Universidad Autónoma de Madrid
Grado Administración y Gestión
Facultad Ciencias Económicas y Empresariales
Departamento de Economía Aplicada
Profesor: Santiago de la Fuente Fernández

ESTADÍSTICA BIDIMENSIONAL

VARIABLE ESTADÍSTICA BIDIMENSIONAL

Cuando se consideran situaciones en la que el estadístico realiza la observación simultanea de dos caracteres en el individuo, se obtienen pares de resultados.
Los distintos valores de las modalidades que pueden adoptar estos caracteres forman un conjunto de pares, que representamos por (X, Y), y llamaremos *variable estadística bidimensional*.
Los dos caracteres observados no tienen por qué ser de la misma clase, pudiendo presentarse distintas situaciones:

- **Dos caracteres cualitativos:** El sexo y color del pelo de una persona.
- **Dos caracteres cuantitativos:** El peso y la estatura de una persona.
- **Uno cuantitativo y otro cualitativo:** La profesión y los años de servicio.

Las variables (X, Y) que representan los valores de dos caracteres cuantitativos, pueden clasificarse:

- **X discreta e Y discreta:** Número de hijos y número de hermanos de una persona.
- **X continua e Y continua:** Perímetro craneal y perímetro torácico de una persona.
- **X discreta e Y continua:** Hijos de una familia y estatura del padre.
- **X continua e Y discreta:** Temperatura y pulsaciones.

ORDENACIÓN DE LOS DATOS: TABLA DE DOBLE ENTRADA

El par (X, Y) es la unidad del estudio y dos pares serán repetidos solo cuando sus respectivas componentes sean iguales. De otra parte, el número de modalidades que adopta el carácter X no tiene por qué ser el mismo que el que adopta el carácter Y:

$$X = (x_1, x_2, \cdots, x_k) \qquad Y = (y_1, y_2, \cdots, y_m)$$

Para ordenar los datos se utiliza una tabla de doble entrada donde tengan cabida los k valores distintos de la variable X y los m valores distintos de la variable Y. En la tabla se puede expresar el número de veces que se repite cada para de valores posibles (x_i, y_j) formado en el producto cartesiano de los dos conjuntos numéricos.

TABLA DE DOBLE ENTRADA

Y / X	y_1	y_2	…	y_j	…	y_m
x_1	n_{11}	n_{12}	…	⋮	…	n_{1m}
x_2	n_{21}	n_{22}	…	⋮	…	n_{2m}
…	…	…	…	⋮	…	…
x_i	……	……	……	… n_{ij}	…	n_{im}
…	…	…	…		…	…
x_k	n_{k1}	n_{k2}	…		…	n_{km}

$N \equiv$ número total observaciones

$n_{ij} \equiv$ frecuencia absoluta, número de veces que aparece repetido el par (x_i, y_j).

La frecuencia relativa del par se define: $f_{ij} = \dfrac{n_{ij}}{N}$

DISTRIBUCIONES MARGINALES

X \ Y	y_1	y_2	…	y_j	…	y_m	$n_{i\bullet}$
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1m}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2m}	$n_{2\bullet}$
…	…	…	…	…	…	…	…
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{im}	$n_{i\bullet}$
…	…	…	…	…	…	…	…
x_k	n_{k1}	n_{k2}	…	n_{kj}	…	n_{km}	$n_{k\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	…	$n_{\bullet j}$	…	$n_{\bullet m}$	$N=\sum_{i=1}^{k} n_{i\bullet}=\sum_{j=1}^{m} n_{\bullet j}$

$$a_{11}=\frac{\sum_{i=1}^{k}\sum_{j=1}^{m} x_i . y_j . n_{ij}}{N} \qquad s_{yx} = a_{11} - a_{10}.a_{01} = a_{11} - \bar{x}.\bar{y} \text{ (covarianza)}$$

- **DISTRIBUCIÓN MARGINAL DE LA VARIABLE X**

X	x_1	x_2	…	x_i	…	x_k
$n_{i\bullet}$	$n_{1\bullet}$	$n_{2\bullet}$	…	$n_{i\bullet}$	…	$n_{k\bullet}$
$f_{i\bullet}=\frac{n_{i\bullet}}{N}$	$f_{1\bullet}$	$f_{2\bullet}$	…	$f_{i\bullet}$	…	$f_{k\bullet}$

$$N=\sum_{i=1}^{k} n_{i\bullet} \qquad \sum_{i=1}^{k} f_{i\bullet}=1$$

$$a_{10}=\bar{x}=\frac{\sum_{i=1}^{k} x_i . n_{i\bullet}}{N} \qquad a_{20}=\frac{\sum_{i=1}^{k} x_i^2 . n_{i\bullet}}{N} \qquad m_{20} = s_x^2 = a_{20} - (a_{10})^2 = a_{20} - (\bar{x})^2$$

