


Fig. 5

X (dosis de fármaco)	Y (Valores de PA)
1	278
2	260
3	198
4	160
5	154

Tabla 3

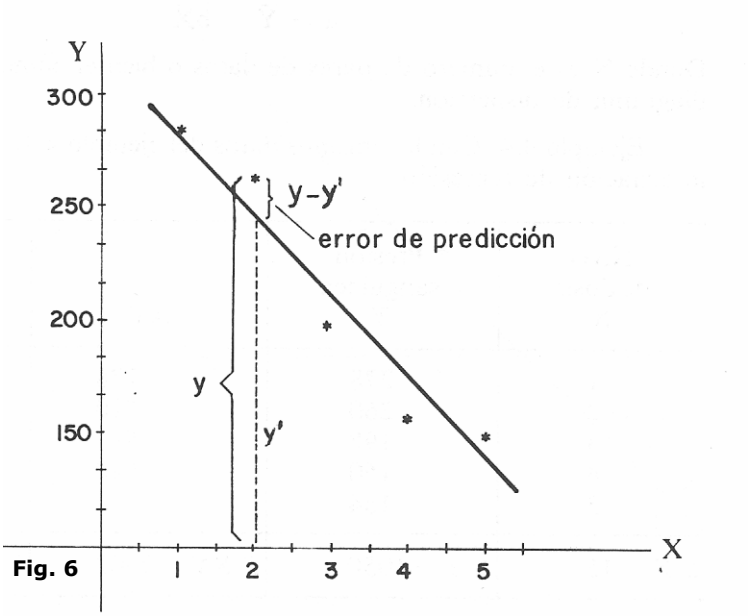


Fig. 6

Nivel de dosis X	Presión Sanguínea	XY	X^2
1	278	278	1
2	260	520	4
3	198	594	9
4	160	640	16
5	154	770	25
$\sum = 15$ **Media de X = 3**	**1050** **Media de Y = 210**	**2802**	**55**

Tabla 4

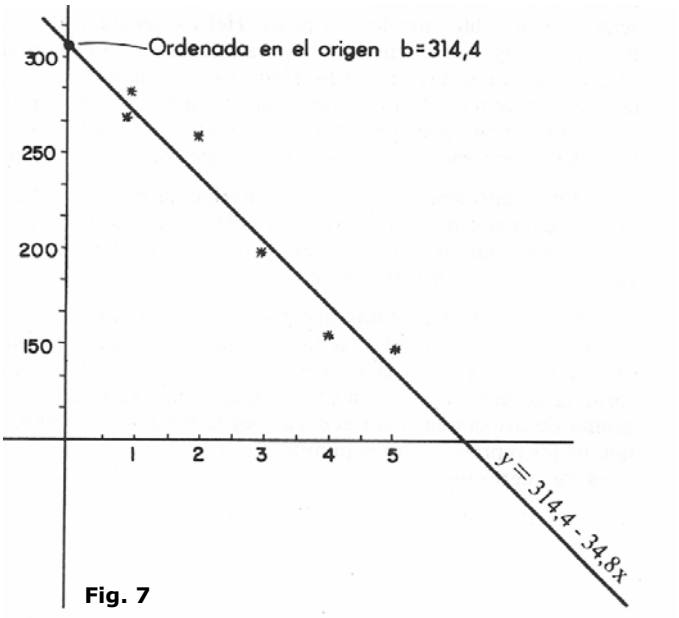


Fig. 7