6 – El modelo de regresión lineal

• Modelos predictivos o de regresión: la representación de la relación entre dos (o más) variables a través de un modelo formal supone contar con una expresión lógico-matemática que, aparte de resumir cómo es esa relación, va a permitir realizar predicciones de los valores que tomará una de las dos variables, la que se asuma como variable de respuesta, a partir de los valores de la que se asuma como variable explicativa.

• En lo que respecta al papel que juegan las variables en el modelo, mientras que en el análisis de la relación entre dos variables no se asumía un rol específico para las mismas (**rol simétrico** de las variables –era lo mismo la relación de la variable *A* con la variable *B*, que la de *B* con *A*), la aplicación de un modelo de regresión supone que una de las 2 variables adopta el papel de variable explicativa y la otra el de variable de respuesta y es, por tanto, que se dice que las variables adoptan un **rol asimétrico** –no es el mismo el modelo de regresión de *B* sobre *A*, que el de *A* sobre *B*.

• En la literatura estadística se han planteado diferentes tipos de modelos predictivos que han dado respuesta a las distintas características de las variables que pueden aparecer implicadas en el mismo, ya sea su escala de medida, la forma de su distribución... El más conocido es el modelo de regresión lineal (variable de respuesta cuantitativa), si bien, otras opciones a tener en cuenta son el modelo de

regresión logística (variable de respuesta categórica) o el modelo de Poisson (variable de respuesta cuantitativa con distribución muy asimétrica), entre otros.

1. Conceptos básicos sobre el análisis de regresión lineal

• El modelo de regresión lineal es el más utilizado a la hora de predecir los valores de una variable cuantitativa a partir de los valores de otra variable explicativa también cuantitativa (modelo de regresión lineal simple). Una generalización de este modelo, el de regresión lineal múltiple, permite considerar más de una variable explicativa cuantitativa en el modelo.

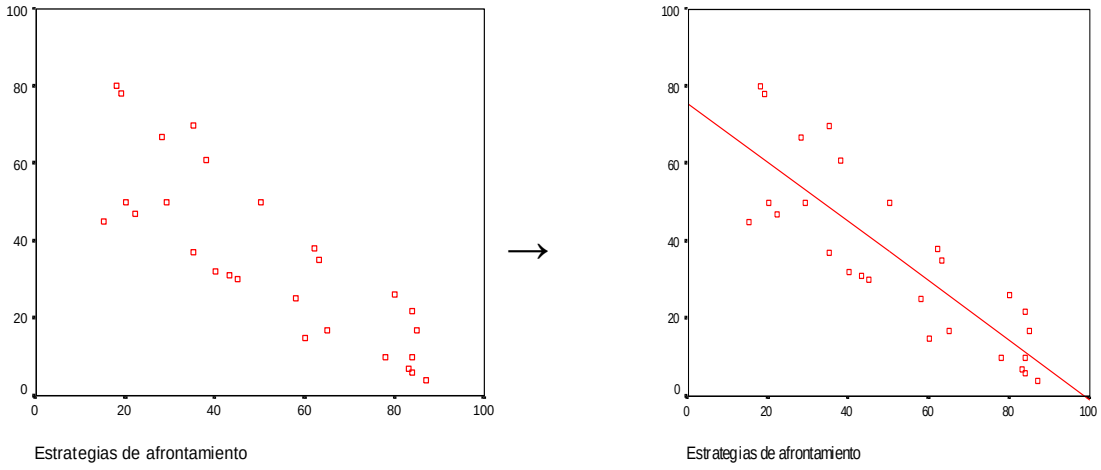
• En concreto, según el modelo de regresión lineal simple, la distribución conjunta de dos variables, una de ellas considerada como variable explicativa (X) y la otra como variable de respuesta (Y), cuya relación sea más o menos lineal (ver diagrama de dispersión) puede ser representada (modelada) por la ecuación de una línea recta:

$$\hat{Y} = B_0 + B_1 \cdot X$$

Ejemplo de aplicación de un modelo de regresión lineal simple a fin de modelar la distribución conjunta de las variables "Estrategias de afrontamiento" y "Estrés" (ver el correspondiente diagrama de dispersión en el gráfico de abajo a la izquierda). En este ejemplo concreto, el modelo de regresión lineal simple se concreta en el ajuste a los datos de la siguiente ecuación (también conocida como recta de regresión):

$$\hat{Y} = 75,4 + (-0,76) \cdot X$$

El cómo han sido obtenido los coeficientes de dicha ecuación (B_0 y B_1) será tratado más adelante. En el gráfico de abajo a la derecha se ha dibujado la recta correspondiente a la citada ecuación.



Una importante ventaja de contar con el modelo de regresión de "Estrés" sobre "Estrategias de afrontamiento" es que a partir del mismo podemos predecir cuál será la puntuación en "Estrés" a

partir de un valor cualquiera de "Estrategias de afrontamiento", por ejemplo, para una puntuación de 50, la puntuación predicha de "Estrés" será de 37,4 ($=75,4-0,76\cdot 50$).