- **DISTRIBUCIÓN MARGINAL DE LA VARIABLE Y**

Y	y_1	y_2	…	y_j	…	y_m
$n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	…	$n_{\bullet j}$	…	$n_{\bullet m}$
$f_{\bullet j}=\frac{n_{\bullet j}}{N}$	$f_{\bullet 1}$	$f_{\bullet 2}$	…	$f_{\bullet j}$	…	$f_{\bullet m}$

$$N=\sum_{j=1}^{m} n_{\bullet j} \qquad \sum_{j=1}^{m} f_{\bullet j}=1$$

$$a_{01}=\bar{y}=\frac{\sum_{j=1}^{m} y_j . n_{\bullet j}}{N} \qquad a_{02}=\frac{\sum_{j=1}^{m} y_j^2 . n_{\bullet j}}{N} \qquad m_{02} = s_y^2 = a_{02} - (a_{01})^2 = a_{02} - (\bar{y})^2$$

Las variables (X, Y) son independientes cuando: $\frac{n_{ij}}{N}=\left(\frac{n_{i\bullet}}{N}\right)\left(\frac{n_{\bullet j}}{N}\right) \ \forall i,j$

Si (X,Y) independientes	$\mapsto \quad s_{yx} = 0$
Si $s_{yx} = 0$	$\mapsto$ (X,Y) No independientes

- **DISTRIBUCIÓN CONDICIONADA DE LA VARIABLE X para un valor** $Y = y_j$

Y / X	y_1	y_2	…	y_j	…	y_m	$n_{i\bullet}$
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1m}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2m}	$n_{2\bullet}$
…	…	…	…	…	…	…	…
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{im}	$n_{i\bullet}$
…	…	…	…	…	…	…	…
x_k	n_{k1}	n_{k2}	…	n_{kj}	…	n_{km}	$n_{k\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	…	$n_{\bullet j}$	…	$n_{\bullet m}$	$N = \sum_{i=1}^{k} n_{i\bullet} = \sum_{j=1}^{m} n_{\bullet j}$

X	x_1	x_2	…	x_i	…	x_k
$n(X/Y=y_j)$	n_{1j}	n_{2j}	…	n_{ij}	…	n_{kj}
$f(X/Y=y_j)$	f_{1j}	f_{2j}	…	f_{ij}	…	f_{kj}

$$f(X/Y=y_j) = \frac{n(x_i/Y=y_j)}{n_{\bullet j}}$$

- **DISTRIBUCIÓN CONDICIONADA DE LA VARIABLE Y para un valor** $X = x_i$

Y / X	Y_1	Y_2	…	y_j	…	y_m	$n_{i\bullet}$
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1m}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2m}	$n_{2\bullet}$
…	…	…	…	…	…	…	…
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{im}	$n_{i\bullet}$
…	…	…	…	…	…	…	…
x_k	n_{k1}	n_{k2}	…	n_{kj}	…	n_{km}	$n_{k\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	…	$n_{\bullet j}$	…	$n_{\bullet m}$	$N = \sum_{i=1}^{k} n_{i\bullet} = \sum_{j=1}^{m} n_{\bullet j}$

Y	y_1	y_2	…	y_j	…	y_m
$n(Y/X=x_i)$	n_{i1}	n_{i2}	…	n_{ij}	…	n_{im}
$f(Y/X=x_i)$	f_{i1}	f_{i2}	…	f_{ij}	…	f_{im}

$$f(Y/X=x_i) = \frac{n(Y/X=x_i)}{n_{i\bullet}}$$

MOMENTOS

Se define el momento respecto al par de valores (c, v) de órdenes r y s:

$$M_{rs}(c,v) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m}(x_i - c)^r (y_j - v)^s \; n_{ij}}{N}$$

Tienen especial interés dos casos particulares para los valores c y v

- **MOMENTOS RESPECTO AL ORIGEN (c, v) = (0, 0)**

$$a_{rs} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m}(x_i - 0)^r (y_j - 0)^s \; n_{ij}}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} x_i^r . y_j^s . n_{ij}}{N}$$

de interés son los particulares:

$$a_{00} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} x_i^0 . y_j^0 . n_{ij}}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij}}{N} = 1$$

$$a_{11} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} x_i^1 . y_j^1 . n_{ij}}{N}$$

$$a_{10} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} x_i^1 . y_j^0 . n_{ij}}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} x_i n_{ij}}{N} = \frac{\sum_{i=1}^{k} x_i n_{i\bullet}}{N} = \bar{x}$$

$$a_{01} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} x^0 . y_j^1 . n_{ij}}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} y_j n_{ij}}{N} = \frac{\sum_{j=1}^{m} y_j n_{\bullet j}}{N} = \bar{y}$$

$$a_{20} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} x_i^2 . y_j^0 . n_{ij}}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} x_i^2 n_{ij}}{N} = \frac{\sum_{i=1}^{k} x_i^2 n_{i\bullet}}{N}$$

$$a_{02} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} x^0 . y_j^2 . n_{ij}}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} y_j^2 n_{ij}}{N} = \frac{\sum_{j=1}^{m} y_j^2 n_{\bullet j}}{N}$$

