T. 9 – El modelo de regresión lineal

1. Conceptos básicos sobre el análisis de regresión lineal
2. Ajuste de la recta de regresión
3. Bondad de ajuste del modelo de regresión

• 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**, dependiente, criterio o *Y*) a partir de los valores de la otra (la que se asuma como **variable explicativa**, independiente, predictora o *X*).

• 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 variables implicadas (**rol simétrico** de las variables), la aplicación de un modelo predictivo 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**.

• En la literatura estadística se han planteado diferentes tipos de modelos predictivos que han dado respuesta a las características (escala de medida, distribución...) de las variables que pueden aparecer implicadas en un determinado modelo. 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. Por otra parte, tal como se verá en un tema posterior, es también posible incluir variables explicativas categóricas en un modelo de regresión lineal si se sigue una determinada estrategia en la codificación de los datos conocida como codificación ficticia.

• En concreto, según el modelo de regresión lineal simple, las puntuaciones de los sujetos en 2 variables -una de ellas considerada como variable predictora (X) y la otra como variable de respuesta (Y)- vienen representadas (modeladas) por la ecuación de una línea recta:

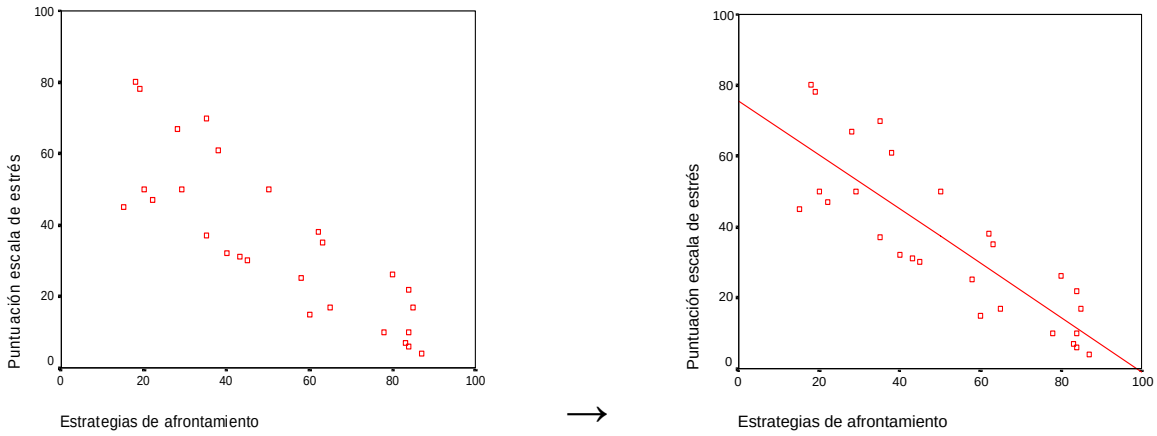
$$\hat{Y} = \beta_0 + \beta_1 \cdot X_1$$

Cuando hay más de una variable explicativa (modelo de regresión lineal múltiple), se utiliza un subíndice para cada una de ellas, por ejemplo, para el caso de dos variables explicativas:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2$$

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". En este ejemplo concreto, el modelo de regresión se concreta en el ajuste a los datos de la siguiente ecuación de regresión (también conocida como recta de regresión):