• Los dos coeficientes de la ecuación del modelo de regresión lineal simple, B_0 y B_1, son conocidos como la constante y la pendiente del modelo, respectivamente. En conjunto reciben el nombre de coeficientes de la ecuación de regresión. Si la ecuación de la recta de regresión es obtenida a partir de una muestra, que no una población, los coeficientes de la ecuación de regresión que obtengamos serán estadísticos, no parámetros, y la ecuación se expresa simbólicamente como:

$$\hat{Y} = B_0 + B_1 \cdot X_1$$

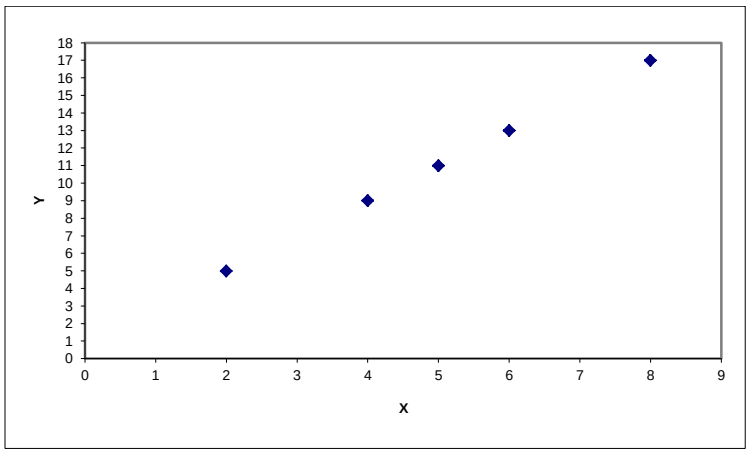
Señalar también que la ecuación de la recta de regresión aparece expresada en algunos libros de texto así:

$$\hat{Y} = A + B \cdot X$$

• Una vez que sean conocidos los valores de B_0 y B_1 del modelo de regresión lineal simple, éste puede ser utilizado como modelo predictivo, esto es, para realizar predicciones de los valores que tomará la variable de respuesta para determinados valores de la variable explicativa. Basta para ello con sustituir en la ecuación de regresión el valor concreto de X que se quiera (X_i). Al hacerlo, se obtendrá el valor predicho para Y para aquellos casos que en la variable X tomen el valor X_i. Este valor es conocido de forma genérica como puntuación predicha, siendo representado simbólicamente como Y_i' o $\hat{Y}_i$.

Ejercicio 1: A partir de la distribución conjunta de las variables cuantitativas X e Y, dibuja en el correspondiente diagrama de dispersión la recta de regresión que creas que mejor se ajusta a la nube de puntos. Aunque todavía no se haya visto como obtener los valores de B_0 y B_1 de la recta de regresión, intenta deducir de forma intuitiva cuáles pueden ser esos valores. Si has planteado la ecuación, utilízala para realizar una predicción de cuáles serán los valores predichos en Y para distintos valores de X (por ejemplo, para $X_i = 3$, para $X_i = 6$, para $X_i = 9$…). También puedes realizar esas predicciones utilizando el gráfico inferior a partir de la recta de regresión que has dibujado.

X	Y
2	5
4	9
5	11
6	13
8	17



• Relaciones deterministas vs. probabilísticas y error de predicción: el anterior ejemplo representa el caso de una relación determinista o perfecta entre X e Y –si se calcula R_{XY}, nos dará igual a 1. En consecuencia, los valores predichos $\hat{Y}$ a partir de X según el modelo de regresión coincidirán exactamente con los valores observados en Y, no cometiéndose ningún error de predicción. Sin embargo, esta situación es inusual en el ámbito de las Ciencias Sociales y de la Salud, donde casi siempre nos encontramos con relaciones entre variables no perfectas ($R_{XY} \neq 1$ o $R_{XY} \neq -1$). En estos casos, cuando se utiliza la ecuación de la recta de regresión para predecir cuál será el valor en Y para un determinado valor X_i, es más que probable que se cometa cierto error en la predicción realizada. A este error se le conoce como error de predicción o residual (E_i) y queda definido, por tanto, como la diferencia entre el verdadero valor de un sujeto en la variable Y (Y_i) y su valor predicho según la ecuación de regresión ($\hat{Y}_i$):

$$E_i = Y_i - \hat{Y}_i$$

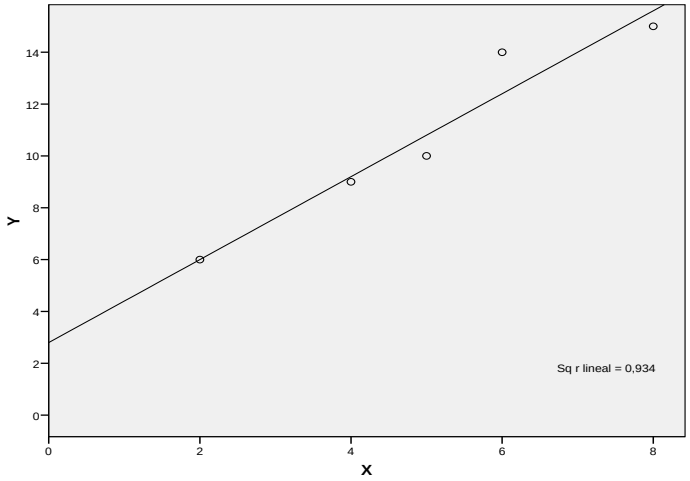
De la expresión anterior se deriva que la puntuación observada de un sujeto en Y se puede obtener sumando a la puntuación predicha, el error de predicción o residual para dicha puntuación, esto es:

$$Y_i = \hat{Y}_i + E_i$$

Ejemplo de los conceptos presentados para dos variables X e Y (N = 5), siendo el modelo de regresión lineal ajustado a la distribución conjunta de ambas variables, el siguiente:

$$\hat{Y} = 2{,}8 + 1{,}6 \cdot X$$

ID	*X*	*Y*
1	2	6
2	4	9
3	5	10
4	6	14
5	8	15

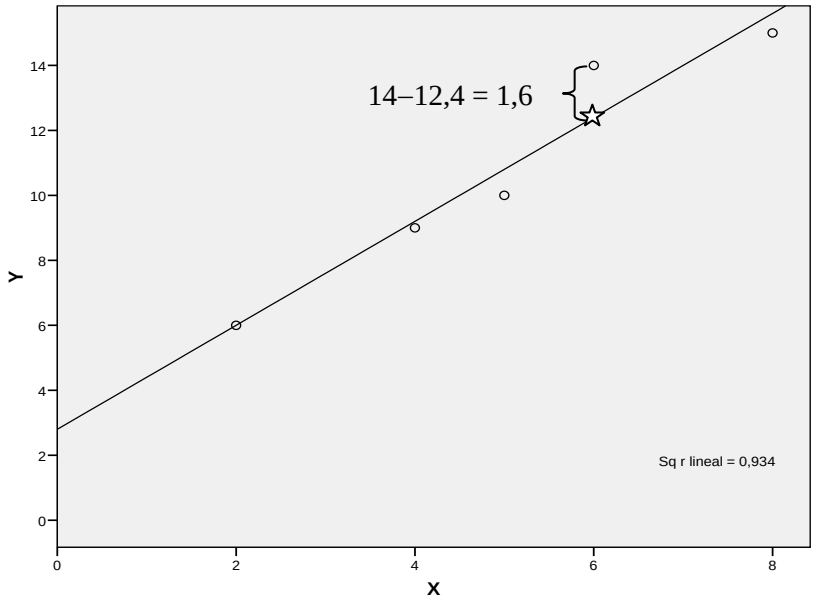


Utilizando la ecuación de regresión ajustada a los datos, ¿qué error cometemos al predecir *Y* a partir de *X* para cada uno de los 5 casos? Por ejemplo, para el cuarto caso en la tabla (X_4 = 6), el valor predicho es 12,4 ($\hat{Y}_4$ = 2,8 + 1,6·6 = 12,4) y, en consecuencia, su error de predicción o residual es 1,6 (E_4 = 14 – 12,4), esto es, la diferencia entre su verdadero valor en la variable *Y* (Y_4 = 14) y su valor predicho según la ecuación de regresión ($\hat{Y}_4$ = 12,4). Del mismo modo, para el resto de casos:

ID	*X*	*Y*	$\hat{Y}$	*E*
1	2	6	6,0	0
2	4	9	9,2	-0,2
3	5	10	10,8	-0,8
4	6	14	12,4	1,6
5	8	15	15,6	-0,6

Adelantar ya que la columna de los errores de predicción constituye un elemento de información clave a la hora de tratar el concepto de bondad de ajuste de un modelo de regresión, algo que se abordará en una sección posterior.

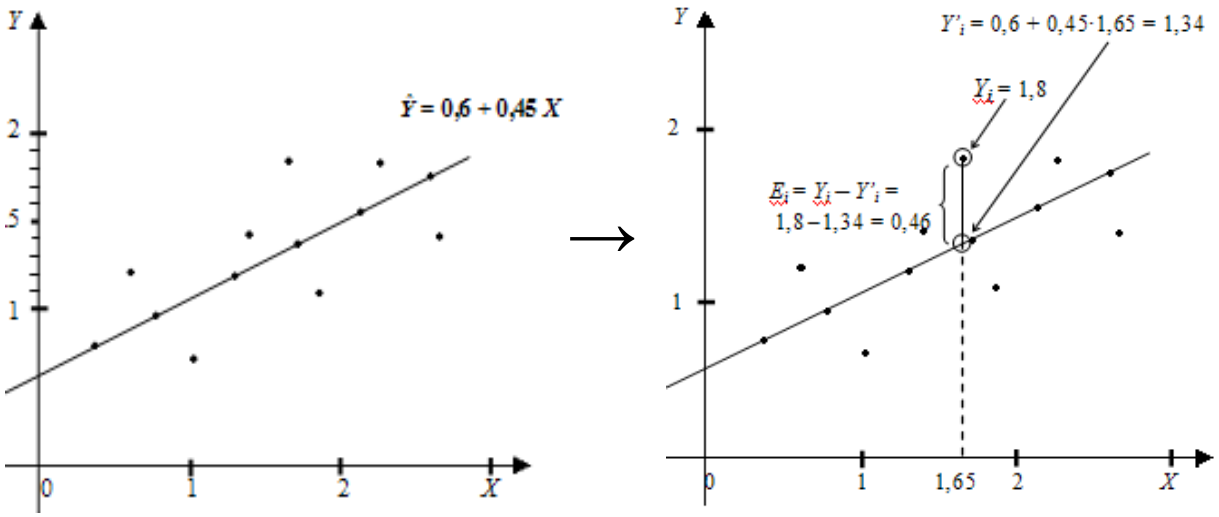
• Gráficamente, el residual correspondiente a cualquier caso (puntos en el diagrama de dispersión) viene representado por la distancia vertical del punto correspondiente a la recta de regresión, tal como se muestra abajo para el caso 4º del ejemplo anterior.