- **MOMENTOS CENTRALES O RESPECTO A LAS MEDIAS** $(c, v) = (\bar{x}, \bar{y})$

$$m_{rs} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m}(x_i - \bar{x})^r (y_j - \bar{y})^s \; n_{ij}}{N}$$

de interés son los particulares:

$$m_{11} = s_{yx} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m}(x_i - \bar{x})\,(y_j - \bar{y})\; n_{ij}}{N} = s_{xy} \quad \text{covarianza}$$

$$m_{20} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m}(x_i - \bar{x})^2 (y_j - \bar{y})^0 \; n_{ij}}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m}(x_i - \bar{x})^2 n_{ij}}{N} = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 n_{i\bullet}}{N} = s_x^2 \quad \text{varianza de X}$$

$$m_{02} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m}(x_i - \overline{x})^0 (y_j - \overline{y})^2 \; n_{ij}}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m}(y_j - \overline{y})^2 \; n_{ij}}{N} = \frac{\sum_{j=1}^{m}(y_j - \overline{y})^2 \; n_{\bullet j}}{N} = s_y^2 \quad \text{varianza de Y}$$

- Se demuestra fácilmente que, $m_{11} = s_{xy} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m}(x_i - \overline{x})\,(y_j - \overline{y})\;\; n_{ij}}{N} = a_{11} - a_{10}.a_{01} = a_{11} - \overline{x}.\overline{y}$

$$m_{11} = s_{xy} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m}(x_i - \overline{x})\,(y_j - \overline{y})\;\; n_{ij}}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m}(x_i.y_j - x_i.\overline{y} - \overline{x}.y_j + \overline{x}.\overline{y})\; n_{ij}}{N} =$$

$$= \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} x_i.y_j.n_{ij}}{N} - \overline{y}.\frac{\sum_{i=1}^{k}\sum_{j=1}^{m} x_i.n_{ij}}{N} - \overline{x}.\frac{\sum_{i=1}^{k}\sum_{j=1}^{m} y_j.n_{ij}}{N} + \overline{x}.\overline{y}.\frac{\sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij}}{N} =$$

$$= \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} x_i.y_j.n_{ij}}{N} - \overline{y}.\frac{\sum_{i=1}^{k} x_i.n_{i\bullet}}{N} - \overline{x}.\frac{\sum_{j=1}^{m} y_j.n_{\bullet j}}{N} + \overline{x}.\overline{y}.\frac{\sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij}}{N} =$$

$$= a_{11} - \overline{y}.\overline{x} - \overline{x}.\overline{y} + \overline{x}.\overline{y} = a_{11} - \overline{x}.\overline{y} = a_{11} - a_{10}.a_{01}$$

DEPENDENCIA ENTRE LAS VARIABLES (X, Y)

Al observar dos caracteres en cada individuo se presenta el problema de determinar la existencia de algún tipo de dependencia entre ellos. En este sentido, conviene destacar dos tipos de dependencia:

- **Dependencia funcional:** Entre dos variables X e Y existe dependencia funcional cuando hay una expresión matemáticas que las relacione. Por ejemplo, los radios de una circunferencia (X) y las longitudes (Y).
- **Dependencia aleatoria:** Entre dos variables X e Y existe dependencia aleatoria cuando no existe una expresión matemática que las relacione. Por ejemplo, la edad de los niños (X) y la edad (Y).

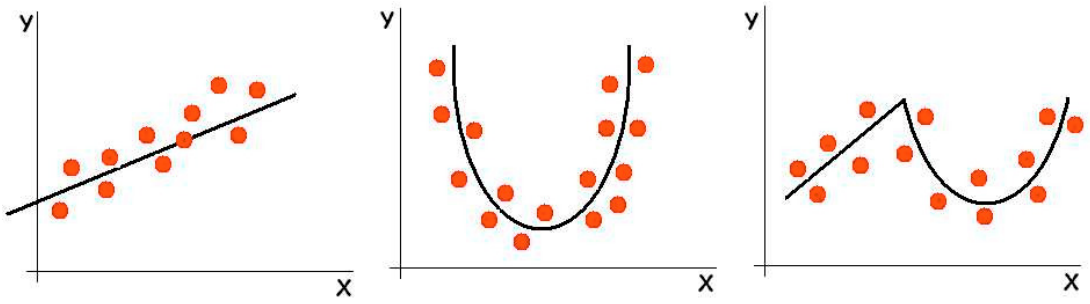
Señalar que existen variables entre las que no existe ningún tipo de dependencia, lo que conlleva a decir que los dos conceptos anteriores no son complementarios.