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



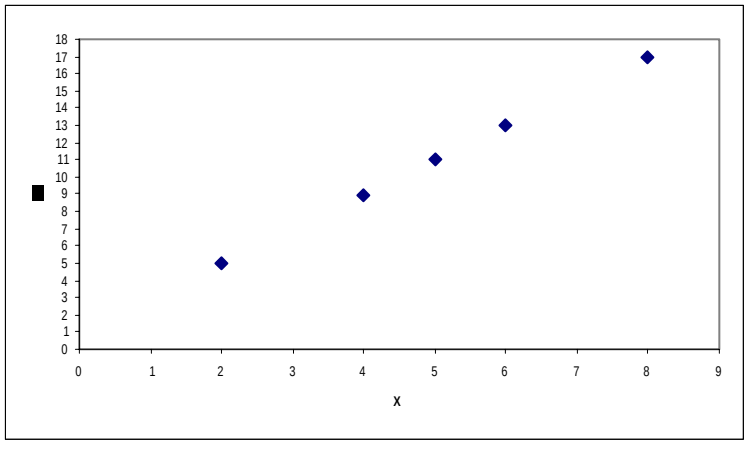
• Los dos parámetros de la ecuación de regresión lineal simple, β_0 y β_1, son conocidos como el origen (también, constante) y la pendiente del modelo, respectivamente. En conjunto reciben el nombre de coeficientes de la ecuación de regresión. Si la ecuación de la recta de regresión es obtenida a partir de una muestra, y no de una población (esto es, los coeficientes de la ecuación de regresión son estadísticos, y no parámetros), la ecuación se expresa como:

$$\hat{Y} = b_0 + b_1 \cdot X_1$$

• Una vez que sean conocidos los valores de β_0 y β_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 según la ecuación de regresión 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 y el correspondiente diagrama de dispersión, dibuja la recta de regresión que mejor se ajuste a la nube de puntos. ¿Cuál será la ecuación de la recta de regresión dibujada?, ¿cuáles serán, por tanto, los valores de β_0 y β_1? Obtener los valores predichos en Y para distintos valores de X (por ejemplo, para X = 3, para X = 6, para X = 9…).

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 (perfecta) entre X e Y, donde $r_{XY} = 1$, 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 -1). En estos casos, cuando se utiliza la recta de regresión para predecir el valor en Y a partir del valor en X de un determinado sujeto (X_i), es probable que se cometa un error en la predicción realizada. A este error se le suele denominar 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 ($\hat{Y}_i$):

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

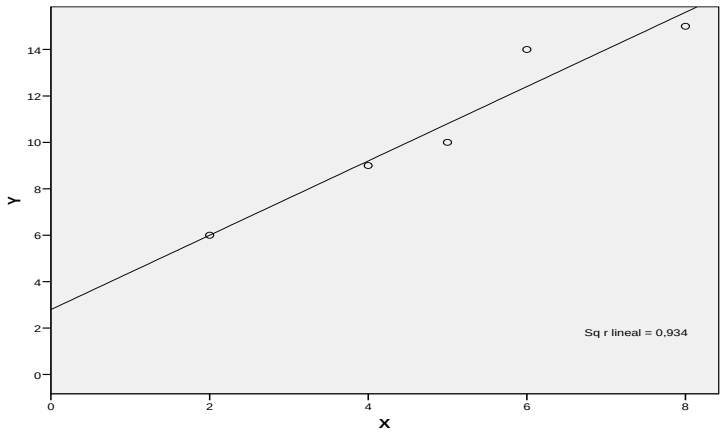
De la expresión anterior se deriva que la puntuación observada de un sujeto en Y se puede obtener sumando a la puntuación predicha el error de predicción o residual para dicha puntuación, esto es:

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

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:

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

X	Y
2	6
4	9
5	10
6	14
8	15



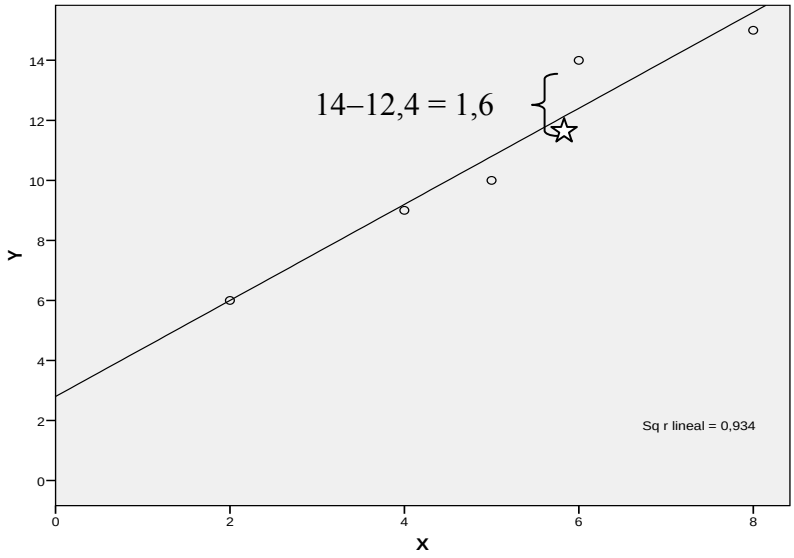
Utilizando la ecuación de regresión ajustada a los datos, ¿qué error cometemos al predecir Y a partir de X para cada uno de los 5 casos?