Otro **ejemplo** (Losilla y cols., 2005) para el caso de dos variables *X* e *Y* cuyo diagrama de dispersión se muestra a continuación (gráfico de la izquierda). El modelo de regresión lineal obtenido para la distribución conjunta de ambas variables viene definido por la siguiente ecuación:

$$\hat{Y} = 0{,}6 + 0{,}45 \cdot X$$

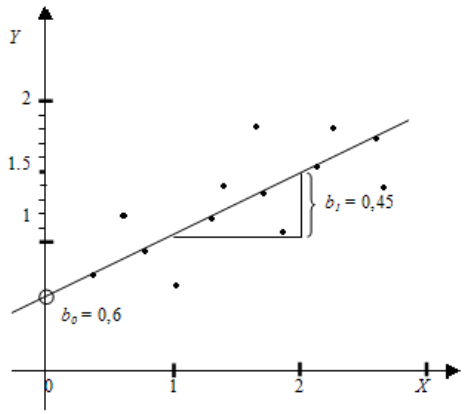
En el gráfico de la derecha se muestra el error de predicción, de acuerdo al modelo de regresión lineal ajustado, para el caso cuya puntuación en *X* y en *Y* es, respectivamente, 1,65 y 1,8. Como puede observarse la puntuación predicha en *Y* para los sujetos que en *X* son 1,65 es igual a 1,34, por lo que el error de predicción resultante es igual a 0,46.



• Interpretación de B_0 y B_1

- La constante de la ecuación de la recta de regresión (B_0) representa el valor predicho en Y cuando la variable X es igual a 0. Se trata de un valor que no entraña especial interés a nivel interpretativo.
- El valor de la pendiente (B_1) representa el cambio esperado en la variable Y por cada unidad de incremento en la variable X. En este sentido, B_1 representa un indicador de la relevancia del efecto que los cambios en X tienen sobre Y. Señalar que, en cuanto que representa el incremento en $\hat{Y}$ por cada incremento de X en una unidad, el valor de la pendiente estará expresado en las mismas unidades que la variable de respuesta Y.

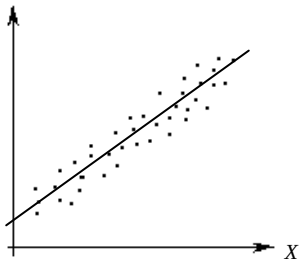
Ejemplo para el caso de 2 variables, X e Y, siendo la ecuación de regresión de Y sobre X la siguiente: $\hat{Y} = 0{,}6 + 0{,}45 \cdot X$. Tal como puede observarse en el diagrama de dispersión, cuando se aumenta en una unidad el valor de X, en el valor de Y predicho por la recta de regresión se produce un aumento de 0,45 unidades.



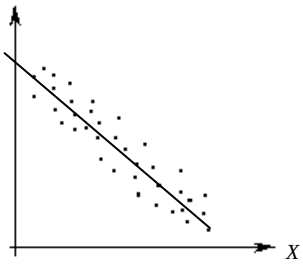
• Valores que puede tomar B_1 : Puede tomar valores tanto positivos como negativos, dependiendo del sentido (directo o inverso) de la relación entre las variables. Su valor, en valor absoluto, será más alto cuanto mayor sea la relación lineal entre las variables.

A continuación se muestran cuatro **ejemplos** que muestran el vínculo directo entre el valor de B_1 y el tipo de relación existente entre las variables. En el ejemplo 1 la relación entre X e Y es directa, por lo que el valor de la pendiente será de signo positivo. En el ejemplo 2 la relación entre X e Y es inversa, por tanto B_1 será menor de 0. Las figuras 3 y 4 evidencian la no existencia de relación lineal entre las variables, en cuyo caso el valor de B_1 será nulo o prácticamente nulo

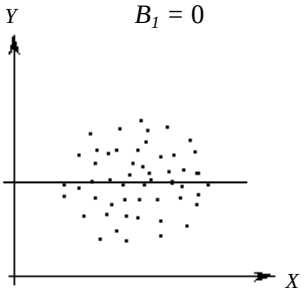
Y $B_1 > 0$ Y $B_1 < 0$



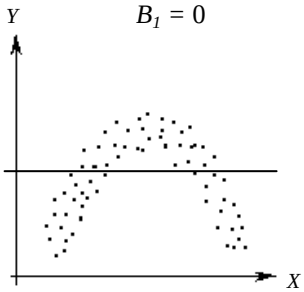
Ejemplo 1. Relación lineal directa.



Ejemplo 2. Relación lineal inversa.



Ejemplo 3. Ausencia de relación.

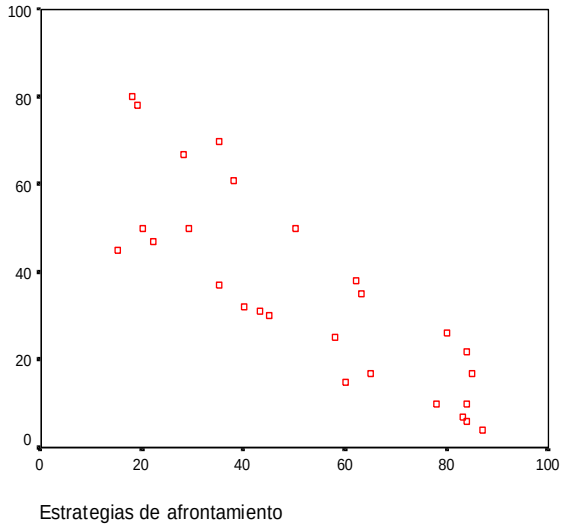


Ejemplo 4. Relación no lineal.

• A continuación se presentan los datos de un estudio cuyo objetivo fue investigar el efecto del nivel de estrategias de afrontamiento (*X*) de los sujetos sobre su nivel de estrés (*Y*). Estos datos van a ser utilizados en los siguientes apartados para ilustrar: (1) cómo obtener el valor de los dos coeficientes del modelo de regresión lineal –lo que se conoce como el *ajuste o identificación del modelo*; (2) cómo utilizar el modelo de regresión obtenido para realizar predicciones en "Estrés" a partir del valor de "Afrontamiento" de los sujetos; y (3) cómo valorar la calidad de dichas predicciones –lo que se conoce como el *análisis de la bondad de ajuste o capacidad predictiva del modelo*.

