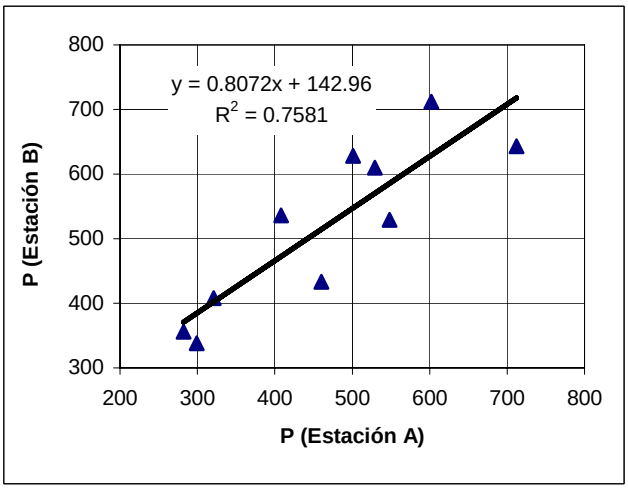
# Correlación lineal y regresión

Si disponemos de dos series de datos emparejadas, con frecuencia es útil conocer si ambas variables están relacionadas, y, en caso afirmativo, encontrar la expresión que refleja dicha relación. Si la ecuación que mejor relaciona dichas variables es la de una recta, decimos que existe **correlación lineal**.

Un ejemplo puede ser la pluviometría registrada en dos estaciones próximas (Tabla adjunta). Si la pluviometría es similar en ambos puntos, sería de gran utilidad cuantificar esa relación, pues de ese modo podríamos evaluar, aunque fuera de modo aproximado, la pluviometría de un lugar a partir de la registrada en el otro.

| **x** | **y** |
|---|---|
| Estación A | Estación B |
| 321 | 408 |
| 548 | 529 |
| 460 | 433 |
| 712 | 643 |
| 602 | 712 |
| 282 | 356 |
| 529 | 610 |
| 408 | 536 |
| 501 | 628 |
| 299 | 338 |
| | |
| 640 | ?? |
| | |




En este ejemplo, la correlación es: $P_{en\ B} = 0{,}807 \cdot P_{en\ A} + 142{,}96$

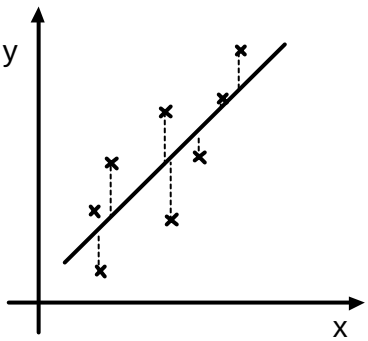
Supongamos que para un año conocemos el valor de P = 640 en el punto **A,** pero no lo tenemos para el punto **B**. Se podrá estimar mediante la relación anterior:

$$P_{en\ B} = 0{,}807 \cdot 640 + 14{,}96 = 531$$

La relación entre dos variables (como las dos columnas de datos anteriores) puede ser lineal, exponencial, polinómica, etc. Es decir: que aunque los puntos no estén alineados puede que tengan una fuerte correlación, pero no lineal (por ejemplo: $y = x^2 + 2{,}3$). En este breve apunte vamos a centrarnos en la posible relacion lineal entre dos variables.

## *Recta de regresión*

Se llama así a la recta que atraviesa la nube de puntos y que mejor se ajusta a ellos. Supongamos que medimos la distancia vertical de cada punto a la recta (líneas de trazos en la figura adjunta). La recta buscada sería aquella para la que la suma de estas distancias fuera mínima.




La ecuación de una recta es:

$$y = a \,.\, x + b$$

Si, por ejemplo, fuera: $y = 0{,}83\ x + 12{,}9$
la pendiente sería 0,83 y la ordenada en el origen (altura a la que la recta corta el eje vertical) sería 12,9

Si llegamos a conocer esa ecuación, podremos llegar a estimar valores de y desconocidos a partir de valores de x conocidos. Otro ejemplo: supongamos que **x** es la altitud de cada estación pluviométrica e **y** es su pluviometría; si establecemos que ambas variables están correlacionadas y obtenemos la **ecuación** de la recta de regresión, conociendo la cota del punto podremos estimar su pluviometria.