Por ejemplo, para el cuarto sujeto en la tabla ($X_4 = 6$), el valor predicho es 12,4 ($\hat{Y}_4 = 2{,}8+1{,}6\cdot 6 = 12{,}4$) y, en consecuencia, su error de predicción o residual es 1,6 ($E_4 = 14-12{,}4$). Del mismo modo, para el resto de casos:

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

Adelantar 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 del modelo de regresión, algo que se abordará en una sección posterior.

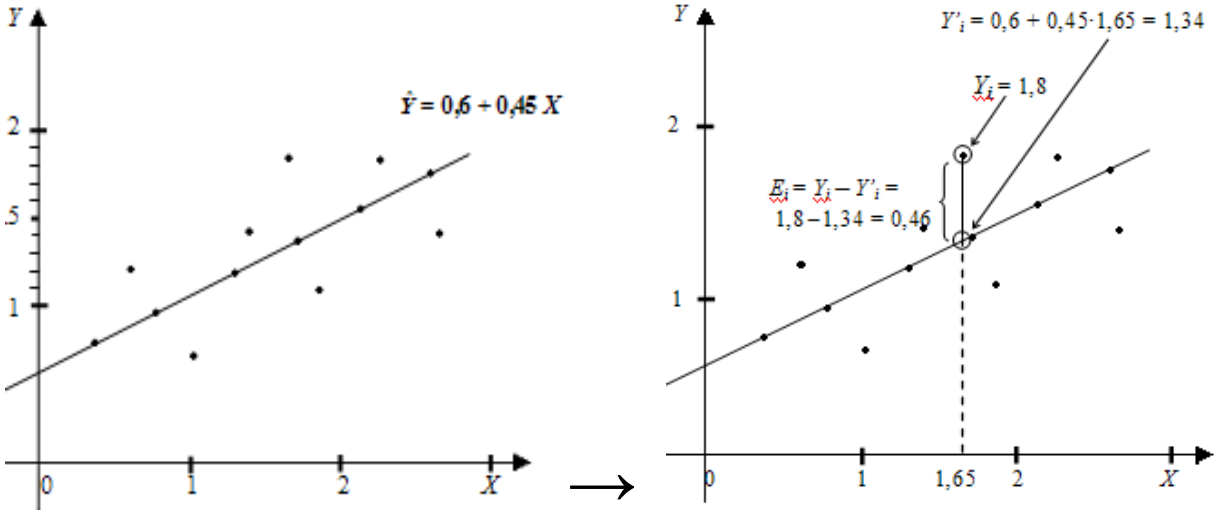
Gráficamente, el residual correspondiente a cualquier punto del diagrama de dispersión viene representado por su distancia vertical a la recta de regresión, tal como se muestra abajo para el caso 4º de la muestra.