Los datos correspondientes a este estudio se muestran en la tabla inferior, en concreto, las puntuaciones recogidas a partir de una muestra de 27 sujetos en una escala observacional de "Estrés" y en un test orientado a evaluar la utilización de estrategias de "Afrontamiento". El rango de puntuaciones en ambas variables podía oscilar entre 0 a 100, significando puntuaciones más altas mayor estrés y mayor capacidad de utilización de mecanismos de afrontamiento, respectivamente. El diagrama de dispersión permite visualizar la distribución conjunta de las puntuaciones en ambas variables.

Caso	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Estrés	61	26	32	22	38	80	17	10	47	15	50	25	50	30	78	10	35	31	4	6	7	17	37	45	50	67	70
Afronta	38	80	40	84	62	18	65	78	22	60	50	58	20	45	19	84	63	43	87	84	83	85	35	15	29	28	35



2. Ajuste de la recta de regresión

• La identificación o ajuste de un modelo de regresión supone obtener los coeficientes que caracterizan al mismo. En el caso del modelo de regresión lineal simple, B_0 y B_1.

• Ello supone aplicar un procedimiento de cálculo (método de estimación) que permita, a partir de los datos disponibles, obtener los coeficientes de la ecuación de una línea recta que represente óptimamente la distribución conjunta de las variables modeladas. Ahora bien, ¿cuál es la línea recta que representa óptimamente a la nube de puntos de un diagrama de dispersión?

• En principio, un criterio natural de bondad de ajuste supone considerar la ecuación de regresión que dé lugar a un menor error en las predicciones. Este aspecto se concreta en que la mejor recta sea aquella para la que la suma de los cuadrados de los errores (*SCE*) tenga un valor más próximo a 0. Así, para este método, conocido como método de los mínimos cuadrados ordinarios, la mejor recta de regresión, de entre todas las posibles que se pueden ajustar a la distribución conjunta de 2 variables, será aquélla para la que la *SCE* sea mínima:

$$\textit{Mejor modelo de regresión} \rightarrow \min(SCE) = \min\left(\sum E_i^2\right) = \min\left(\sum (Y_i - \hat{Y}_i)^2\right)$$

• Tras realizar las derivaciones matemáticas pertinentes, sobre las que no se va a entrar aquí, las fórmulas de obtención de los coeficientes del modelo de regresión lineal simple ($\hat{Y} = B_0 + B_1 \cdot X$) que van a satisfacer que la *SCE* sea mínima (criterio del método de los mínimos cuadrados ordinarios) son las siguientes:

$$B_1 = R_{XY} \cdot \frac{S_Y}{S_X} \qquad B_0 = \bar{Y} - B_1 \cdot \bar{X}$$

Como puede observarse, la obtención de B_0 implica haber calculado previamente B_1.

Ejercicio 2:

a) Obtener el valor de los coeficientes B_0 y B_1 para el ejemplo sobre las variables "Afrontamiento" y "Estrés" (ver enunciado más arriba, p. 7-8), teniendo en cuenta la siguiente información sobre estas variables: R_{xy}= –0,847; S_X = 24,80; S_Y = 22,37; $\overline{X}$ = 52,22 e $\overline{Y}$ = 35,56

b) Plantear la ecuación de la recta de regresión.

c) ¿Qué predicción de estrés haríamos para el sujeto nº 8, el cual tiene una puntuación de 78 en la escala de afrontamiento (X_i = 78)? ¿Cuál sería el error de predicción (E_i) para este sujeto?

d) Interpretar los coeficientes de la recta de regresión

e) Dibujar (de forma aproximada) la recta de regresión sobre el diagrama de dispersión de las variables presentado anteriormente.

f) A continuación se muestran los *outputs* obtenidos con el programa SPSS del análisis de regresión para este ejemplo. Identificar en los mismos los resultados obtenidos anteriormente.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	.847[a]	.717	.705	12.14

a. Variables predictoras: (Constante), Estrategias de afrontamiento

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	75.425	5.532		13.634	.000
	Estrategias de afrontamiento	-.763	.096	-.847	-7.951	.000

a. Variable dependiente: Puntuación escala de estrés

3. Bondad de ajuste

• La bondad de ajuste de un modelo de regresión se refiere al grado en que éste es conveniente como modelo que representa a la distribución conjunta de las variables implicadas en el mismo. Tal como hemos visto, al ajustar un modelo de regresión lineal simple a la distribución conjunta de 2 variables obtendremos la mejor recta de regresión de entre todas las posibles que se pueden ajustar a esa distribución, ahora bien, ello no significa que sea buena. Así, puede ocurrir que la distribución

conjunta de 2 variables sea difícil de modelar debido a la inexistencia de relación entre las variables (ver, por ejemplo, el caso de la Figura 1 a continuación), o bien, que el modelo de regresión lineal no sea el más adecuado para ese propósito (ver, por ejemplo, el caso de la Figura 2).

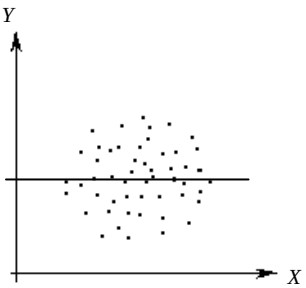


Figura 1: Ausencia de relación.

