

INT-0310

c.1

INSTITUTO LATINOAMERICANO DE  
PLANIFICACION ECONOMICA Y SOCIAL  
Santiago, 27 de mayo de 1963

✓  
80/63

RESUMEN SOBRE REGRESION LINEAL \*

\* Preparado por el Profesor Juan Ayza para el Programa de Capacitación del Instituto, Curso Básico de Planificación.



## RESUMEN SOBRE REGRESION LINEAL

### Nomenclatura

Diferenciamos las propiedades de la "población" original, de las muestrales, y éstas de las deducciones que obtengamos.

Las propiedades de la población original las expresamos con letras griegas. Así una relación lineal entre dos variables, en la población original, la expresaremos

$$Y = \alpha + \beta X$$

Las muestras vendrán en pares de observaciones que indicamos con minúsculas. Para diferenciar una observación de otra, cuando se requiere pondremos un subíndice que indique el número de la observación, así el par de valores observados en las variables Y y X, en la observación i, será  $y_i, x_i$ . A veces no será preciso diferenciar una observación de otra (en ciertas sumas) y entonces prescindiremos del subíndice.

Si de las diversas observaciones deducimos una recta, más aceptable, a los datos de la muestra de la "población original". La relación deducida será

$$Y = a + bX$$

La recta deducida, será "más aceptable" en relación con el estudio estadístico de la muestra, y las hipótesis que se hagan respecto a la población original.

### Ecuaciones normales

Son las relaciones de carácter estadístico, que nos permiten relacionar las observaciones con los parámetros  $a$  y  $b$ , a determinar.

---

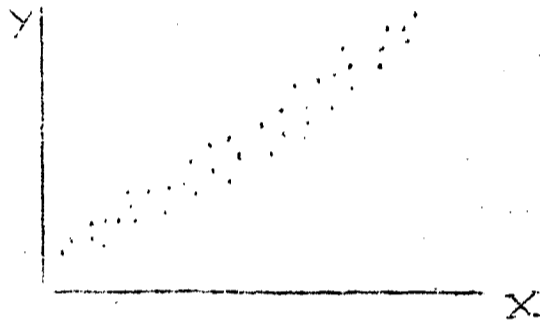
Nota: Para ampliación de conceptos y aplicaciones prácticas debe revisarse el folleto "Análisis de Correlación" por Peuro Vusković (1962).

/Podemos basarnos

Podemos basarnos en el método de los momentos o en el de los mínimos cuadrados. En ambos casos obtendremos las mismas ecuaciones normales.

La representación gráfica de la nube de puntos, donde cada punto corresponde a un par observado  $x_i, y_i$ , puede ser la siguiente:

Gráfico 1



El método de los momentos, de más fácil comprensión para ingenieros considera a cada punto como extremo de una fuerza paralela al eje Y y de módulo  $y_i$ . El problema consiste en encontrar una distribución uniforme (a lo largo de una recta) de otras fuerzas, que tengan la misma proyección y el mismo momento; condición de equilibrio.

Es decir

$$\sum Y = \sum y$$

$$\sum YX = \sum yx$$

de donde se deducen las dos ecuaciones normales:

$$\begin{aligned} \sum y &= Na + b \sum x \\ \sum yx &= a \sum x + b \sum x^2 \end{aligned}$$

/El método

El método de los mínimos cuadrados se basa en aceptar un tipo de distribución de las desviaciones, con la cual debe cumplirse

$$\sum (y - Y)^2 = \text{mínimo}$$

o sea

$$\sum (y - a - bx)^2 = \text{mínimo}$$

Por tanto

$$\frac{\partial \sum (y - a - bx)^2}{\partial a} = 0 = -2 \sum (y - a - bx)$$

$$\frac{\partial \sum (y - a - bx)^2}{\partial b} = 0 = -2 \sum (y - a - bx)x$$

de donde se obtienen las dos ecuaciones normales

$\sum y = Na + b \sum x$
$\sum yx = a \sum x + b \sum x^2$

que cumplen con otras condiciones del mínimo.