Otro **ejemplo** (Losilla y cols., 2005) para el caso de las variables X e Y cuyo diagrama de dispersión se muestra a continuación, siendo la correspondiente ecuación de regresión:

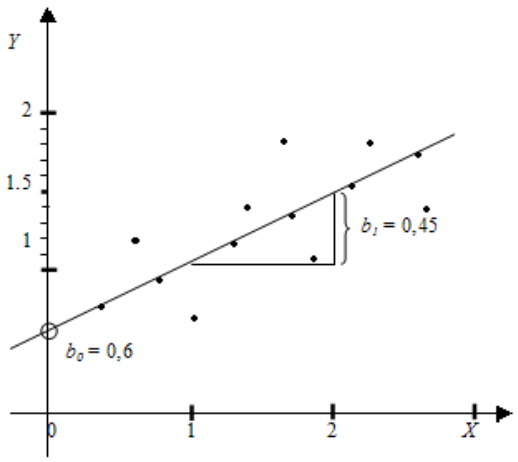
$$\hat{Y} = 0{,}6 + 0{,}45\cdot X$$

A la derecha se muestra el error de predicción, según el modelo de regresión ajustado, para el sujeto cuya puntuación en *X* y en *Y* es, respectivamente, 1,65 y 1,8.



• Interpretación de β_0 y β_1: El origen (o constante) de la ecuación de la recta de regresión (β_0) representa el valor predicho en *Y* cuando la variable *X* es igual a 0; por su parte, más interesante resulta el valor de la pendiente (β_1), el cual representa la inclinación de la recta de regresión respecto al eje de abscisas, más concretamente, cuánto cambio se produce en $\hat{Y}$ por cada unidad de incremento en *X*. En este sentido, β_1 representa un indicador de la relevancia del efecto que los cambios en *X* tienen sobre *Y*.

Ejemplo para el caso de 2 variables *X* e *Y*, siendo la ecuación de regresión: $\hat{Y} = 0{,}6 + 0{,}45 \cdot X$



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*.

• Valores que puede tomar β_1 : Puede tomar valores tanto positivos como negativos, siendo mayores en valor absoluto cuanto mayor sea la pendiente de la recta de regresión. Sería igual a 0 si la recta de regresión fuese horizontal. A continuación se muestran 4 ejemplos que muestran el vínculo directo entre el valor de β y el tipo de relación existente entre las variables:

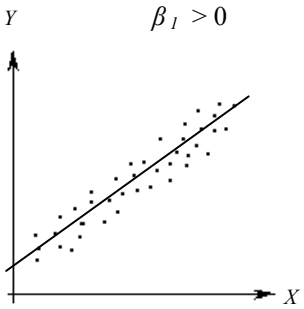


Figura A. Relación lineal positiva (directa).

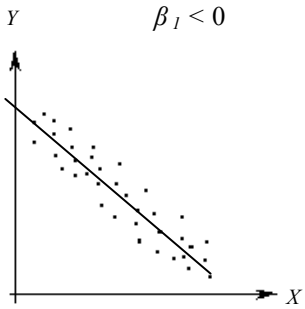


Figura B. Relación lineal negativa (inversa).

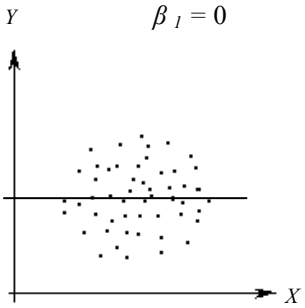


Figura C. Ausencia de relación.

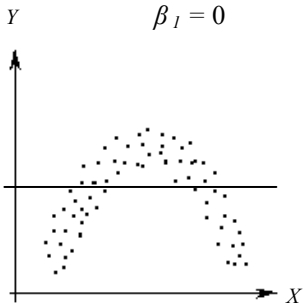


Figura D. Relación no lineal: curvilínea.

