22. Relación entre variables

Comencemos con un ejemplo.

Ejemplo 1: En una ciudad con graves problemas de obesidad en la población, se solicitó a un grupo de 60 adolescentes que registrara durante un mes la cantidad de horas que dedicaban cada día a actividades sedentarias (mirar televisión, estudiar o utilizar la computadora) y las promediaran. La tabla 22.1 presenta la edad en años (Edad), el género (Varón, Mujer), el promedio de horas por día dedicadas a actividades sedentarias (Horas) y un número (Id) para identificar a cada participante:

PROMEDIO DE HORAS POR DÍA DEDICADAS A MIRAR TELEVISIÓN, ESTUDIAR O UTILIZAR LA COMPUTADORA. TABLA 22.1

Varones						Mujeres					
Id	Edad	Horas	Id	Edad	Horas	Id	Edad	Horas	Id	Edad	Horas
1	11,2	5,5	16	14,6	5,3	31	11,1	4,3	46	13,9	5,0
2	11,4	5,4	17	15,0	5,2	32	11,2	5,1	47	14,2	4,4
3	11,4	4,5	18	15,4	7,0	33	11,2	4,7	48	14,4	5,6
4	11,5	4,8	19	15,6	5,9	34	11,5	4,5	49	14,9	4,4
5	11,6	5,0	20	15,9	6,6	35	11,6	4,7	50	15,1	5,2
6	11,7	5,5	21	16,2	6,3	36	11,6	4,8	51	15,4	5,1
7	11,9	4,3	22	16,5	5,8	37	11,8	4,4	52	15,6	5,1
8	12,6	5,7	23	17,0	6,9	38	11,9	4,7	53	15,9	5,3
9	12,8	4,7	24	17,3	6,9	39	12,3	5,0	54	16,2	4,7
10	13,2	5,4	25	17,4	6,2	40	12,8	4,7	55	16,4	4,9
11	13,8	5,6	26	17,5	5,5	41	12,8	5,1	56	16,6	6,7
12	13,8	5,5	27	17,8	6,0	42	12,9	5,2	57	17,2	5,0
13	14,0	6,6	28	17,9	6,5	43	13,1	5,8	58	17,4	5,8
14	14,3	5,5	29	18,2	6,4	44	13,5	5,2	59	17,9	5,6
15	14,5	5,4	30	18,3	5,7	45	13,6	5,1	60	18,1	5,8

¿Qué información nos pueden brindar los datos de la tabla 22.1? Podemos comparar la distribución de las variables "Edad" y "Horas", en dos grupos definidos por la variable categórica "Género" con dos categorías: Varón, Mujer.

Medidas resumen. TABLA 22.2

	Varones		Mujeres	
	Edad	Horas	Edad	Horas
Mínimo	11,20	4,30	11,10	4,30
1er. Cuartil	12,65	5,33	12,00	4,70
Mediana	14,55	5,65	13,75	5,05
Media	14,68	5,73	14,07	5,06
3er. Cuartil	16,88	6,38	15,82	5,20
Máximo	18,30	7,00	18,10	6,70

No se observan diferencias llamativas entre las distribuciones de las edades de varones y mujeres, tanto mirando la tabla 22.2 como la figura 22.1. Sin embargo para las horas, todos los valores para los varones entre el 1er cuartil y el máximo (5,33-7,00) son mayores que todos los valores para las mujeres entre el mínimo y el 3er. cuartil (4,30-5,20). Esto sugiere que en general los adolescentes de esta ciudad le dedican más horas a las actividades sedentarias que las adolescentes.