De estas ecuaciones por determinantes, indicaremos el valor de a y b.

$$a = \frac{\begin{vmatrix} \sum y & \sum x \\ \sum yx & \sum x^2 \end{vmatrix}}{\begin{vmatrix} N & \sum x \\ \sum x & \sum x^2 \end{vmatrix}}$$

$$b = \frac{\begin{vmatrix} N & \sum y \\ \sum x & \sum yx \end{vmatrix}}{\begin{vmatrix} N & \sum x \\ \sum x & \sum x^2 \end{vmatrix}}$$

b puede

b puede calcularse de esta manera, en cuyo caso a se obtiene más fácilmente de la primera ecuación normal

$$a = \frac{1}{N} \left( \sum y - b \sum x \right)$$

o también  $a = \bar{y} - b\bar{x}$

Donde representamos los promedios aritméticos con una barra.

$$\bar{x} = \frac{1}{N} \sum x \qquad \bar{y} = \frac{1}{N} \sum y$$

Definimos otros dos símbolos  $s_x$  y  $s_{xy}$ . Al primero lo llamamos desviación standard.

$$s_x = \left[ \frac{1}{N} \sum (x - \bar{x})^2 \right]^{1/2}$$

$$s_{xy} = \frac{1}{N} \sum (x - \bar{x})(y - \bar{y})$$

$$Ns_x^2 = \sum (x - \bar{x})^2$$

$$Ns_{xy} = \sum (x - \bar{x})(y - \bar{y})$$

a estas últimas expresiones las llamamos varianza y covarianza, respectivamente.

A continuación deduciremos expresiones de mayor simplicidad en el cálculo de varianza y covarianza, y mostraremos después su relación con los parámetros de regresión a y b.

$$/Ns_x^2 =$$

$$Ns_x^2 = \sum (x^2 - 2x\bar{x} + \bar{x}^2)$$

$$= \sum x^2 - 2\bar{x} \sum x + N\bar{x}^2$$

$$= \sum x^2 - 2N\bar{x}^2 + N\bar{x}^2$$

$$Ns_x^2 = \sum x^2 - N\bar{x}^2$$

$$Ns_x^2 = \sum x^2 - \frac{(\sum x)^2}{N}$$

$$N^2 s_x^2 = N \sum x^2 - (\sum x)^2$$

$$N^2 s_x^2 = \begin{vmatrix} N & \sum x \\ \sum x & \sum x^2 \end{vmatrix}$$

$$Ns_{xy} = \sum (xy - x\bar{y} - \bar{x}y + \bar{x}\bar{y})$$

$$= \sum xy - \bar{y} \sum x - \bar{x} \sum y + N\bar{x}\bar{y}$$

$$= \sum xy - N\bar{y}\bar{x} - N\bar{x}\bar{y} + N\bar{x}\bar{y}$$

$$= \sum xy - N\bar{x}\bar{y}$$

$$Ns_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{N}$$

$$N^2 s_{xy} = N \sum xy - (\sum x)(\sum y)$$

$$N^2 s_{xy} = \begin{vmatrix} N & \sum y \\ \sum x & \sum xy \end{vmatrix}$$

Si comparamos estos determinantes con los de b, deducimos

$$b = \frac{N^2 s_{xy}}{N^2 s_x^2}$$

$$b = \frac{s_{xy}}{s_x^2}$$

Si reemplazamos a por su expresión en términos de promedios, en la ecuación de la recta, obtendremos

$$/Y =$$

$$Y = a + bX$$

$$a = \bar{y} - b\bar{x}$$

$$Y - \bar{y} = b(X - \bar{x})$$

ecuación de la recta que pasa por el punto  $\bar{x}, \bar{y}$ . Es decir la línea de regresión pasa por  $\bar{x}, \bar{y}$ .

La misma ecuación también la podemos expresar:

$$Y - \bar{y} = \frac{s_{xy}}{s_x^2} (X - \bar{x})$$

#### ¿Dos líneas de regresión?

Todo lo anteriormente deducido partía de ajustar la ecuación

$$Y = a + bX$$

donde  $Y = f_1(X)$ , se considera la función directa. Podríamos despejar X, obteniendo así la llamada función inversa de Y, o sea  $X = f_2(Y)$ , que sería otra relación lineal, por ejemplo

$$X = a' + b'Y$$

esta función puede obtenerse matemáticamente de la precedente, operación que no introduce novedad alguna. Pero otra solución alternativa sería plantear el problema nuevamente. Es decir, con los mismos pares de observaciones  $x_i, y_i$ , se desea encontrar el valor de los parámetros que nos dé la relación lineal

$$X = a' + b'Y$$

/Aplicando los



Aplicando los mismos razonamientos que en el caso anterior mutatis mutandis, se llega a fórmulas similares. La principal conclusión es que, en general, la recta así obtenida no es la misma que la obtenida para la primera ecuación, o lo que es lo mismo, la deducida de esta despejando X. Veamos primero las fórmulas a que se llega, que pondremos por simple analogía, con la diferencia de que definimos una desviación standard  $s_y$ .