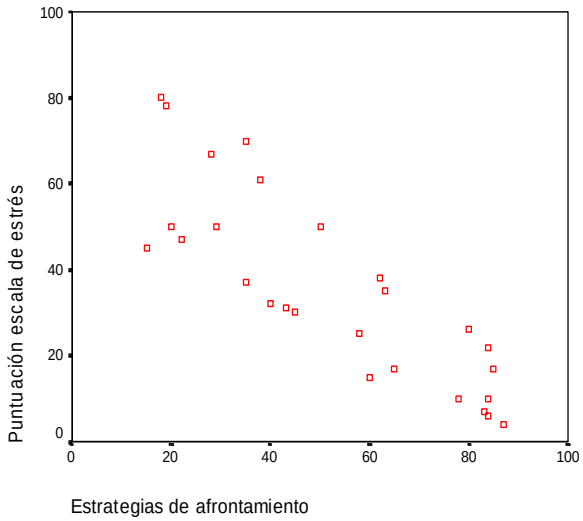
En la figura A la relación entre X e Y es positiva ($\beta_1>0$), lo cual indica que cada incremento de una unidad en X producirá un incremento en $\hat{Y}$ igual al valor de la pendiente. En la figura B la relación es inversa ($\beta_1<0$), por tanto, cada incremento de una unidad en X producirá un decremento en $\hat{Y}$ igual al valor de la pendiente. En la figura C y la figura D, $\beta_1=0$ y, por tanto, la recta de regresión es paralela al eje de abscisas, poniendo de manifiesto que no existe relación lineal entre X e Y.

Ejemplo: A continuación se presentan los datos de un estudio cuyo objetivo fue investigar el efecto de las estrategias de afrontamiento (X) de los sujetos sobre su nivel de estrés (Y). En los siguientes apartados veremos cómo obtener el valor de los dos coeficientes del modelo de regresión lineal (lo que se conoce como el ajuste o identificación del modelo), cómo utilizarlo

para realizar predicciones en "Estrés" a partir del valor de "Afrontamiento" de los sujetos, y cómo valorar la calidad de dichas predicciones (lo que se conoce como el análisis de la bondad de ajuste o capacidad predictiva del modelo).

En la tabla inferior se muestran 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 mecanismos de afrontamiento. El rango de puntuaciones en ambas variables puede oscilar entre 0 a 100, significando puntuaciones más altas mayor estrés y mayor capacidad de utilización de mecanismos de afrontamiento, respectivamente.

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

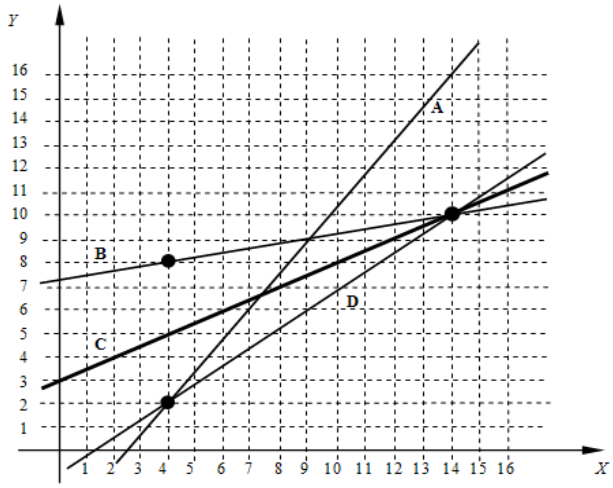


2. Ajuste de la recta de regresión

• 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, β_0 y β_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 la línea recta que represente óptimamente la distribución conjunta de las variables modeladas. Ahora bien, ¿cuál es la línea recta que representa óptimamente a una nube de puntos?, en definitiva, ¿cuál es la que ofrece una mayor bondad de ajuste?

Ejemplo: para los 3 pares de valores en las variables *X* e *Y* representados gráficamente abajo se han superpuesto 4 posibles rectas de regresión, ¿cuál sería la recta de regresión que elegiríamos como mejor?, ¿por qué?



X	*Y*
4	2
4	8
14	10

• 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. Ahora bien, pueden considerarse diferentes procedimientos a la hora de hacer operativa la evaluación de la magnitud de los errores de predicción. Por ejemplo, la tabla inferior (Losilla y cols., 2005) ilustra gráficamente la diferencia entre el uso de tres métodos a la hora de evaluar la magnitud de los errores de predicción de un determinado modelo de regresión: la suma de los errores (*SE*); la suma de los valores absolutos de los errores (*SAE*); y la suma de los cuadrados de los errores (*SCE*). Para cualquiera de ellos, tendrá un mejor ajuste la ecuación de regresión que tenga un valor más próximo a 0.