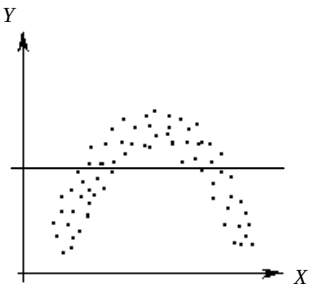
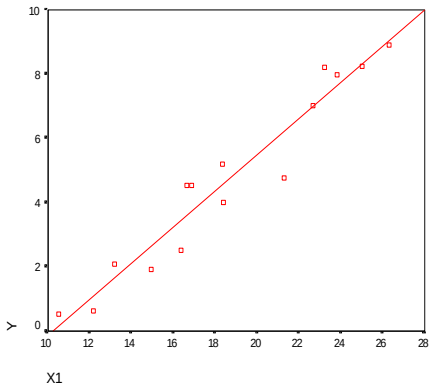
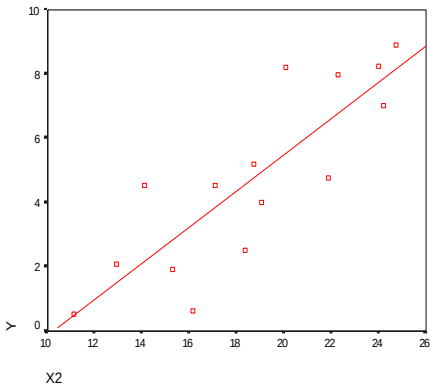


Figura 2: Relación no lineal.

Ejemplo: la relación entre los dos pares de variables (*X1, Y1)* y *(X2, Y2)* que aparece representada en los dos siguientes diagramas de dispersión (Losilla y cls., 2005) es ajustada, *casualmente*, por la misma ecuación de regresión lineal ($Y' = -5,74 + 0,56 \cdot X$). Sin embargo, tal como se puede intuir a nivel visual, la bondad de ajuste de la ecuación de la figura de la izquierda será mejor que la de la figura de la derecha.



Modelo 1: $Y1' = -5,74 + 0,56 \cdot X2$



Modelo 2: $Y2' = -5,74 + 0,56 \cdot X2$

• Existen diferentes aproximaciones en la <u>evaluación de la bondad del ajuste</u> de un modelo a la realidad que ese modelo pretende representar. Una elemental consiste en comparar las puntuaciones predichas por el modelo de regresión ($\hat{Y}_i$) con las puntuaciones reales a partir de las que ha sido estimado (Y_i). El índice más utilizado en esta aproximación es, precisamente, el conocido como la suma de cuadrados de los errores de predicción (o residuales) (*SCE o* $SC_{Y \cdot X}$), el cual ya fue introducido en el apartado anterior como criterio de referencia del método de estimación de mínimos cuadrados ordinarios en la estimación de los coeficientes de la ecuación de regresión:

$$SCE\,(o\,SC_{Y \cdot X}) = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

• La suma de cuadrados de los residuales puede oscilar entre 0 y cualquier valor positivo. Si este sumatorio da 0, el modelo de regresión se ajusta perfectamente a los datos; cuanto mayor sea su valor, ello significará que más erróneas son las predicciones de la ecuación de regresión y, por lo tanto, peor su bondad como modelo predictivo. Consecuencia de esta ausencia de un techo numérico, este índice puede resultar difícil de interpretar en la práctica.

• Dos índices derivados del anterior:

(1) el que se obtiene como media aritmética del cuadrado de los errores de predicción, esto es, el resultado de dividir la *SCE* por *n*, el cual se denomina como varianza de los errores ($S^2_{Y \cdot X}$).

$$S^2_{Y \cdot X} = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n}$$

(2) la raíz cuadrada del anterior, esto es, la desviación típica de los errores de predicción ($S_{Y \cdot X}$), más conocido como el error típico de estimación:

$$S_{Y \cdot X} = \sqrt{S^2_{Y \cdot X}} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n}}$$

Ambos índices adolecen del mismo problema de interpretación que *SCE*.

• Otro índice que supera el problema interpretativo de los dos anteriores ha sido propuesto tras tomar como punto de referencia una relación básica que se da cuando se ajusta un modelo de regresión lineal a 2 (o más) variables. Es la que se conoce como igualdad de la descomposición de la varianza de *Y*, la cual se deriva del axioma que establece que la puntuación observada en la variable de respuesta es igual a la predicha según el modelo de regresión más el error de predicción cometido: $Y_i = \hat{Y}_i + E_i$. A partir de la anterior igualdad se puede derivar algebraicamente la siguiente: $SC_Y = SC_{Y'} + SC_{Y \cdot X}$, o lo que es lo mismo:

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (Y_i' - \bar{Y})^2 + \sum_{i=1}^{n} (Y_i - Y_i')^2$$

Si cada uno de los términos de la expresión anterior lo dividimos por *n*, tendremos la misma igualdad expresada en forma de varianzas:

$$S^2_Y = S^2_{Y'} + S^2_{Y \cdot X}$$

Así, la varianza en las puntuaciones de la variable de respuesta (*Y*) es igual a la varianza explicada por el modelo de regresión (varianza de las puntuaciones predichas) más la varianza no explicada por el modelo de regresión (varianza de los errores o residuales).

• Consecuencia de la igualdad de descomposición de la varianzas, se puede plantear de forma inmediata un índice de bondad de ajuste del modelo de regresión como razón de la varianza explicada por el modelo de regresión ($S^2_{Y'}$) respecto a la varianza total (S^2_Y) $\rightarrow$ $S^2_{Y'}/S^2_Y$.