REGRESIÓN O AJUSTE

La observación de una variable estadística bidimensional (X, Y) comporta la representación de los puntos obtenidos en una **nube o diagrama de dispersión**. El problema general de regresión se plantea en el intento de ajustar una función de ecuación conocida (recta, parábola, exponencial, hipérbola, polinómica, etc.) a la nube de puntos con el interés de poder obtener una **predicción** aproximada de una de las variables a partir de la otra.

Naturalmente, que entre todas las funciones que se pueden elegir para ajustar a la nube de puntos, hemos de seleccionar la óptima, esto es, la que mejor encaje sobre los puntos que tenemos, para lo cual recurriremos al **método de los mínimos cuadrados**.

MÉTODO: Dependiendo de la forma que adopte la nube de puntos, en un principio sabremos si hemos de emplear una recta, una parábola, una función mixta, etc.



Una vez elegida la función, se estiman los parámetros correspondientes de la misma a partir de los datos observados. Por ejemplo, si la función elegida es una parábola:

$$y = a + b\,x + c\,x^2 \qquad \text{hemos de estimar} \quad a,\ b,\ c$$

Por último, una vez realizada la estimación hay que comprobar si efectivamente el ajuste era el idóneo o no. Para ello se emplean cualquiera de los tests construidos para estudiar la bondad del ajuste. El modelo más utilizado es el de la χ^2 (chi-cuadrado).

REGRESIÓN LINEAL MÍNIMO CUADRÁTICA

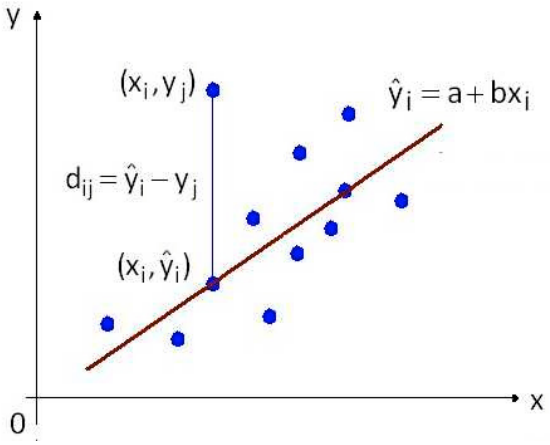
En el supuesto de que sea la recta la función que mejor se comporta con arreglo a la forma de la nube de puntos, nos encontramos ante una problema de regresión lineal, distinguiendo entre:

- **Recta de regresión de Y sobre X:** Obteniendo valores aproximados de la Y conocidos los de la X
- **Recta de regresión de X sobre Y:** Obteniendo valores aproximados de la X conocidos los de la Y

RECTA DE REGRESIÓN DE Y SOBRE X

En cada par (X,Y) al valor observado x_i le corresponde un valor observado y_j y otro valor teórico $\hat{y}_i$ que sería el que le correspondería en la recta como función, es decir: $\hat{y}_i = a + b\,x_i$

A la distancia entre estos dos valores (teórico y experimental), la denotamos por $d_{ij} = \hat{y}_i - y_j$



Para obtener los parámetros **a** y **b**, se toman las distancias (errores) al cuadrado para que no se contrarresten los signos positivos y negativos, haciendo mínima su suma: $M = \sum_{i,j} d_{i,j}^2 = \sum_{i,j} (\hat{y}_i - y_j)^2$

Por otra parte, para simplificar el mecanismo para obtener la recta de regresión de Y *(variable dependiente)* sobre X *(variable independiente)*, se descartan multiplicidades y suponemos que cada par se repite una sola vez.