ecuaciones normales

$$\begin{aligned} \sum x &= Na' + b' \sum y \\ \sum xy &= a' \sum y + b' \sum y^2 \end{aligned}$$

$$a' = \frac{\begin{vmatrix} \sum x & \sum y \\ \sum xy & \sum y^2 \end{vmatrix}}{\begin{vmatrix} N & \sum y \\ \sum y & \sum y^2 \end{vmatrix}}$$

$$b' = \frac{\begin{vmatrix} N & \sum x \\ \sum y & \sum xy \end{vmatrix}}{\begin{vmatrix} N & \sum y \\ \sum y & \sum y^2 \end{vmatrix}}$$

$$a' = \frac{1}{N} (\sum x - b' \sum y)$$

$$a' = \bar{x} - b' \bar{y}$$

$$Ns_y^2 = \sum (y - \bar{y})^2$$

$$Ns_y^2 = \sum y^2 - \frac{(\sum y)^2}{N}$$

$$N^2 s_y^2 = \begin{vmatrix} N & \sum y \\ \sum y & \sum y^2 \end{vmatrix}$$

$$/b' =$$

$$b' = \frac{N^2 s_{xy}}{N^2 s_y^2}$$

$$b' = \frac{s_{xy}}{s_y^2}$$

$$X = a' + b'Y$$

$$a' = \bar{x} - b'\bar{y}$$

$$X - \bar{x} = b'(Y - \bar{y})$$

o sea, que esta recta pasa también por el punto promedio  $\bar{x}, \bar{y}$ .

Pero esta recta no coincide con la primera, salvo casos especiales, como veremos a continuación.

Para que coincidan las rectas

$$X = a' + b' Y$$

$$Y = a + bX$$

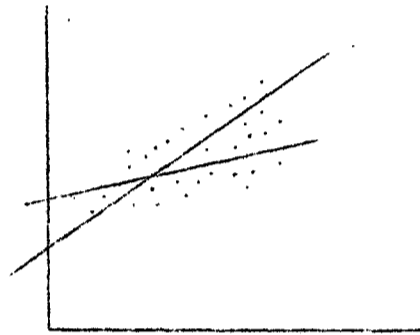
se requiere: 1)  $b' = \frac{1}{b}$       2)  $a' = -\frac{a}{b}$

un examen rápido de las expresiones para  $b$  y  $b'$ , en determinantes nos indica que sólo en condiciones especiales coincidirán las pendientes de ambas rectas. Algo similar sucede si examinamos las expresiones de  $a$  y  $a'$ . Las condiciones son justamente las que resulten de igualar las expresiones anteriores.

Nos encontramos ante el hecho de que procediendo en manera similar tendremos, en el caso general, dos rectas de regresión distintas que gráficamente podemos representar

/Gráfico 2

Gráfico 2



A la que representa

$y = a + bX$  se le llama regresión de Y sobre X.

Grado de asociación entre las variables

Consideremos nuevamente el gráfico 2. Ambas rectas pasan por el mismo punto  $\bar{x}$ ,  $\bar{y}$ . Si cambiamos el origen de coordenadas a ese punto, tendremos expresiones más simples para ambas rectas. Estas expresiones son justamente las ecuaciones que obtuvimos ya:

$$(Y - \bar{y}) = b (X - \bar{x})$$

$$(X - \bar{x}) = b' (Y - \bar{y})$$

donde cada una de las dos diferencias sería la nueva variable, referida al eje de coordenadas con origen  $\bar{x}$ ,  $\bar{y}$ . Pero estas rectas no tienen por qué ser funciones inversas. Es decir, no son simétricas respecto al eje de  $45^\circ$ . Las funciones simétricas deben tener esta propiedad. Sin embargo, si variamos las escalas convenientemente podemos lograr esa simetría. Veamos primero cuales son las condiciones de simetría. Si tenemos dos variables Z y W, relacionadas linealmente  $Z = rW$  r mide la pendiente de la recta, que pasa por el origen, respecto al eje W.