## *Coeficiente de correlación de Pearson (r)*

Este coeficiente nos informa del grado de relación entre dos variables. Si la relación es lineal perfecta, **r** será 1 ó -1. El coeficiente **r** será positivo si la relación es positiva (al aumentar **x** aumenta **y**), y **r** será negativo en el caso contrario (si al aumentar **x**, disminuye **y**).

En general, valores (absolutos) de **r** > 0,80 se consideran altos, aunque esto depende del número de parejas de datos con las que hemos realizado el cálculo y del nivel de seguridad con el que queramos extraer nuestras conclusiones.

No vamos a entrar en el estudio del nivel de significación del coeficiente **r** , pero como indicación: para 11 parejas de datos, y si admitimos un 5% de posibilidades de equivocarnos, con **r**>0,553 ya podemos decir que ambas series de datos **no son independientes** (parece que tienen algún tipo de relación). Si tuviéramos 50 parejas de datos, nos bastaría **r**>0,273 para sacar la misma conclusión (siempre considerando el valor absoluto de **r**)

Si nos ponemos más estrictos, y queremos sacar la conclusión de que las dos series no son independientes con un 99% de seguridad (sólo un 1% de posibilidad de error), con 11 parejas necesitamos que **r**>0,684 y con 50 parejas **r**>0,354

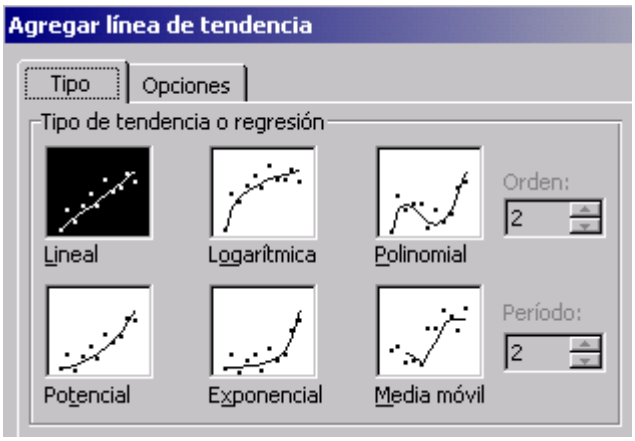
Precauciones:

1. El que estemos seguros de que ambas series están relacionadas, **no** quiere decir que la relación sea tan estrecha como para estimar valores de **y** desconocidos a partir de valores de **x** conocidos; éso dependerá del error de estimación que aceptemos.

2. La existencia de una correlación no indica relación causa-efecto.

## *Cálculo de la recta de regresión y del coeficiente r con Excel*

Vamos a elaborar el gráfico de la primera figura, con la ecuación y el coeficiente ***r.*** (puedes copiar las parejas de valores del ejemplo inicial).

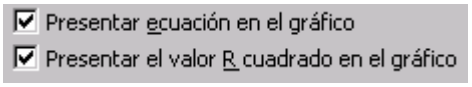



1. Seleccionar las dos columnas de datos

2. **Insertar** > **Gráfico** (con el menu Insertar o con el botón correspondiente de una de las barras). Tipo de gráfico: **XY (dispersión)**

3. Click con botón derecho sobre uno cualquiera de los puntos; en el menu que surge, elegir **Agregar línea de tendencia.** Elegir Lineal

4. Sin cerrar el cuadro, en la pestaña **Opciones**, marcar los recuadros de :




Obtenemos un gráfico como el que aparece en la página anterior.

**Atención**: Excel calcula $r^2$, no **r** . ($r^2$ se llama *coeficiente de determinación*).

El cuadro de Excel nos recuerda que quizá la correlación lineal sea mala, pero otro tipo de correlación puede ser buena (logarítmica, polinómica, potencial, exponencial,...)

* * * * *

Finalmente, con la ecuación obtenida, rellenamos los datos que faltan. En el ejemplo de la página anterior, en Excel habría que escribir lo siguiente:

B14 = A14 * 0,8072 +142,96

(B14 sería la celda a la que le falta el dato, A14 la que está a su izquierda)

> En la sección *Complementos* existe un documento con las nociones imprescindibles para escribir fórmulas en Excel.

Con la mayoría de las calculadoras científicas puede calcularse la ecuación de la recta y el coeficiente **r**. Siempre es instructivo realizar los cálculos a mano ¡al menos una vez en la vida! (Ver, por ejemplo: DOWNIE & HEATH, *Métodos Estadísticos Aplicados*, Ed. Castillo)