Ejemplo: en la tabla inferior se muestra el resultado de aplicar los 3 métodos considerados a cada una de las 4 ecuaciones de regresión ajustadas a los datos del ejemplo anterior, ¿cuál de ellos hace corresponder como mejor modelo a aquél que hemos elegido anteriormente de forma gráfica?, ¿qué ventajas e inconvenientes encontramos a estos métodos?

	Método *SE* ΣE_i	Método *SAE* $\Sigma \lvert E_i \rvert$	Método *SCE* ΣE_i^2
Recta A: $Y = -3{,}6 + 1{,}4 \cdot X$	0+6+(–6) = 0	0+6+6 = 12	$0+6^2+(-6)^2 = 72$
Recta B: $Y = 7{,}2 + 0{,}2 \cdot X$	–6+0+0 = –6	6+0+0 = 6	$(-6)^2+0+0 = 36$
Recta C: $Y = 3 + 0{,}5 \times X$	3+(–3)+0 = 0	3+3+0 = 6	$3^2+(-3)^2+0 = 18$
Recta D: $Y = -1{,}2 + 0{,}8 \times X$	0+6+0 = 6	0+6+0 = 6	$0+6^2+0 = 36$

(SE: Sumatorio de los errores; SAE: Sumatorio de valores absolutos de los errores; SCE: Sumatorio de cuadrados de los errores)

• Como puede observarse, el método *SE* enmascara la posible existencia de errores de gran magnitud que, al sumarse y ser de distinto signo, se compensan entre sí dando lugar a un valor de *SE* que puede llegar a ser bajo o incluso nulo. Tanto el criterio *SAE* como el *SCE* salvan este inconveniente, sin embargo, el método *SCE* se ve favorecido por la existencia de errores que, en general, sean tan bajos como sea posible, pues los errores individuales altos, al elevarse a cuadrado, se convierten en números muy grandes. En resumen, la ventaja del método *SCE* estriba en que su valor será más bajo cuando globalmente los errores para todas las observaciones sean pequeños, algo que resulta deseable para una recta que represente a todos los datos y que pueda utilizarse a la hora de realizar predicciones.

• Dadas la ventaja del método *SCE* frente a otros a la hora de evaluar la magnitud de los errores de predicción, éste ha venido en constituirse como el método más popular a la hora de estimar los coeficientes de la ecuación de regresión. 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, de acuerdo al método de mínimos cuadrados ordinarios, las fórmulas de obtención de los parámetros de la ecuación de regresión que van a satisfacer que la *SCE* sea mínima son las siguientes:

$$\beta_1 = \rho_{XY} \cdot \frac{\sigma_Y}{\sigma_X} \qquad \beta_0 = \mu_Y - \beta_1 \cdot \mu_X$$

• Y en el caso que los mismos deban ser estimados a partir de datos muestrales, los mejores estimadores puntuales de los anteriores parámetros son los siguientes estadísticos:

$$\hat{\beta}_1 \rightarrow b_1 = r_{XY} \cdot \frac{s_Y}{s_X} \quad o \quad r_{XY} \cdot \frac{s'_Y}{s'_X} \qquad \hat{\beta}_0 \rightarrow b_0 = \overline{Y} - b_1 \cdot \overline{X}$$

• A partir de lo anterior, la ecuación de la recta de regresión quedaría expresada a nivel muestral como $\hat{Y}_i = b_0 + b_1 \cdot X_i$, si bien, también aparece en algunos libros de texto como $\hat{Y}_i = a + b \cdot X_i$.

Ejercicio 2:

a) Obtener el valor de los coeficientes b_0 y b_1 para el ejemplo sobre afrontamiento y estrés, teniendo en cuenta los siguientes resultados: r_{xy}= –0,847; s_X = 24,8; 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 un sujeto con 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 del modelo de regresión

• La bondad de ajuste de un modelo de regresión se refiere al grado en que éste es conveniente como modelo que representa a 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 como modelo que represente a ambas variables. 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 A), o bien, que el modelo de regresión lineal no sea el más adecuado para ese propósito (ver, por ejemplo, el caso de la Figura B).

