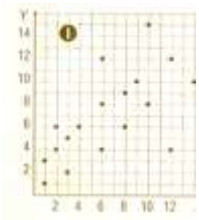
DOCUMENTO 11: RECTA DE REGRESIÓN

Cuando la nube de puntos sugiere algún tipo de dependencia entre las dos variables X e Y, condensándose los puntos alrededor de una cierta línea, podemos tratar de encontrar la recta que mejor se ajuste a la distribución. Esta recta se llama **recta de regresión.**



¿Qué entendemos por la línea que mejor se ajusta al diagrama de dispersión? Aquella línea que haga que la suma de las desviaciones de los puntos de la nube de puntos respecto de los correspondientes de la línea sea lo menor posible.
En estas condiciones, diremos que es la línea que menos se separa de la nube de puntos.

Para encontrar la ecuación de la recta que mejor se ajuste a la nube, el método más utilizado es el de los mínimos cuadrados. Este método consiste en hacer mínima la suma de los cuadrados de las diferencias entre los valores observados experimentalmente y los teóricos que se obtengan mediante la recta.

De la aplicación de este método se deduce que la recta pasa por el punto ($\overline{x}, \overline{y}$). Su ecuación es:

$$y - \overline{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \overline{x})$$

Esta ecuación se llama recta de regresión de Y sobre X. Sustituyendo en esta ecuación los valores de X podemos obtener, con cierta aproximación, los valores esperados para la variable Y que llamamos estimaciones o previsiones.

Si lo que queremos es estimar los valores de X partiendo de los de Y, utilizaremos la ecuación de la recta de regresión de X sobre Y, que es:

$$x - \overline{x} = \frac{\sigma_{xy}}{\sigma_y^2}(y - \overline{y})$$

Las rectas de regresión de Y sobre X y de X sobre Y son distintas, por eso hay que saber qué variable es la dependiente, pues X e Y no son intercambiables.

Recuerda el ejemplo que vimos en el Documento 9, que relacionaba los gastos en publicidad (en miles de euros) y las ventas de una compañía (en miles de euros). Para la misma podría resultar interesante averiguar qué ventas podría conseguir si invierte 18000 euros en publicidad, o bien, cuánto necesitaría invertir para que sus ventas ascendieran a 23000 euros.

En el primer caso sabiendo que x = 18000 euros queremos averiguar el valor de y. Por lo tanto, para realizar nuestras estimaciones utilizaríamos la recta de regresión de Y sobre X.

En el segundo caso sabiendo que y = 23000 euros queremos averiguar el valor de x. Por lo tanto, para realizar nuestras estimaciones utilizaríamos la recta de regresión de X sobre Y.

Ejemplo: Una persona se somete a una dieta de adelgazamiento durante 5 semanas. A continuación se detalla su peso al término de cada una de esas semanas:

Semanas de dieta (X)	1	2	3	4	5
Peso en Kg. (Y)	88,5	87	84	82,5	79

a) ¿Es razonable suponer que existe correlación lineal entre el peso y la dieta? ¿Cómo puede expresarse esa correlación?
b) ¿Qué peso es de esperar que alcance esa persona si sigue la dieta durante dos semanas más? ¿Y si hace la dieta durante 25 semanas?

¿Qué fiabilidad podemos conceder a las estimaciones obtenidas a través de la recta de regresión?

- Si r es próximo a 0, no tiene sentido realizar previsiones.
- Si r es próximo a 1 ó -1, probablemente los valores reales sean próximos a nuestras estimaciones.
- Si r = 1 ó r = -1, las estimaciones realizadas coincidirán con los valores reales.

Debemos tener en cuenta, además, que:

- La recta de regresión debe usarse para hacer estimaciones en valores próximos a los considerados. Pretender una estimación en puntos lejanos puede conducir a situaciones absurdas.
- La fiabilidad aumenta al aumentar los datos. Una recta obtenida a partir de pocos datos genera grandes riesgos, aunque r sea muy alto.

ACTIVIDADES

4.- Una distribución bidimensional en la que los valores de x son: 12, 15, 17, 21, 22 y 25, tiene una correlación r = 0,99 y su recta de regresión es: y = 10,5 + 3,2x.
Calcula los valores de y para x = 20, x = 13, x = 30 y x = 100. ¿Cuáles de las estimaciones anteriores son fiables, cuál es poco fiable y cuál no se debe hacer?