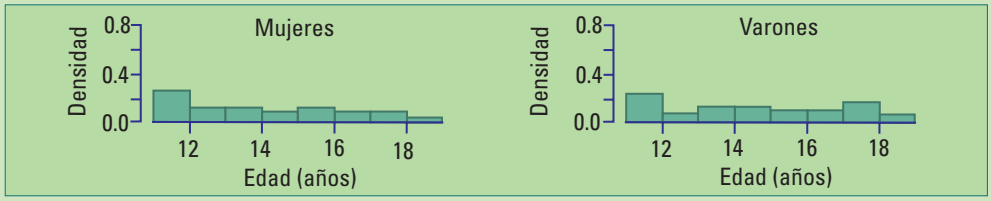


Figura 22.1. *Histogramas de las edades de mujeres y varones.*

Para las mujeres la media muestral de la cantidad de horas por día dedicadas a actividades sedentarias es 5,06. En los varones es 5,73 horas, aproximadamente 3/4 de hora más. La figura 22.2 refuerza esta situación. Los intervalos correspondientes a la mayor cantidad de horas (de 6 a 7) presentan mayor densidad de datos para los varones que para las mujeres. El intervalo más poblado para las mujeres es entre 4,5 y 5,0 horas y en los varones entre 5 y 5,5 horas.

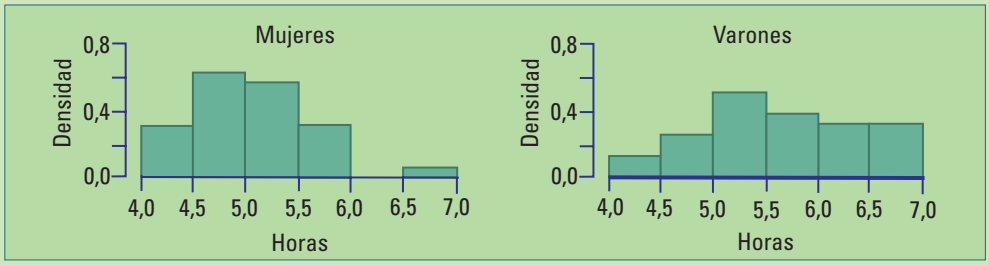


Figura 22.2. *Histogramas de las horas por día dedicadas a actividades sedentarias de mujeres y varones.*

El gráfico caja de la figura 22.3 para mujeres muestra además un valor atípico; se trata de un valor muy alejado del resto. Se destaca también que la caja correspondiente a los varones se encuentra desplazada hacia arriba (hacia los valores mayores) en comparación con las mujeres; por lo tanto, más del 75% de los valores de horas para los varones son mayores que más del 75% de los valores menores de las horas para mujeres. Esto es lo mismo que notamos al describir las medidas resumen.

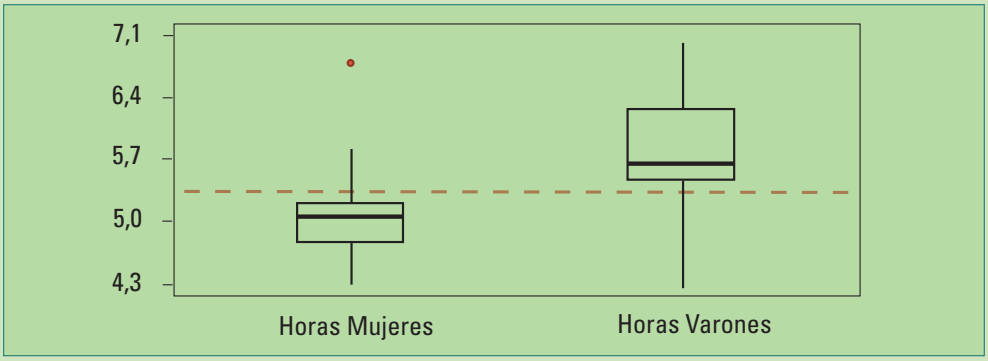


Figura 22.3. *Gráficos caja para la cantidad de horas por día dedicadas a actividades sedentarias de varones y mujeres. Se destacan un valor atípico y los valores de las cajas que no se superponen. La caja para mujeres se encuentra por debajo de la de los varones.*

Hasta aquí comparamos los valores de una variable continua por vez en dos grupos definidos por una variable categórica.