Cuanto mayor sea la varianza explicada respecto a la varianza total, más próximo a 1 será el valor de este índice, poniendo de manifiesto una mayor capacidad predictiva de la variable explicativa respecto a la variable de respuesta. Este índice, conocido en la literatura estadística como coeficiente de determinación, se expresa simbólicamente como R^2_{XY} , pues también puede ser obtenido elevando al cuadrado el coeficiente de correlación de Pearson entre la variable predictora y la variable de respuesta. Por último, R^2_{XY} puede también obtenerse como razón de la suma de cuadrados explicada por el modelo de regresión ($SC_{Y'}$) y la suma de cuadrados de la variable de respuesta (SC_Y):

$$R^2_{XY} = \frac{S^2_{Y'}}{S^2_Y} = \quad o \quad R^2_{XY} = \frac{SC_{Y'}}{SC_Y}$$

• El coeficiente de determinación (R^2_{XY}) representa la proporción de varianza de *Y* explicada por las variables implicadas en el modelo de regresión ajustado a los datos (*X* en el modelo de regresión lineal simple). En cuanto que una razón, este coeficiente oscilará siempre entre 0 y 1, de modo que cuanto más próximo sea R^2_{XY} a 1, indicará mejor bondad de ajuste del modelo de regresión a la distribución conjunta de las variables. Si R^2_{XY} es igual a 1, el ajuste será perfecto.

• Otro propuesta de índice de bondad de ajuste complementaria a la anterior, aunque mucho menos utilizada en la práctica, es el conocido como coeficiente de alienación, el cual también oscila entre 0 y 1, si bien, en este caso valores próximos a 1 indican peor bondad de ajuste del modelo a los datos.

$$CALN = \frac{S^2_{Y \cdot X}}{S^2_Y} \quad o \quad CALN = \frac{SC_{Y \cdot X}}{SC_Y}$$

Obviamente, $CALN = 1 - R^2_{XY}$

• El siguiente cuadro resume todos los conceptos vistos en este apartado:

Varianza de Y = Varianza de Y explicada por X + Varianza de Y **no** explicada por X
(varianza de las puntuaciones predichas) (varianza de los errores o residuales)

$$\mathbf{S_Y^2} = \mathbf{S_{\hat{Y}}^2} + \mathbf{S_{Y\times X}^2}$$

$$\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{n} = \frac{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}{n} + \frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{n}$$

$\frac{\mathbf{S_{\hat{Y}}^2}}{\mathbf{S_Y^2}}$ = **Proporción de la varianza de Y que es explicada por X = Coeficiente de determinación = $\mathbf{R_{XY}^2}$**

$\frac{\mathbf{S_{Y\square X}^2}}{\mathbf{S_Y^2}}$ = **Proporción de la varianza de Y que NO es explicada por X = Coeficiente de alienación = 1 - $\mathbf{R_{XY}^2}$**

Ejemplo de cálculo de las varianzas asociadas a la igualdad de descomposición de la varianza en un análisis de regresión de una variable *Y* sobre una variable *X*:

X	*Y*
4	2
8	11
11	9
2	3
5	10

Sabiendo que: $\bar{X} = 6$; $S_X = 3{,}16$; $\bar{Y} = 7$; $S_Y = 3{,}74$; $R_{XY} = 0{,}69$

y que la ecuación de la recta de *Y* sobre *X*: $\hat{Y} = 2{,}08 + 0{,}82 \cdot X$

veamos cómo se obtienen los valores predichos ($\hat{Y}_i$) y los residuales (E_i) para cada caso:

X	*Y*	$\hat{Y}$	E_i	E_i^2	$(\hat{Y}_i-\bar{Y})^2$
4	2	5,36	-3,36	11,29	2,69
8	11	8,64	2,36	5,57	2,69
11	9	11,1	-2,1	4,41	16,81
2	3	3,72	-0,72	0,52	10,76
5	10	6,18	3,82	14,59	0,67

$S_{Y\cdot X}^2 = 36{,}4/5 = 7{,}28$ $S_{\hat{Y}}^2 = 33{,}62/5 = 6{,}72$

A partir de los valores predichos se puede obtener:

- La varianza de los errores (o residuales) → $S_{Y\cdot X}^2 = 7{,}28$
- La varianza de las puntuaciones predichas → $S_{Y'}^2 = 6{,}72$

Descomposición de la varianza de *Y*:

$S_Y^2 = 3{,}74^2 = 14$

$14 = 6{,}72 + 7{,}28$

$\downarrow \quad \downarrow \quad \downarrow$

$S_Y^2 = S_{Y'}^2 + S_{Y \cdot X}^2$

Coeficiente de determinación (proporción de la varianza de *Y* explicada por *X*):

$R^2 = 6{,}72/14 = 0{,}48 \quad (= 0{,}69^2)$

Coeficiente de alienación (proporción de la varianza de *Y* no explicada por *X*):

CALN = 7,28/14 = 0,52 (= 1– 0,48)

Ejercicio 3: Dadas dos variables *Q* y *T* de las que se sabe que la varianza de *T* es 10 ($S_T^2 = 10$) y que en el análisis de regresión de *T* sobre *Q*, la varianza de los errores es 8 ($S_{T \cdot Q}^2 = 8$). A partir de esta información, intenta obtener el coeficiente de correlación de Pearson entre *Q* y *T*.