Considerando que $\hat{y}_i = a + b\,x_i$, $M = \sum_{i,j} d_{i,j}^2 = \sum_{i,j} (a + b\,x_i - y_j)^2$

Para hallar los valores de **a** y **b** que hagan mínima esta función hemos de hallar las derivadas, igualando a cero las ecuaciones resultantes:

$$\frac{\vartheta M}{\vartheta a} = 2\sum_{i,j}(a + b\,x_i - y_j) = 0 \qquad \Rightarrow \quad \sum_{i,j}(a + b\,x_i - y_j) = 0$$

$$\frac{\vartheta M}{\vartheta b} = 2\sum_{i,j}(a + b\,x_i - y_j)(x_i) = 0 \quad \Rightarrow \quad \sum_{i,j}(a + b\,x_i - y_j)(x_i) = 0$$

Por las propiedades del sumatorio, se obtienen las ***ecuaciones normales*** de la regresión:

$$\begin{cases} \sum_i a + b\sum_i x_i - \sum_j y_j = 0 \\ a\sum_i x_i + b\sum_i x_i^2 - \sum_{i,j} x_i\, y_j = 0 \end{cases} \Rightarrow \begin{cases} \sum_i a + b\sum_i x_i = \sum_j y_j \\ a\sum_i x_i + b\sum_i x_i^2 = \sum_{i,j} x_i\, y_j \end{cases}$$

Dividiendo las expresiones anteriores por N (número total de datos), habiendo supuesto que la frecuencia absoluta de cada par (X, Y) es la unidad, resulta:

$$\left.\begin{array}{l} a\dfrac{\sum_i 1}{N} + b\dfrac{\sum_i x_i}{N} = \dfrac{\sum_j y_j}{N} \\ a\dfrac{\sum_i x_i}{N} + b\dfrac{\sum_i x_i^2}{N} = \dfrac{\sum_{i,j} x_i\, y_j}{N} \end{array}\right\}$$

Considerando los momentos, se tiene: $\begin{cases} a + b\,\bar{x} = \bar{y} \\ a\,\bar{x} + b\,a_{20} = a_{11} \end{cases}$

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

sustituyendo en la ecuación $a\,\bar{x} + b\,a_{20} = a_{11}$, resulta:

$$(\bar{y} - b\bar{x})\,\bar{x} + b\,a_{20} = a_{11} \;\mapsto\; b(a_{20} - \bar{x}^2) = a_{11} - \bar{x}\,\bar{y} \;\mapsto\; \boxed{b = \frac{a_{11} - \bar{x}\,\bar{y}}{a_{20} - \bar{x}^2} = \frac{s_{xy}}{s_x^2}}$$

$$\boxed{a = \bar{y} - \frac{s_{xy}}{s_x^2}\,\bar{x}}$$

Finalmente, sustituyendo los valores obtenidos en la ecuación de la recta $y = a + b\,x$

$$y = \bar{y} - \frac{s_{xy}}{s_x^2}\,\bar{x} + \frac{s_{xy}}{s_x^2}\,x \quad \Rightarrow \quad \boxed{y - \bar{y} = \frac{s_{xy}}{s_x^2}\,(x - \bar{x})}$$

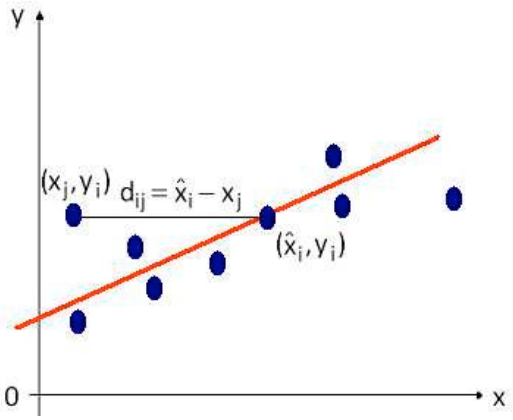
NOTA.- En el supuesto de que no hubiéramos partido de las hipótesis iniciales para el desarrollo, es decir, si hay multiplicidades de (x_i, y_j) y si cada par se repite n_{ij} veces, la ecuación a minimizar sería

$$M = \sum_{i,j} d_{i,j}^2\; n_{ij} = \sum_{i,j}(a + b\,x_i - y_j)^2\, n_{ij}$$

RECTA DE REGRESIÓN DE X SOBRE Y

Si en lugar de tomar las distancias d_{ij} sobre las verticales (esto es, sobre la Y) se toman sobre las horizontales (sobre la X) y se utiliza el mismo método de los mínimos cuadrados, por un proceso idénticamente igual se llega a la **ecuación de regresión de X sobre Y**:

$$x - \overline{x} = \frac{s_{xy}}{s_y^2}(y - \overline{y})$$



COEFICIENTES DE REGRESIÓN LINEAL

- La **recta de regresión de Y sobre X**: $y - \overline{y} = \frac{s_{xy}}{s_x^2}(x - \overline{x})$, donde el coeficiente de regresión lineal

$b_{yx} = \frac{s_{xy}}{s_x^2}$ es la pendiente de la recta.

