Práctica 5.1

Un grupo de Enseñanza Secundaria ha elaborado una encuesta sobre las horas diarias que emplean en el estudio y la calificación obtenida en Matemáticas en el último examen.

Han recogido los resultados en la siguiente tabla:

Horas de estudio	0	0	1	1	1	1	1	2	2	2	2	2	3	4	4	5
Calificación	2	1	3	4	3	2	2	4	5	7	8	6	5	8	10	7

Además de estudiar el grado de asociación entre las dos variables, que ya se explicó en el tema anterior mediante el coeficiente de correlación, nos puede interesar hacer pronósticos: *¿Qué nota puedo esperar si estudio 2 horas y meda?* Para realizar esos pronósticos usaremos las técnicas de Regresión.

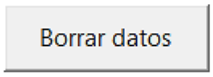
Puede ser interesante, en primer lugar, determinar el grado de paralelismo que existe entre ambas variables. Para averiguarlo, abre un modelo similar al que usamos en el tema 4. Te repetimos las instrucciones:

- Selecciona la tabla en la página de inicio de este tema (*tema5.htm*) y pide **Copiar** con la combinación de teclas **CTRL+C**.

- Abre el archivo ***regresion.ods***, selecciona la hoja ***Borrador***, y en cualquier celda y pide **Pegado Especial** como **Formato HTM.**

- Selecciona en esa hoja *Borrador* **sólo los datos numéricos**.

- Borra la zona de entrada de datos de la hoja ***Cálculo*** con el botón correspondiente



- Pega los datos en esa zona mediante **Pegado Especial**, pero acordándote de activar ***Transponer***, para que queden en columna.

Si no te apetece realizar las operaciones anteriores, escribe los datos de forma manual.

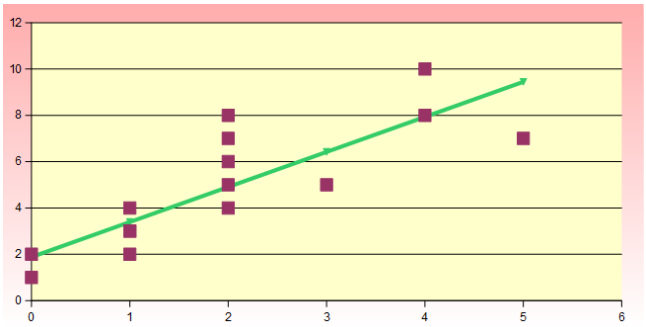
Columna de la variable X	Columna para la Y	Estimación lineal
0	2	1,88
0	1	1,88
1	3	3,39
1	4	3,39
1	3	3,39
1	2	3,39
1	2	3,39
2	4	4,91
2	5	4,91
2	7	4,91
2	8	4,91
2	6	4,91
3	5	6,42
4	8	7,93
4	10	7,93
5	7	9,45

Con esta operación descubriremos que el Coeficiente de Correlación es alto y positivo: **0,824**, luego podemos afirmar que

Existe un grado de asociación importante y positiva entre las horas de estudio y las calificaciones que se dan en nuestro grupo.

En el Gráfico de dispersión se observa una tendencia de los datos de pasar de abajo a la izquierda hasta arriba a la derecha, es decir, que a menos horas corresponden calificaciones bajas y más horas mejores notas. Para recalcar esta tendencia, aunque suponga un pequeño

adelanto en la teoría, se le ha dibujado una línea recta que la representa.



Esta recta que hemos dibujado es la llamada **recta de regresión lineal**, y es el principal objeto del estudio de este tema.

