EJERCICIOS RESUELTOS TEMA 7: Regresión y correlación

Ejercicio 1

En determinado barrio se desea saber si existe alguna relación entre la edad de los vecinos y la "percepción de inseguridad en el barrio", medida en una escala del 0-10 donde el 0 representa "totalmente seguro" y el 10 representa "totalmente inseguro". Se realiza un pequeño pre-test con 10 individuos, obteniendo los siguientes datos:

Edad	Inseguridad
34	4,5
27	3
65	7
20	3,5
53	8
49	5
42	4
31	4
55	5,5
61	7,5

Se pide:
a) Estudiar la relación entre las dos variables "edad" y "percepción de inseguridad" a través de regresión lineal.
b) Representar gráficamente la nube de puntos y la recta que las relaciona. ¿Qué podemos decir de esta relación?
c) ¿Qué puntuación sobre la inseguridad en el barrio obtendría un individuo de 25 años? ¿Y un individuo de 70?
d) Estudiar la correlación de las variables e interpretar los coeficientes.
e) ¿Qué valor de la edad presenta mayor residuo? ¿Cuál es residuo para la edad de 42 años? ¿Y para la edad de 31?

Solución

a) Tenemos dos variables y queremos observar su relación. Para ello, debemos definir cuál de ellas será x y cuál y. En general, para esta definición debemos decidir qué variable es "independiente" (que llamaremos x) y cuál es "dependiente" (que llamaremos y), de tal forma que podamos formular una hipótesis que relacione ambas variables. Es importante tener en cuenta que esta "dependencia" no hay que considerarla simplemente como una causa-efecto, pues las relaciones entre variables, sobre todo en investigación social, son bastante más complejas. Pero sí podemos comprobar, y ese será nuestro objetivo, cómo una variable se relaciona con la otra de tal forma que cambios en una variable se asocian con cambios en la otra. En este caso, tiene sentido estudiar si la percepción de inseguridad cambia en función de la edad, o dicho de otra manera, si la edad influye en la percepción de inseguridad. Por eso, consideraremos:

x: edad
y: percepción de inseguridad en el barrio

Para estudiar la relación entre x e y mediante un modelo lineal necesitamos hallar la recta de regresión:

$y = a + bx$

Y para ello es imprescindible obtener los coeficientes *a* y *b* de la recta. Sabemos que *a* es el punto de corte de la recta en el eje de ordenadas (u "ordenada en el origen") y *b* es la pendiente de la recta, de tal forma que su signo determinará el sentido positivo o negativo de esa pendiente. Primero hallamos *b*, para lo que necesitamos calcular primero la covarianza S_{xy} y la varianza de *x* (S_x^2):

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

La covarianza es: $S_{xy} = \frac{\sum x_i y_i}{n} - \bar{x}\bar{y}$. Hallamos entonces los términos que la fórmula requiere:

x	*y*	$x_i y_i$
34	4,5	153
27	3	81
65	7	455
20	3,5	70
53	8	424
49	5	245
42	4	168
31	4	124
55	5,5	302,5
61	7,5	457,5
	Total:	2480

$\sum x_i y_i = 2480$ $n = 10$

$\bar{x} = \frac{437}{10} = 43{,}7$ $\bar{y} = \frac{52}{10} = 5{,}2$

Ya podemos calcular la covarianza:

$$S_{xy} = \frac{\sum x_i y_i}{n} - \bar{x}\bar{y} = \frac{2480}{10} - 43{,}7 \cdot 5{,}2 = 20{,}76$$

Como la covarianza es ≠ 0 podemos afirmar que existe alguna relación entre las variables, aunque con ella es difícil saber su intensidad, pues la covarianza no tiene un valor máximo o mínimo. Serán otros los coeficientes que nos permitan determinar esa intensidad, y que calcularemos en otro apartado de este ejercicio.

Para el cálculo de *b* necesitamos también conocer la varianza de *x*:

$$S_x^2 = \frac{\sum(x_i - \bar{x})^2}{n} = 209{,}41$$

NOTA: también, podemos utilizar esta otra fórmula de la varianza, equivalente a la anterior, que nos simplifica bastante los cálculos:

$$S_x^2 = \frac{\sum x_i^2}{n} - \bar{x}^2$$

$$b = \frac{S_{xy}}{S_x^2} = \frac{20{,}76}{209{,}41} = 0{,}0991356$$

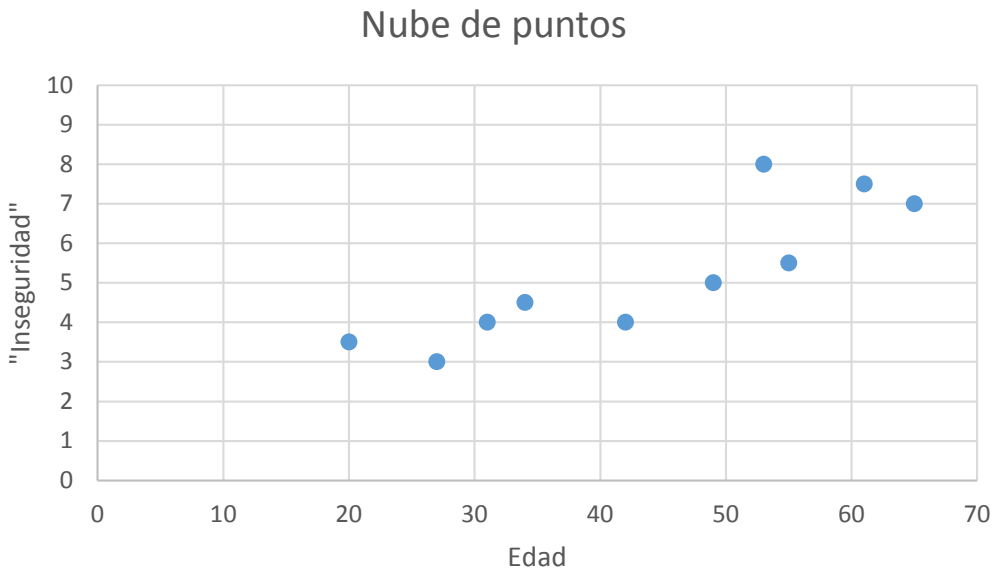
Una vez hemos obtenido *b*, podemos calcular *a*:

$a = \bar{y} - b\bar{x} = 5{,}2 - 0{,}0991356 \cdot 43{,}7 = 0{,}8677742$

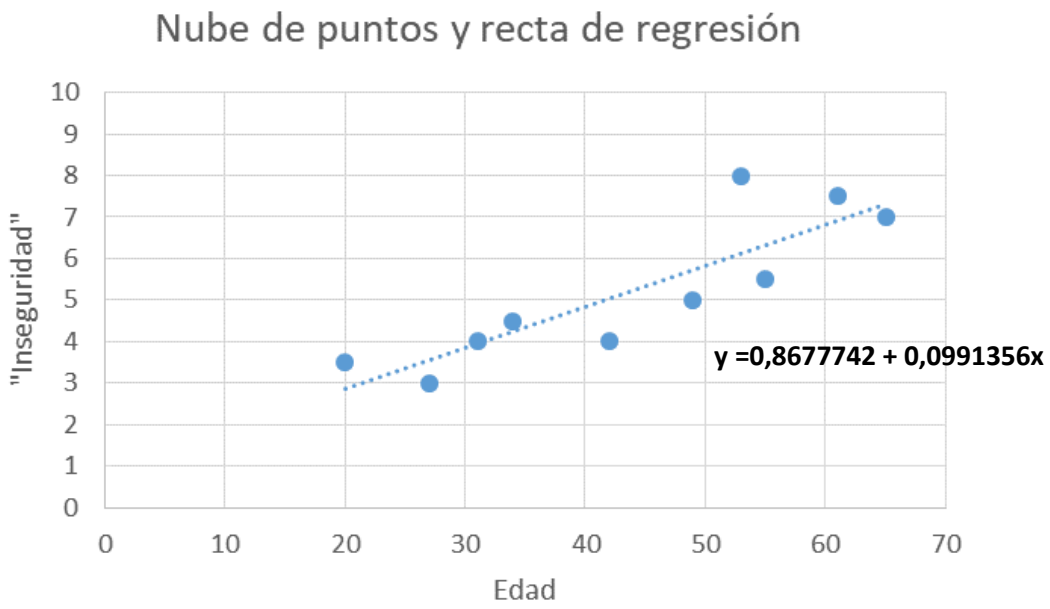
Recta de regresión:

$$\boldsymbol{y = 0{,}8677742 + 0{,}0991356x}$$

b) En primer lugar representamos la nube de puntos teniendo en cuenta cada par de valores (*x*, *y*) de los 10 individuos del estudio:



La nube de puntos adopta una forma ascendente, algo más alineada en las edades más jóvenes y centrales, y más dispersa, aunque también ascendente a partir de los 50 años. Esta primera impresión a través de la nube de puntos, nos indica una cierta relación positiva entre las dos variables. Vamos a representar la recta de regresión que hemos calculado. Para ello, solo tenemos que dar un par de valores cualesquiera a *x* dentro del rango de la variable y calcular la correspondiente *y* a partir de la recta de regresión (recordemos que una recta siempre se puede dibujar como la línea que pasa entre dos puntos):



La recta de regresión tiene pendiente positiva (algo que también sabemos porque b>0). Ello indica que aumentos en X indican también aumentos en Y determinados por la magnitud de la pendiente de la recta (b). Por tanto, la recta indica que conforme aumenta la edad, también aumenta la percepción de inseguridad en el barrio. Sin embargo, la recta solo establece la posible relación entre X e Y, pero no sabemos si es un buen ajuste teniendo en cuenta los datos observados en la realidad. Para ello, necesitamos otras medidas que cuantifiquen la correlación, es decir la intensidad de esa relación.

c) La recta permite "predecir" los valores de Y (en nuestro caso, de la percepción de inseguridad), según los valores que tome X (la edad). Para ello, solo tenemos que sustituir en la recta de regresión los valores de X.

Para $x = 25$

$$y = 0{,}8677742 + 0{,}0991356 \cdot 25 = 3{,}35$$

Un individuo de 25 años puntuaría con alrededor de 3,35 puntos la inseguridad en el barrio

Para $x = 70$

$$y = 0{,}8677742 + 0{,}0991356 \cdot 70 = 7{,}8$$

Un individuo de 70 años puntuaría la inseguridad en el barrio con 7,8 puntos. Vemos, por tanto, que hay una notable diferencia entre los individuos jóvenes respecto a los mayores en su percepción de inseguridad, aumentando según aumenta la edad.

d) Al estudiar la correlación estudiamos la intensidad de la relación entre las dos variables. Es muy importante estudiar la correlación para saber si el ajuste de la recta de regresión es o no un buen indicador de la relación entre las variables. Si la correlación (la variación conjunta de x, *y*) es baja, entonces la recta no explica mucho de la relación entre las variables. En cambio, si

la correlación es alta, la recta de regresión puede ser útil para ver en qué medida los cambios en *x* influyen en *y*. Vamos a estudiar la correlación calculando en primer lugar, **el coeficiente de correlación de Pearson (r)**:

$$r = \frac{S_{xy}}{S_x S_y}$$

Para ello necesitamos la covarianza y las desviaciones típicas de ambas variables S_x y S_y

$S_{xy} = 20{,}76$

$S_x = \sqrt{209{,}41} = 14{,}471$ (la podemos calcular directamente a partir de la varianza de *x* obtenida anteriormente)

$$S_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n}} = 1{,}6613$$

$$r = \frac{S_{xy}}{S_x S_y} = \frac{20{,}76}{14{,}471 \cdot 1{,}6613} = 0{,}8635$$

Dado que el coeficiente de correlación *r* varía entre +1 (máxima relación positiva) y -1 (máxima relación negativa), obtener r = 0,86 indica una correlación bastante alta entre las variables. Por tanto, para este colectivo de vecinos, la edad correlaciona de forma bastante importante con la percepción de inseguridad.

Sin embargo, aún podemos dar un paso más viendo exactamente qué proporción de la variabilidad de Y es explicada por la recta de regresión. Esto lo podemos calcular a través del **coeficiente de determinación (r^2):**

$r^2 = (0{,}8635)^2 = 0{,}7456$

Esto indica que un 74,56% de la variación de Y es explicado por la recta de regresión. Es decir, prácticamente las tres cuartas partes del comportamiento de Y a partir de X se explica por la relación establecida en la recta de regresión. El ajuste de la recta es, por tanto, razonablemente bueno. Ese poco más de 25% del comportamiento de Y que no puede explicarse por la recta se deberá a otros factores (variables) que no hemos contemplado en el modelo.