Ejercicio 4: En una muestra de 10 alumnos de enseñanza secundaria se han medido dos variables: (1) rendimiento en el curso, cuantificado como el promedio de las calificaciones de las asignaturas del curso (*Y*); (2) el promedio de horas de estudio semanal durante el curso, obtenido a partir de auto-informe de los propios estudiantes (*X*). Los datos recogidos (*N* = 10) son los que se muestran a continuación:

X	*Y*
5	3
12	6
7	4
9	5
15	9
10	6
12	6
8	5
18	9
14	7

Sabiendo que $\bar{X} = 11$, $\bar{Y} = 6$, S_X = 3,77, S_Y = 1,84 y que R_{XY} = 0,964, obtener: (a) la ecuación del modelo de regresión lineal de *Y* sobre *X*; (b) los valores predichos por la ecuación de regresión para cada sujeto ($\hat{Y}_i$); (c) los errores de predicción o residuales para cada sujeto (E_i); (d) la varianza de los errores ($S_{Y \cdot X}^2$); (e) la varianza de *Y* (S_Y^2); (f) la varianza de las puntuaciones predichas ($S_{\hat{Y}}^2$); (g) comprobar que es cierta la igualdad de la descomposición de la varianza ($S_Y^2 = S_{\hat{Y}}^2 + S_{Y \cdot X}^2$); (h) el coeficiente de determinación [de dos formas: (h.1) a partir de las varianzas; (h.2) a partir del coeficiente de correlación entre *X* e *Y*]; (i) interpretar los coeficientes de la recta de regresión obtenidos (B_0 y B_1); (j) estimar cuál será la puntuación media obtenida a final de curso por un estudiante que dedique 16 horas de estudio a la semana de promedio.

Ejercicio 5: A continuación se muestran el *output* del análisis de regresión realizado con *SPSS* para los datos del ejercicio anterior. Identifica en los mismos los resultados obtenidos anteriormente.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	.964(a)	.930	.921	.546

ANOVA

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	31.613	1	31.613	105.935	.000(a)
	Residual	2.387	8	.298		
	Total	34.000	9			

Coeficientes

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	T	Sig.	Intervalo de confianza para *B* al 95%	
		B	Error típ.	Beta			Límite inferior	Límite superior
1	(Constante)	.810	.533		1.519	.167	-.419	2.039
	Horas_estudio	.472	.046	.964	10.292	.000	.366	.578

Ejercicio 6: En un ejemplo introducido anteriormente con las variables de "Afrontamiento" y "Estrés", sabemos que $R_{XY} = -0{,}847$ y que $S_Y = 22{,}37$. Preguntas: (a) ¿cuál será el valor del coeficiente de determinación?, ¿cómo se interpreta dicho valor?; (b) ¿cuál es el valor de la varianza de *Y* explicada por el modelo de regresión (en este caso, por la variable "Afrontamiento")?; (c) ¿cuál el de la varianza de los residuales?

4. La regresión lineal múltiple

• Una generalización del modelo de regresión lineal simple es el de regresión lineal múltiple, que permite considerar más de una variable explicativa –en principio, cuantitativas– en el modelo de regresión. La formulación del modelo es:

$$\hat{Y} = B_0 + B_1 \cdot X_1 + B_2 \cdot X_2 + ... + B_j \cdot X_j$$

• La importancia relativa de cada variable explicativa es evaluada a través de su coeficiente b_j correspondiente. Cabe señalar que para que dichos coeficientes sean comparables, ya que su magnitud depende de la unidad de medida de la variable *X* a la que multiplican, se ha de comparar su valor estandarizado (los denominados coeficientes tipificados o Beta).

Ejemplo de aplicación de un modelo de regresión lineal múltiple para explicar la "Satisfacción con la carrera" en una muestra de estudiantes universitarios.

$$SatisfaccionCarrera' = B_0 + B_1 \cdot HorasEstudio + B_2 \cdot RelaccionProfesores + B_3 \cdot NotaAcceso$$

A continuación se muestran los resultados obtenidos con SPSS al ajustar el modelo de regresión anterior a un conjunto de datos de estas cuatro variables: ¿Qué porcentaje de la varianza de "Satisfacción con la carrera" es explicada a partir de este modelo? ¿Cuál de las variables explicativas es más relevante?

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,362[a]	,131	,116	1,385

a. Variables predictoras: (Constante), Nota media de acceso, Horas de estudio, Relación con profesores

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.	Intervalo de confianza de 95,0% para B	
		B	Error típ.	Beta			Límite inferior	Límite superior
1	(Constante)	6,386	1,305		4,894	,000	3,809	8,963
	Horas de estudio	,031	,017	,135	1,823	,070	-,003	,066
	Relación con profesores	,238	,058	,303	4,090	,000	,123	,353
	Nota media de acceso	-,186	,193	-,070	-,960	,339	-,568	,196

a. Variable dependiente: Satisfacción con la carrera

5. Descripción estadística de la relación entre dos variables: tabla resumen

• A lo largo de los temas 7, 8 y 9 se han presentado diversos procedimientos estadísticos, tanto numéricos como gráficos, adecuados para describir la relación entre dos variables. En la tabla siguiente se resume esta información en función de la escala de medida de las variables implicadas.

	Gráficos	Índices numéricos
Categórica-categórica	Gráfico de barras agrupado Gráfico de barras apiladas agrupado	*Ji-cuadrado* de Pearson *Phi* de Pearson *V* de Cramer
Categórica-cuantitativa	Gráfico de caja y bigotes agrupado Polígono de frecuencias agrupado Panel de histogramas Gráfico de medias	Diferencia de medias *d* de Cohen *f* de Cohen
Cuantitativa-cuantitativa	Diagrama de dispersión	Covarianza Coef. de correlación de Pearson Coeficiente de determinación Ecuación de regresión lineal Coef. de correlación de Spearman (si relación no lineal)

Referencias

Losilla, J. M., Navarro, B., Palmer, A., Rodrigo, M. F. y Ato, M. (2005). *Del contraste de hipótesis al modelado estadístico.* Documenta Universitaria. [www.edicionsapeticio.com]