Recta de regresión de Y sobre X, según el coeficiente de regresión $b_{yx} \equiv \begin{cases} >0 & \text{creciente} \\ =0 & \text{horizontal} \\ <0 & \text{decreciente} \end{cases}$

- La **recta de regresión de X sobre Y**: $x - \overline{x} = \frac{s_{xy}}{s_y^2}(y - \overline{y})$, donde el coeficiente de regresión lineal

$b_{xy} = \frac{s_{yx}}{s_y^2}$ es la pendiente de la recta.

Recta de regresión de X sobre Y, según el coeficiente de regresión $b_{xy} \equiv \begin{cases} >0 & \text{creciente} \\ =0 & \text{vertical} \\ <0 & \text{decreciente} \end{cases}$

CORRELACIÓN

Así como la regresión estudia la posible predicción de los valores de una variable a partir de la otra, la correlación estudia el tipo de dependencia que existe entre ambas variables, intentando cuantificarla mediante el cálculo de los coeficientes de correlación.
A continuación se estudian los coeficientes de determinación y correlación lineal.

COEFICIENTE DE CORRELACIÓN LINEAL

- El **coeficiente de correlación lineal** R es un número abstracto que determina el grado de ajuste entre una nube de puntos y una recta de regresión. Se define como la media geométrica de los coeficientes de correlación lineal

$$r = \sqrt{b_{yx} \cdot b_{xy}} = \sqrt{\frac{s_{xy}}{s_x^2}\frac{s_{xy}}{s_y^2}} = \frac{s_{xy}}{s_x s_y}$$

RELACIÓN ENTRE LOS COEFICIENTES DE REGRESIÓN Y DE CORRELACIÓN

- **Recta de regresión de Y sobre X**: $y-\bar{y}=\frac{s_{xy}}{s_x^2}(x-\bar{x})$, coeficiente de regresión lineal $b_{yx}=\frac{s_{xy}}{s_x^2}$

  (pendiente de la recta), coeficiente de correlación $r=\frac{s_{xy}}{s_x\,s_y}$ (grado de ajuste)

$$\left.\begin{array}{l} b_{yx}=\dfrac{s_{xy}}{s_x^2} \\ r=\dfrac{s_{xy}}{s_x\,s_y} \end{array}\right\} \mapsto \left\{\begin{array}{l} s_{xy}=b_{yx}\,s_x^2 \\ s_{xy}=r\,s_x\,s_y \end{array}\right\} \mapsto b_{yx}\,s_x^2=r\,s_x\,s_y \mapsto \boxed{b_{yx}=r\,\frac{s_y}{s_x}}$$

se observa que los dos coeficientes (regresión lineal y correlación lineal) tienen el mismo signo.

- **Recta de regresión de X sobre Y**: $x-\bar{x}=\frac{s_{xy}}{s_y^2}(y-\bar{y})$, coeficiente de regresión lineal $b_{xy}=\frac{s_{xy}}{s_y^2}$

  (pendiente de la recta), coeficiente de correlación $r=\frac{s_{xy}}{s_x\,s_y}$ (grado de ajuste)

$$\left.\begin{array}{l} b_{xy}=\dfrac{s_{xy}}{s_y^2} \\ r=\dfrac{s_{xy}}{s_x\,s_y} \end{array}\right\} \mapsto \left\{\begin{array}{l} s_{xy}=b_{xy}\,s_y^2 \\ s_{xy}=r\,s_x\,s_y \end{array}\right\} \mapsto b_{xy}\,s_y^2=r\,s_x\,s_y \mapsto \boxed{b_{xy}=r\,\frac{s_x}{s_y}}$$

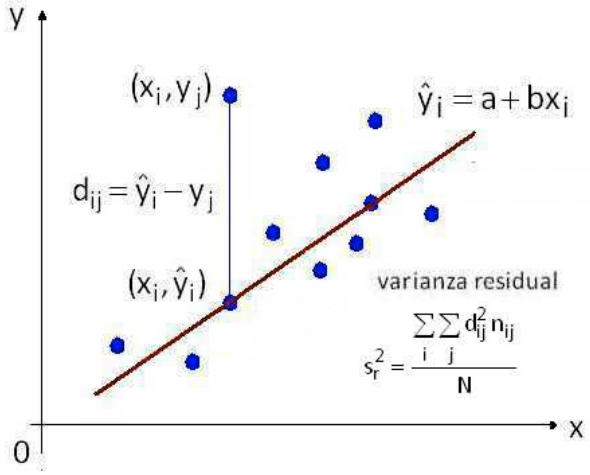
los dos coeficientes (regresión lineal y correlación lineal) tienen el mismo signo.

VARIANZA RESIDUAL

Es la dispersión de los errores cometidos entre los residuos, dispersión de la suma de las distancias de los valores observados (o experimentales) y los valores teóricos (en la recta de regresión).