La función inversa será  $Z = r'W$  y entre ambas pendientes debe cumplirse la condición

$$rr' = 1$$

o sea que la función inversa puede expresarse también así:

$$W = r' Z.$$

Volvamos a nuestras ecuaciones referidas a los nuevos ejes y reemplacemos los parámetros  $b$  por su expresión en términos de  $s_{xy}$ ,  $s_x$  y  $s_y$ , buscando lograr la simetría señalada.

$$(Y - \bar{y}) = b (X - \bar{x})$$

$$(X - \bar{x}) = b' (Y - \bar{y})$$

$$(Y - \bar{y}) = \frac{s_{xy}}{s_x} (X - \bar{x})$$

$$(X - \bar{x}) = \frac{s_{xy}}{s_y} (Y - \bar{y})$$

/ (Y -  $\bar{y}$ )

$$(Y - \bar{y}) = \frac{s_{xy}}{s_x} \cdot \frac{(X - \bar{x})}{s_x} \qquad (X - \bar{x}) = \frac{s_{xy}}{s_y} \cdot \frac{(Y - \bar{y})}{s_y}$$

Si dividimos las ecuaciones entre  $s_y$  y  $s_x$ , respectivamente, habremos conseguido la simetría buscada.

$$\frac{(Y - \bar{y})}{s_y} = \frac{s_{xy}}{s_x s_y} \cdot \frac{(X - \bar{x})}{s_x} \qquad \frac{(X - \bar{x})}{s_x} = \frac{s_{xy}}{s_x s_y} \cdot \frac{(Y - \bar{y})}{s_y}$$

donde  $r = \frac{s_{xy}}{s_x s_y}$        $Z = \frac{(Y - \bar{y})}{s_y}$        $W = \frac{(X - \bar{x})}{s_x}$

En resumen, mediante un cambio en el origen de coordenadas, y una modificación de las escalas hemos simplificado las fórmulas de las dos líneas de regresión a

$$Z = rW$$

y

$$W = rZ$$

que son dos rectas que pasan por el nuevo origen, y una de ellas es función inversa de la otra.

La pendiente  $r$  de estas rectas modificadas, tiene mucha importancia y se llama coeficiente de correlación.

Si el grado de asociación lineal de las variables  $X$  e  $Y$  fuera excelente, las dos líneas de regresión tenderían a coincidir, y en consecuencia  $r$  tendería a 1 ó a -1. Si el grado de asociación fuera mínimo, en el peor de los casos las rectas de regresión serían normales entre sí y  $r$  tendería a cero.

Es fácil demostrar

$$r = \sqrt{bb'} \text{ y toma el mismo signo de } s_{xy}$$

También se demuestra que el valor de  $r$  no cambia ni con el origen de coordenadas ni con las unidades de medida. Es decir, es invariante a estas transformaciones.

Variación alrededor de la línea de regresión

Si una vez ajustada la recta

$$Y = a + bX$$

estudiamos los residuos (distancias verticales de cada punto a la recta)

$$|d_i| = |y_i - Y| = |Y - y_i|$$

y formamos  $\sum d^2$ , en forma análoga a las anteriores podemos definir una desviación correspondiente  $s_{ey}$

$$Ns_{ey}^2 = \sum d^2$$

$$Ns_{ey}^2 = \sum (y - Y)^2$$

$$= \sum [y - \bar{y} - b(x - \bar{x})]^2$$

$$= \sum (y - \bar{y})^2 - 2b \sum (y - \bar{y})(x - \bar{x}) + b^2 \sum (x - \bar{x})^2$$

$$Ns_{ey}^2 = Ns_y^2 - 2bNs_{xy} + b^2 s_x^2$$

$$s_{ey}^2 = s_y^2 - 2bs_{xy} + b^2 s_x^2$$

/Pero de

Pero de  $b = \frac{s_{xy}}{s_x^2}$

se deduce  $b^2 s_x^2 = b s_{xy}$

y  $b s_{xy} = r^2 s_y^2$

entonces  $s_{ey}^2 = s_y^2 - 2r^2 s_y^2 + r^2 s_y^2$

$$s_{ey}^2 = (1 - r^2) s_y^2$$

Con objeto de relacionarla a la anterior desviación  $s_{ey}$ , estudiaremos ahora la varianza de la línea de regresión, definiendo su desviación correspondiente  $s_y$