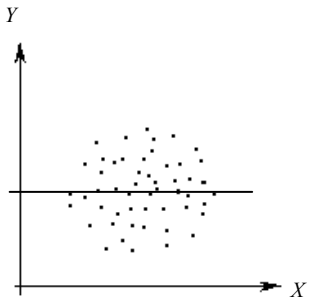


Figura A: Ausencia de relación.

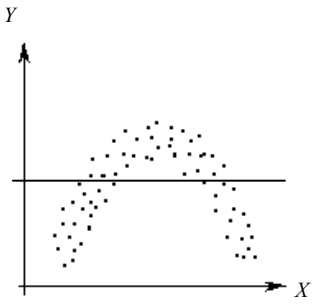
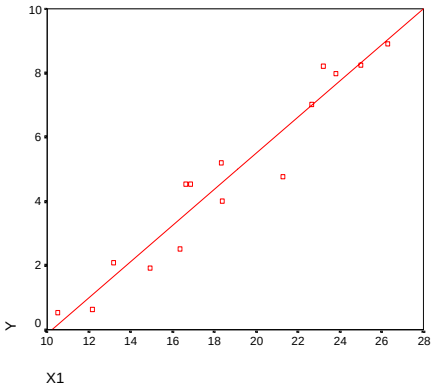
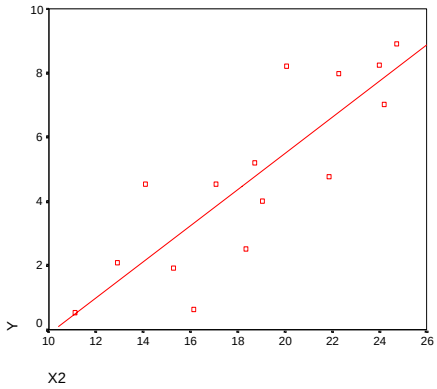


Figura B: Relación no lineal.

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 evaluación de la bondad del ajuste 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 parámetros 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.

• Un índice derivado del anterior es 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 ($S^2_{Y \cdot X}$). De nuevo, este índice adolece del mismo problema de interpretación que *SCE*.

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

• 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:

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

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

$$s_Y^2 = s_{Y'}^2 + s_{Y \cdot X}^2$$

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). (Y si se hubiese dividido por *n-1*, lo mismo con cuasi-varianzas:)

• Consecuencia de la igualdad de descomposición de la varianzas, se puede plantear un índice de la bondad de ajuste como razón de la varianza explicada por el modelo de regresión ($s_{Y'}^2$) respecto a la varianza total (s_Y^2):

$$s_{Y'}^2 \Big/ s_Y^2$$

La anterior razón, conocido como coeficiente de determinación (R^2), puede también expresarse en forma de razón de cuasi-varianzas o de sumas de cuadrados:

$$R^2 = \frac{s_{Y'}^2}{s_Y^2} = \frac{s_{Y'}^{2'}}{s_Y^{2'}} = \frac{SC_{Y'}}{SC_Y}$$

• El coeficiente de determinación (R^2) 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 a 1, indicará mejor bondad de ajuste del modelo de regresión a la distribución conjunta de las variables. Si R^2 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.

$$CALN = \frac{SC_{Y \cdot X}}{SC_Y} = \frac{s^2_{Y \cdot X}}{s^2_Y} = \frac{s'^2_{Y \cdot X}}{s'^2_Y}$$