* Las diferencias se toman al cuadrado para que no se puedan contrarrestar los signos positivos y negativos.

$$s_r^2=\frac{\sum_{i,j} d_{i,j}^2\,n_{ij}}{N}$$



Para simplificar el mecanismo suponemos que el centro de gravedad se encuentra en el origen $(\bar{x}=0,\ \bar{y}=0)$, con lo que la ecuación de la recta $y=a+bx$ se reduce a $y=bx$, partiendo que cada par (x_i, y_j) se repite una sola vez (descartando multiplicidades).

Con las hipótesis planteadas: $(\bar{x}=0,\ \bar{y}=0)\quad b=\dfrac{s_{xy}}{s_x^2}$

$$\begin{cases} s_{xy} = a_{11} - \bar{x}\,\bar{y} = a_{11} \\ s_x^2 = a_{20} - \bar{x}^2 = a_{20} \\ s_y^2 = a_{02} - \bar{y}^2 = a_{02} \end{cases}$$

Con lo cual,

$$\sum_{i,j} d_{i,j}^2 = \sum_{ij} (\hat{y}_i - y_j)^2 = \sum_{i,j} (b x_i - y_j)^2 = \sum_{i,j} (b^2 x_i^2 + y_j^2 - 2b x_i y_j) = b^2 \sum_i x_i^2 + \sum_j y_j^2 - 2b \sum_{ij} x_i y_j$$

$$s_r^2 = \frac{\sum_{i,j} d_{i,j}^2}{N} = b^2 \frac{\sum_i x_i^2}{N} + \frac{\sum_j y_j^2}{N} - 2b \frac{\sum_{ij} x_i y_j}{N} = b^2 a_{20} + a_{02} - 2b a_{11} = b^2 s_x^2 + s_y^2 - 2b s_{yx}$$

$$s_r^2 = b^2 s_x^2 + s_y^2 - 2b s_{yx} = b(b s_x^2 - 2 s_{xy}) + s_y^2 = \frac{s_{xy}}{s_x^2}\left[\frac{s_{xy}}{s_x^2} s_x^2 - 2 s_{xy}\right] + s_y^2 =$$

$$= \frac{s_{xy}}{s_x^2}\left[s_{xy} - 2 s_{xy}\right] + s_y^2 = -\frac{s_{xy}^2}{s_x^2} + s_y^2 = s_y^2\left[1 - \frac{s_{xy}^2}{s_x^2 s_y^2}\right] = s_y^2\left[1 - r^2\right]$$

- ✓ La cota máxima de la varianza residual s_r^2 es la varianza que tratamos de explicar mediante el modelo de regresión, es decir, la varianza de la variable dependiente. En este caso, $s_r^2 = s_y^2$, hecho que sucede cuando $r=0$, esto es cuando las **variables son incorreladas**.

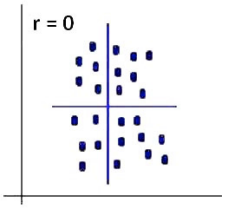
- ✓ La cota mínima de la varianza residual s_r^2 se obtendrá cuando las variables tienen una **dependencia funcional** $r^2 = 1$

- ✓ $\%\ \text{variaciones no explicado} = 100\,\dfrac{s_r^2}{s_y^2}$

- ✓ Una forma de definir el **coeficiente de determinación**: $r^2 = 1 - \dfrac{s_r^2}{s_y^2} \qquad 0 \le r^2 \le 1$

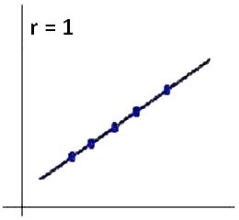
INTERPRETACIÓN COEFICIENTE DE CORRELACIÓN LINEAL

Se hace una interpretación a partir de la relación con la varianza residual $s_r^2 = s_y^2(1-r^2)$:



- Si $r=0 \Rightarrow s_r^2 = s_y^2$ y $b_{yx}=0$ y $b_{xy}=0$.

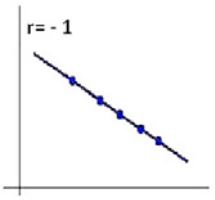
Las dos rectas son perpendiculares y las variables son **INCORRELADAS**



- Si $r=1 \Rightarrow s_r^2 = 0$.

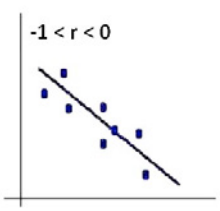
Todos los puntos se encuentran situados sobre la recta de regresión, existiendo entre las dos variables una **DEPENDENCIA FUNCIONAL** (recta de regresión creciente).

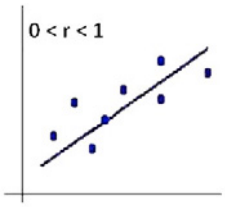
- Si $r=-1 \Rightarrow s_r^2=0$.