Recta de regresión lineal

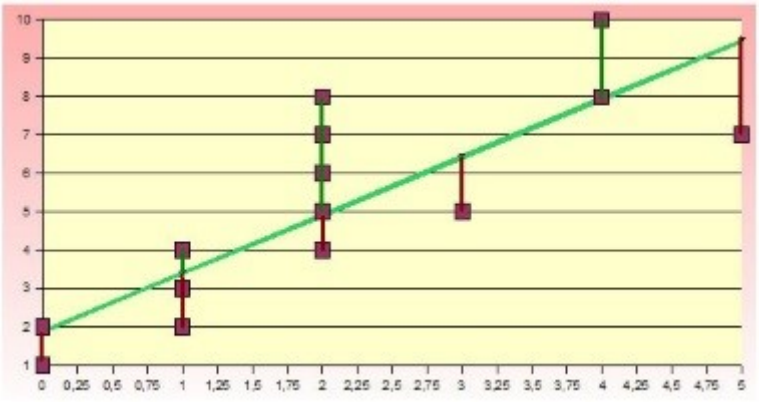
Dada una distribución bidimensional simple, con datos X-Y cuantitativos, llamaremos **recta de regresión** correspondiente a esa distribución a aquella que mejor se adapta al gráfico de dispersión XY, también llamado *Nube de puntos*. Este acercamiento se define de forma rigurosa como

La recta de regresión de Y sobre X es aquella que minimiza la suma de cuadrados de las diferencias entre los valores de Y y los correspondientes Y' (para el mismo valor de X) medidos en dicha recta.

Puedes ir consultando la teoría de forma simultánea al desarrollo de esta actividad

En la imagen se han vuelto a dibujar la nube de puntos y la recta, y algunas diferencias (en verde las positivas y en rojo las negativas) entre los datos verdaderos y los que estarían medidos en la misma recta. Esas diferencias son las que

deben cumplir que la suma de sus cuadrados sea la mínima posible. En la práctica lo que se pretende es que el ajuste entre recta y nube sea el mejor posible.

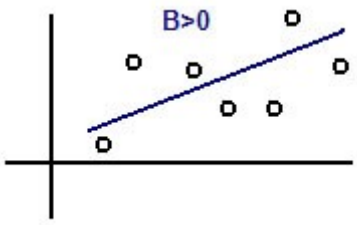


A partir de esta propiedad, y aplicando cálculos un poco complejos que no vienen al caso en este curso, se puede determinar la ecuación de la línea recta que cumple la propiedad deseada. Esta ecuación es del tipo siguiente:

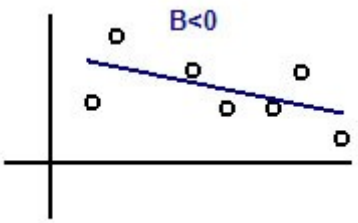
$Y' = A + BX$

donde el coeficiente B representa la tasa de cambio o ***pendiente*** y el coeficiente A es el valor correspondiente a X=0, y la llamaremos ***ordenada en el origen.*** porque es el punto en el que la línea recta corta al eje Y***.*** La variable **Y'** se escribe así para distinguir el verdadero valor de una medida, que sería Y, de su valor correspondiente en la línea de regresión, al que representaremos por Y' y le llamaremos ***pronóstico o predicción***.

Según el signo de la pendiente B, hablaremos de relación ***positiva o creciente***



y de relación ***negativa o decreciente***.



La fórmula para B se demuestra que es

$$B = \frac{S_{xy}}{S_x^2}$$

es decir, la ***covarianza*** existente entre X e Y dividida entre la ***varianza de X***

y la de A

$$A = \bar{Y} - B\bar{X}$$

que podemos expresar como la diferencia entre la media de Y y la de X multiplicada por B

Aunque no demostremos las fórmulas, sí podemos comprobar que estos valores son los adecuados para conseguir que la suma de errores al cuadrado sea mínima.

Abre el documento ***minimocuad.pdf*** para profundizar en el tema.

Si volvemos a nuestra cuestión inicial, esta técnica que presentamos nos permitirá obtener una fórmula para los pronósticos. Si has escrito bien los datos en el archivo ***regresion.ods*** y has obtenido el coeficiente de correlación 0,824, basta que leas al lado la ecuación de regresión de esa encuesta:

Ecuación de regresión
Y=1,881 + 1,513X

Podemos interpretar esta relación como sigue: *si multiplicamos las horas de estudio por 1,513 y le sumamos 1,881, obtendremos un pronóstico para la calificación que esperamos*.