Obviamente, $CALN = 1 - R^2$

• Destacar que, en el caso del modelo de regresión lineal simple, el coeficiente de determinación puede ser también calculado elevando al cuadrado el coeficiente de correlación de Pearson entre la variable predictora y la variable de respuesta $\rightarrow$ $R^2 = r^2_{XY}$, lo cual puede facilitar enormemente el cálculo de R^2 si se conoce r_{XY} . En resumen:

$$R^2 = \frac{SC_{Y'}}{SC_Y} = \frac{s^2_{Y'}}{s^2_Y} = \frac{s'^2_{Y'}}{s'^2_Y} = r^2_{XY}$$

Ejemplo de cálculo de la recta de regresión de Y sobre X a partir de los siguientes 5 pares de puntuaciones en ambas variables:

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

$\overline{X} = 6;\ S_X = 3{,}16;\ \overline{Y} = 7;\ S_Y = 3{,}74;\ r_{XY} = 0{,}69$

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

Obtención valores predichos $\hat{Y}_i$ para cada sujeto:

X	Y	$\hat{Y} = 2,08 + 0,82 \cdot X$	$E\ (Y_i - \hat{Y}_i)$	$(Y_i - \hat{Y}_i)^2$	$(\hat{Y}_i - \overline{Y})^2$
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
				$s^2_{Y \cdot X} = 36,4/5 = 7,28$	$s^2_{Y'} = 33,62/5 = 6,72$

A partir de los valores predichos se puede obtener:

- La varianza de los errores (o residuales) $\rightarrow$ $s^2_{Y \cdot X} = 7,28$

- La varianza de las puntuaciones predichas $\rightarrow$ $s^2_{Y'} = 6,72$

Descomposición de la varianza de *Y*:

$s^2_Y = 3,74^2 = 14$

$14 = 6,72 + 7,28$

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

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

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*):

$CALN = 7,28/14 = 0,52 \quad (= 1 - 0,48)$

Ejercicio 3: Al estudiar la relación entre dos variables *X* e *Y*, sabemos que la varianza de *Y* es 10 y la varianza de los errores es 8. ¿Cuál es el valor del coeficiente de determinación y del de alienación?, ¿y el del coeficiente de correlación de Pearson entre *X* e *Y*?

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

Obtener a partir de los mismos: (1) medias y desviaciones típicas de las dos variables [‘a mano’ o, mejor, con la calculadora]; (2) el coeficiente de correlación de Pearson entre ambas variables [ídem]; (3) la ecuación del modelo de regresión lineal de Y sobre X [ídem]; (4) los valores predichos por la ecuación de regresión para cada sujeto ($\hat{Y}_i$); (5) los errores de predicción o residuales para cada sujeto (E_i); (6) la varianza de los errores ($s^2_{Y \cdot X}$); (7) la varianza de Y (s^2_Y); (8) la varianza de las puntuaciones predichas ($s^2_{\hat{Y}}$) [‘a mano’ o, mejor, con la calculadora]; (9) comprobar que es cierta la igualdad de la descomposición de la varianza ($S^2_Y = S^2_{\hat{Y}} + S^2_{Y \cdot X}$); (10) el coeficiente de determinación [de dos formas: (10.1) a partir de las varianzas; (10.2) a partir del coeficiente de correlación entre X e Y]; (11) interpretar las estimaciones puntuales de los parámetros de la ecuación de regresión obtenidos (b_0 y b_1); (12) estimar según el modelo de regresión obtenido cuál será la puntuación media obtenida a final de curso para 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 el programa SPSS para los datos del ejercicio anterior. Identificar en los mismos los resultados obtenidos en el ejercicio anterior (apartados 2, 3 y 6 a 10).

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			

a Variables predictoras: (Constante), Horas_estudio

b Variable dependiente: Rendimiento_curso

Coeficientes(a)

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 el ejemplo anterior de las variables de "Afrontamiento" y "Estrés" sabemos que r_{XY} = –0,847 y que S_Y = 22,37. ¿Cuál será el valor del coeficiente de determinación?; ¿cómo se interpreta dicho valor?; ¿cuál es el valor de la varianza de Y explicada por el modelo de regresión (en este caso, por la variable "Afrontamiento")?, ¿y cuál el de la varianza de los residuales?

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]