5.- Los datos de la evolución del crecimiento del PIB y del empleo en España (en porcentaje) durante los últimos nueve años están recogidos en la siguiente tabla:

PIB	4,7	3,7	2,3	0,7	-1,2	2,1	2,8	2,4	3,1
Empleo	4,1	2,6	0,2	-1,9	-4,3	-0,9	2,8	3,3	3

a) Dibuja la nube de puntos y estudia la relación entre ambas variables. Si existe correlación lineal, calcula el coeficiente de Pearson.

b) ¿Cuánto crecerá el empleo suponiendo que el PIB crecerá un 3,4% el próximo año

6.- Una empresa dedicada a la elaboración y venta de ropa de jóvenes ha realizado los gastos en publicidad y ha obtenido las ventas que figuran en la siguiente tabla:

Publicidad(en miles de euros)	45	48	51	60	63	72	78	84	90	108
Ventas (en miles de euros)	1202	1232	1382	1442	1502	1622	1683	1803	1863	1953

Si denominamos X a la variable *Gastos en publicidad* e Y a *Beneficios de ventas*, halla:

a) El coeficiente de correlación lineal. Analiza la dependencia entre ambas variables.
b) La recta de regresión de X sobre Y.
c) La empresa decide invertir el próximo año 150253ε en publicidad. Si se mantiene la misma tendencia de los años anteriores, ¿cuál es el volumen de ventas esperado?

7.- Se ha solicitado a un grupo de 50 individuos información sobre el número de horas que dedica diariamente a dormir y a ver la televisión. La clasificación de las respuestas ha permitido elaborar la siguiente tabla:

Nº de horas dormidas X	6	7	8	9	10
Nº de horas televisión Y	4	3	3	2	1
f_i	3	16	20	10	1

a) ¿Existe algún tipo de correlación entre ambas variables? ¿En qué te basas para responder a la pregunta anterior?
b) Si una persona ve la televisión diariamente durante 5 horas, ¿cuánto tiempo cabe esperar que dedica a dormir? Valora la fiabilidad de tu estimación.

8.- Los datos correspondientes al número de incendios forestales registrados en España (X) y el número de hectáreas afectadas (Y) durante 15 años se recogen en la siguiente tabla:

X(miles) / Y (miles)	[0,5)	[5,10)	[10,15)	[15,20)
[0,100)	0	0	2	0
[100,200)	1	4	1	1
[200,300)	0	2	1	0
[300,400)	0	0	0	1
[400,500)	0	0	1	1

a) Haz la distribución marginal de ambas variables ¿Cuál de ellas presenta mayor dispersión?
b) Dibuja la nube de puntos y, si existe correlación lineal halla el coeficiente de Pearson y la recta de regresión lineal de Y sobre X.
c) ¿Cuántas hectáreas cabe esperar que se quemen un año en el que se produzcan 12500 incendios? Valora la predicción efectuada.

9.- Si la pendiente de una recta de regresión es negativa, entonces, necesariamente:

a) la correlación es débil. c) La correlación es directa (positiva)
b) La correlación es inversa (negativa d) La correlación es muy fuerte.

10.- ¿Qué tipo de correlación existe en cada caso?

a) r = 1 b) r = 21 c) r = -0,1 d) r = 0,5

11.- Realiza un diagrama de dispersión aproximado para las distribuciones bidimensionales con correlación: a) r = 0,9 b) r = -0,8 c) r = 0,1 d) r = 1

12.- ¿En cuál de las distribuciones anteriores no es bueno el ajuste por una recta?

13.- Se observaron las edades de 5 niños/as y sus pesos respectivos, obteniéndose los siguientes resultados:

Edad, en años (X)	2	4,5	6	7,2	8
Peso, en Kg. (Y)	15	19	25	33	34

a) Halla el coeficiente de correlación lineal y las rectas de regresión de Y sobre X y de X sobre Y.
b) ¿Qué peso correspondería a un niño/a de 5 años? ¿Qué edad correspondería a un peso de 36 Kg?

14.- Una compañía discográfica ha recopilado la siguiente información sobre el número de conciertos dados, durante el verano, por 15 grupos musicales y las ventas de discos de estos grupos (expresados en miles de CDs):

Conciertos (y) CDs (x)	[10, 30)	[30, 50)	[50,70)
[1, 5)	3	0	0
[5,10)	1	4	1
[10, 20)	0	1	5

a) Representa el diagrama de dispersión e indica qué tipo de relación existe entre ambas variables.
b) Haz las distribuciones marginales y calcula la media y desviación típica de X e Y.

15.- En una muestra de 64 familias se estudió el número de miembros en edad laboral, x, y el número de ellos que están en activo, y, .Los resultados son los de la siguiente tabla:

X Y	1	2	3
1	6	0	0
2	10	2	0
3	12	5	1
4	16	8	4

a) Representa el diagrama de dispersión e indica qué tipo de relación existe entre ambas variables.
b) Haz las distribuciones marginales y calcula la media y desviación típica de X e Y.