$$Ns_Y^2 = \sum (Y - \bar{Y})^2 = \sum (Y - \bar{y})^2$$

pues la línea pasa por  $\bar{y}$ , como vimos. Además  $Y - \bar{y} = b(x - \bar{x})$

$$Ns_Y^2 = b^2 \sum (X - \bar{x})^2$$

$$= b^2 Ns_x^2$$

$$s_Y^2 = b^2 s_x^2$$

y  $s_Y^2 = r^2 s_y^2$

Esto indica también la importancia de  $r$ , al representar una relación entre la varianza explicada por  $Y$ , y la varianza total de las observaciones  $y$ .

/Por consiguiente

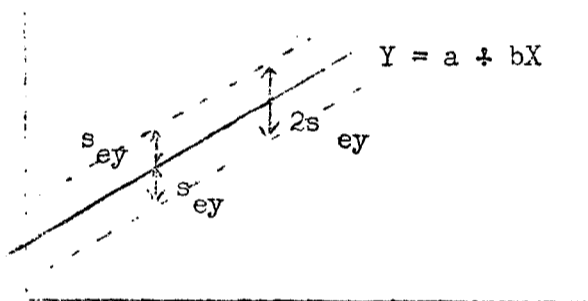
Por consiguiente

$$s_{ey}^2 = s_y^2 - s_Y^2$$

lo que explica el nombre de varianza residual o no explicada a  $\sum d^2$ .

Si la distribución es normal dentro de  $2s_{ey}$ , alrededor de la línea de regresión debe encontrarse aproximadamente 2/3 de los puntos.

Gráfico 3



Formatos para



Formatos para el cálculo práctico de la línea de regresión y otros indicadores de importancia

Interesa fundamentalmente calcular  $b, a, r$  y  $s_{ey}$ , en este orden

$$Y = a + bX$$

	x	y	x <sup>2</sup>	xy	y <sup>2</sup>
(N observaciones)					
$\sum$	$\sum x$	$\sum y$	$\sum x^2$	$\sum xy$	$\sum y^2$
$\sum / N$	$\bar{x}$	$\bar{y}$	$\frac{1}{N} \sum x^2$	$\frac{1}{N} \sum xy$	$\frac{1}{N} \sum y^2$

Calculamos en primer lugar  $s_x^2, s_{xy}, s_y^2$

$$s_x^2 = \frac{1}{N^2} \begin{vmatrix} N & \sum x \\ \sum x & \sum x^2 \end{vmatrix} = \begin{vmatrix} 1 & \bar{x} \\ \bar{x} & \frac{1}{N} \sum x^2 \end{vmatrix} \quad \boxed{s_x^2}$$

$$s_{xy} = \frac{1}{N^2} \begin{vmatrix} N & \sum y \\ \sum x & \sum xy \end{vmatrix} = \begin{vmatrix} 1 & \bar{y} \\ \bar{x} & \frac{1}{N} \sum xy \end{vmatrix} \quad \boxed{s_{xy}}$$

$$/s_y^2$$

$$s_y^2 = \frac{1}{N^2} \left| \begin{array}{c|c} N & \bar{Y} \\ \hline \sum Y & \sum Y^2 \end{array} \right| = \left| \begin{array}{c|c} 1 & \bar{y} \\ \hline \bar{y} & \frac{1}{N} \sum Y^2 \end{array} \right|$$

$$s_y^2$$

Con los datos anteriores calculamos  $b$  y  $a$

$$b = \frac{s_{xy}}{s_x^2}$$

$$b$$

$$a = \bar{y} - b\bar{x}$$

$$a$$

Además

$$b' = \frac{s_{xy}}{s_y^2}$$

de donde

$$r = \sqrt{bb'}$$

$$r$$

con el signo de  $s_{xy}$

y

$$r^2 = bb'$$

$$r^2$$

Calculamos ahora

$$s_{ey}^2 = (1 - r^2) s_y^2$$

$$s_{ey}^2$$

de donde deducimos la desviación residual. Con distribución normal, dentro de  $2s_{ey}$  alrededor de la línea  $Y = a + bx$  deben caer 2/3 de los puntos, aproximadamente.

$$s_{ey}$$