Todos los puntos se encuentran situados sobre la recta de regresión, existiendo entre las dos variables una **DEPENDENCIA FUNCIONAL** (recta de regresión decreciente).

- Si $-1<r<0$ ó $0<r<1$





Las variables están tanto más correladas en cuanto el coeficiente se aproxima más a -1 ó 1, respectivamente.
En ambos casos, existe una **DEPENDENCIA ALEATORIA** entre las variables.

Resaltar que no puede darse el caso de que una recta de regresión sea creciente y la otra decreciente $\begin{cases} b_{yx}>0 \Leftrightarrow r>0 \Leftrightarrow b_{xy}>0 \\ b_{yx}<0 \Leftrightarrow r<0 \Leftrightarrow b_{xy}<0 \end{cases}$

DESCOMPOSICIÓN DE LA VARIABILIDAD: COEFICIENTE DE CORRELACIÓN - VARIANZA RESIDUAL

Sea $\hat{y}_i$ el valor teórico que correspondería en la recta de regresión de Y sobre X: $\hat{y}_i=a+bx_i$.
Elevando al cuadrado la descomposición $(y_i-\bar{y})=(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y})$, se obtiene:

$$SCT=\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}\left[(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y})\right]^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+2\overbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i).(\hat{y}_i-\bar{y})}^{=0}$$

Observemos que, $\sum_{i=1}^{n}(y_i-\hat{y}_i).(\hat{y}_i-\bar{y})=\sum_{i=1}^{n}(y_i-a-bx_i).(a+bx_i-\bar{y})=$

$$=a\underbrace{\sum_{i=1}^{n}(y_i-a-bx_i)}_{=0}+b\underbrace{\sum_{i=1}^{n}x_i(y_i-a-bx_i)}_{=0}+\bar{y}\underbrace{\sum_{i=1}^{n}(y_i-a-bx_i)}_{=0}$$

con lo cual,

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{SCT}=\underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{SCR}+\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{SCE}$$

suma cuadrados total | suma cuadrados residual | suma cuadrados explicada

Por otro lado, $\overbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}^{SCT}=\overbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}^{SCR}+\overbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}^{SCE} \Rightarrow 1=\underbrace{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}}_{SCR/SCT}+\underbrace{\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}}_{r^2=SCE/SCT}$

Una vez estimado el modelo es conveniente obtener una medida acerca de la bondad del ajuste realizado. Un estadístico que facilita esta medida es el Coeficiente de Determinación (r^2), que se

define: $$\boxed{r^2 = \frac{SCE}{SCT} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

El Coeficiente de Determinación permite, además, seleccionar entre modelos clásicos que tengan el mismo número de regresores, ya que la capacidad explicativa de un modelo es mayor cuanto más elevado sea el valor que tome este coeficiente.

De otra parte, $$r^2 = 1 - \frac{SCR}{SCT} = 1 - \frac{\overbrace{\sum(y_i - \hat{y}_i)^2 / N}^{s_r^2}}{\underbrace{\sum(y_i - \bar{y})^2 / N}_{s_y^2}} = 1 - \frac{s_r^2}{s_y^2} \mapsto \overbrace{s_r^2 = s_y^2(1 - r^2)}^{\text{varianza residual}}$$

Considerando la recta de regresión de Y sobre X, el coeficiente de determinación r^2 puede expresarse:

$$r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\sum\left[\frac{s_{xy}}{s_x^2}(x_i - \bar{x})\right]^2}{\sum(y_i - \bar{y})^2} = \frac{\left[\frac{s_{xy}}{s_x^2}\right]^2 \frac{\sum(x_i - \bar{x})^2}{N}}{\frac{\sum(y_i - \bar{y})^2}{N}} = \frac{s_{xy}^2}{s_x^2 s_y^2} \mapsto \overbrace{r = \frac{s_{xy}}{s_x s_y}}^{\text{coeficiente correlación}}$$

El coeficiente de correlación lineal r es un número abstracto que determinará el grado de ajuste entre una nube de puntos y una recta de regresión. Se define como la media geométrica de los coeficientes de regresión lineal:

$$r = \sqrt{b_{yx}\, b_{xy}} = \sqrt{\frac{s_{yx}}{s_x^2}\frac{s_{yx}}{s_y^2}} = \frac{s_{yx}}{s_x\, s_y}$$

Adviértase que si la varianza residual es cero, $s_r^2 = 0$, se tiene, $s_r^2 = s_y^2(1 - r^2) = 0 \mapsto 1 - r^2 = 0$

con lo cual, $r^2 = 1 \Rightarrow s_{xy} = \pm s_x\, s_y$