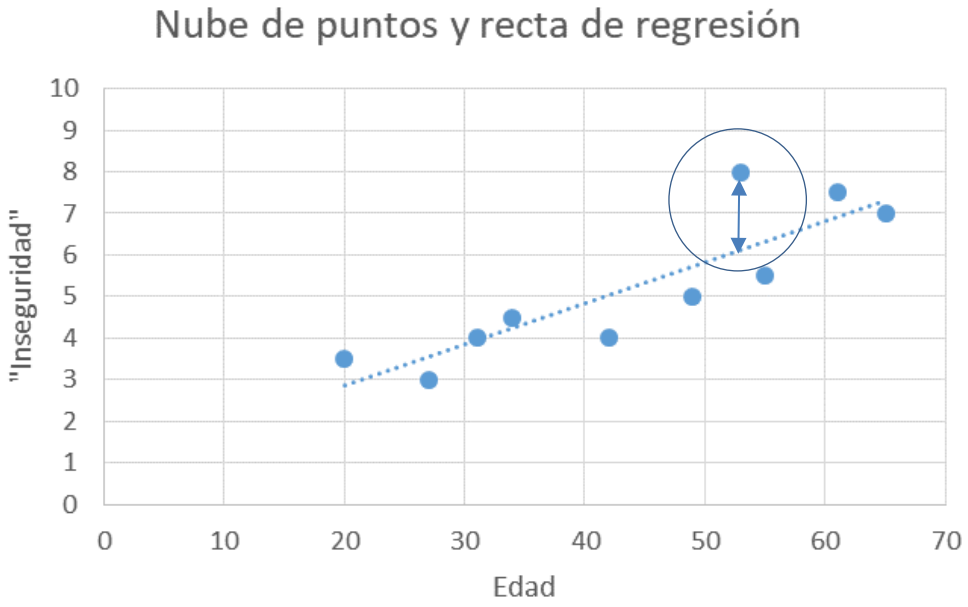
e) Sabemos que la recta de regresión es un ajuste estimado de los valores de Y a partir de los valores de X, pero mientras r no sea ± 1, la relación expresada por la recta no será perfecta y habrá residuos (ε_i), es decir, diferencias entre cada valor de y_i observado en la realidad y su correspondiente $\hat{y}_i$ calculado mediante la recta de regresión, debido a factores o variables más allá de la "edad" que pueden influir en la "percepción de inseguridad" y que no hemos incluido en nuestro modelo. Por eso, la recta de regresión se expresa como:

$$\hat{y}_i = a + bx_i + \varepsilon_i$$

Y los residuos como:

$$\varepsilon_i = (y_i - \hat{y}_i)$$

Como tenemos pocos casos, podemos ver fácilmente qué valor de la "edad" presenta mayor residuo a través de la representación gráfica:



En el gráfico podemos apreciar que el valor observado de *y* que más se separa de la recta de regresión y, por tanto del $\hat{y}_i$ estimado por la recta es el correspondiente a $x = 53$. Podemos decir, entonces, que el individuo con 53 años es aquel para el que la recta "pronostica peor" su percepción.

Para calcular el residuo en la edad de 42 años, debemos calcular ε para $x = 42$, hallando previamente el valor estimado de $\hat{y}_i$ para esa edad en la recta de regresión y comparándolo con el valor observado de *y* en nuestros datos

$$\hat{y} = 0{,}8677742 + 0{,}0991356x$$

$$\hat{y} = 0{,}8677742 + 0{,}0991356 \cdot 42 = 5{,}03$$

$$\varepsilon_i = (y_i - \hat{y}_i) = 4 - 5{,}03 = -1{,}03$$

El resultado indica que para el individuo con edad de 42 años, su puntuación real queda 1,03 puntos por debajo de lo que pronostica la recta.

Para calcular el residuo en la edad de 31 años, realizamos la misma operación:

$$\hat{y} = 0{,}8677742 + 0{,}0991356x$$

$$\hat{y} = 0{,}8677742 + 0{,}0991356 \cdot 31 = 3{,}94$$

$$\varepsilon_i = (y_i - \hat{y}_i) = 4 - 3{,}94 = 0{,}06$$

En este caso, el ajuste es mucho mejor, casi perfecto. El residuo es muy pequeño y la recta de

regresión pronostica muy bien la percepción de inseguridad en el individuo de esa edad.

Ejercicio 2

En determinada zona se está investigando la relación entre el consumo de alcohol y el de otros estupefacientes entre la población juvenil. Se sabe que el consumo medio de alcohol es de 2,3 veces por semana (con una desviación típica de 2,5) mientras que el consumo medio de otros estupefacientes es de 1,3 veces por semana (con una desviación típica de 3). El coeficiente de determinación es del 45,1%. Halle la recta que relaciona el consumo de alcohol y el consumo de otros estupefacientes para esa población de jóvenes.

Solución

En primer lugar, determinaremos cuál será la variable "x" y la variable "y":

x: consumo de alcohol
y: consumo de otros estupefacientes

Debemos hallar la recta:

$y = a + bx$

Para ello, debemos calcular los parámetros «a» y «b» que la definen.

Tenemos los siguientes datos:

$\bar{x} = 2,3$ $\bar{y} = 1,3$
$S_x = 2,5$ $S_y = 3$
$r^2 = 0,451$

Sabemos que el coeficiente de correlación es $r = \frac{S_{xy}}{S_x S_y}$

A partir de r podremos calcular entonces la covarianza (S_{xy}), necesaria para obtener el parámetro *b* de la recta:

$r = \sqrt{0,451} = 0,6715653$

$0,6715653 = \frac{S_{xy}}{2,5 \cdot 3}$ $0,6715653 \cdot 7,5 = S_{xy}$ $S_{xy} = 5,0367398$

Por tanto, *b* será:

$$b = \frac{S_{xy}}{S_x^2} = \frac{5,0367398}{2,5^2} = 0,8058783$$

Ahora hallamos el parámetro *a*:

$a = \bar{y} - b\bar{x}$ = 1,3 – 0,8058783 · 2,3 = -0,5535202

La recta de regresión que relaciona el consumo de alcohol con el consumo de estupefacientes en esa población juvenil es:

y = -0,5535202 + 0,8058783x

Ejercicio 3

Responda a las siguientes cuestiones, argumentándolas únicamente desde la estadística teórica:

a) ¿Sería estadísticamente posible encontrar una correlación r = 0,9 en una población de asalariados y asalariadas entre las variables "sexo" y "salario"?

b) ¿Y una correlación r = -0,3?

c) Observe esta recta de regresión:

y = -3,5 + 2,7x

¿Podemos decir, teniendo en cuenta el signo del parámetro *a*, que la relación entre las dos variables es negativa?

Solución

a) No. El coeficiente de correlación está definido para variables medidas a nivel de intervalo y como la variable "sexo" es del tipo nominal, no tiene sentido hablar de correlación estadística, independientemente de cual sea la relación entre ambas variables. Las rectas de regresión lineal sólo tienen sentido con variables cuyo nivel de medición es de intervalo (cuantitativas continuas).

b) No. Por las mismas razones que en el punto anterior, r carece de sentido para explicar la relación entre sexo y salario, independientemente de su signo.

c) No. El signo de parámetro *a* no tiene ninguna relación con la pendiente de la recta de regresión, que es lo que determina la variación de *y* en función de las variaciones de *x*. La pendiente viene determinada por el parámetro *b*, y en este caso es positivo, por lo que aumentos en *x* conducirán a aumentos en y. Si el signo de *b* fuera negativo, a medida que aumenta *x*, disminuiría *y*.

Ejercicio 4

Un estudio relaciona los salarios y la edad de los empleados de una gran empresa. La correlación observada entre las dos variables es r = 0,6. La media de edad de los trabajadores es de 36 años, con una desviación típica de 6 años y, el salario medio es de 1.350 euros, con una desviación típica de 400 euros. Calcule los coeficientes a y b de la ecuación de la recta de regresión que relaciona x (edad) con y (salario), que permita estimar el salario correspondiente a una determinada edad. A partir de esa ecuación, establezca el salario que correspondería a una persona de 45 años.

Solución

Sabemos que el coeficiente de correlación se puede calcular mediante la expresión:

$$r = \frac{S_{xy}}{S_x S_y}$$

Conocemos r, S_x y S_y de modo que podemos calcular la covarianza:

$$S_{xy} = r \cdot S_x \cdot S_y = 0{,}6 \cdot 6 \cdot 400 = 1.440$$

Conocida la covarianza podemos calcular el parámetro b de la recta pedida:

$$b = \frac{S_{xy}}{S_x^2} = \frac{1440}{6^2} = 40$$

Sabemos que los valores medios de las variables pertenecen a la recta de regresión, por lo que podemos utilizar los valores de la media de X y de Y para obtener el parámetro a, una vez conocido b:

$$\bar{y} = a + b\bar{x}$$

Y despejando a:

$$a = \bar{y} - b\bar{x} = 1.350 - 40 \cdot 36 = -90$$

Por tanto la recta pedida será:

$$y = -90 + 40x$$

Sin necesidad de representar esta recta de regresión, vemos que relaciona ambas variables de forma positiva, pues la pendiente b > 0. Ello indica que los empleados de mayor edad ganan más que los más jóvenes en una cuantía determinada en cada caso por la recta de regresión

Por último, para conocer el salario que correspondería a una edad de 45 años sustituimos en la recta:

$y = -90 + 40 \cdot 45 = 1.710$

El salario correspondiente a un empleado de 45 años es de 1710 euros.