Nos preguntamos ahora:

- ¿Habrá alguna relación entre la cantidad de horas dedicadas a actividades sedentarias y la edad?
- ¿Los más chicos le dedicarán mayor o menor cantidad de horas que los más grandes a ese tipo de actividades?

Se trata en este caso de relacionar los valores de dos variables cuantitativas continuas (horas, edad).

□ 22.1. Diagrama de dispersión

La forma gráfica más habitual de describir la relación entre dos variables cuantitativas es utilizando un **diagrama de dispersión.** Cada **punto** corresponde a **un par de valores** (uno para cada variable), medidos sobre el mismo individuo.

En general, si una de las variables puede pensarse como explicativa de la otra (**variable explicativa**), siempre se la grafica en el eje horizontal (eje x) y la otra (**variable respuesta**) en el eje vertical (eje y).

En el ejemplo de la figura 22.4, la edad es la **variable explicativa.** Pensamos que la edad puede explicar, aunque sea en parte, la cantidad de horas diarias dedicadas a actividades sedentarias (**variable respuesta,** graficada en el eje y).

Cada punto representa a un varón. Está determinado por su edad y la cantidad de horas diarias dedicadas a las actividades sedentarias relevadas (mirar televisión, estudiar o utilizar la computadora).

Para ilustrar como se realiza el gráfico, se destaca el punto correspondiente a Id = 13, Edad =14, Horas = 6,6.

En un diagrama de dispersión observamos el patrón general de la relación entre las variables mirándolo de izquierda a derecha.

Si a medida que **x aumenta** (es decir, nos corremos hacia la derecha del gráfico):

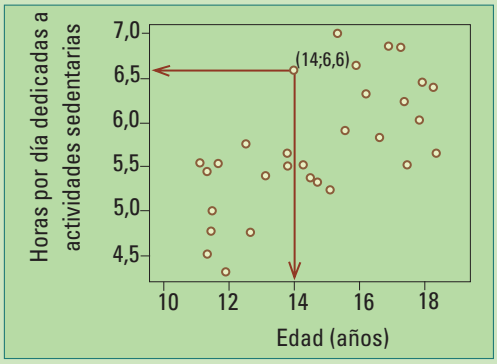


Figura 22.4. *Diagrama de dispersión de las horas en función de la edad para los varones. Se destaca el punto correspondiente a Id = 13, Edad =14, Horas = 6,6.*

- en promedio también lo hace y (los valores de y se encuentran más arriba); esto indica una **asociación** lineal **positiva** entre las variables.
- en promedio y decrece (los valores de y se encuentran más abajo, esto indica una **asociación** lineal **negativa** entre las variables.
- no puede determinarse una tendencia de crecimiento o decrecimiento en los valores de y; esto significa que no hay una asociación lineal entre las variables.

Más formalmente diremos:

- Dos variables están asociadas en forma **positiva** cuando los valores que están por **encima del promedio** de una de ellas tienden a ocurrir mayoritariamente con valores por **encima del promedio** de la otra. Lo mismo ocurre con los que se encuentran por debajo del promedio.
- Dos variables están asociadas en forma **negativa** cuando valores por **encima del promedio** de una suelen estar acompañados por valores por **debajo del promedio** de la otra y viceversa.

La figura 22.5 muestra el diagrama de dispersión dividido en cuatro cuadrantes determinados por la media muestral de las horas (= 5,73) y la media muestral de las edades (= 14,68).

¿Cuántos puntos hay en cada uno de los cuadrantes de la figura 22.5? 11 puntos en el Cuadrante I, 2 puntos en el Cuadrante II, 15 puntos en el Cuadrante III y 3 puntos en el Cuadrante IV.

La mayoría de los puntos se encuentran en el primer y tercer cuadrante. **En el primero** las edades están **por encima de su media muestral** y lo mismo ocurre con las horas. **En el tercero** tanto las edades como las horas son **menores** que **sus respectivas medias** muestrales. Por lo tanto se trata de una **asociación** lineal **positiva.**

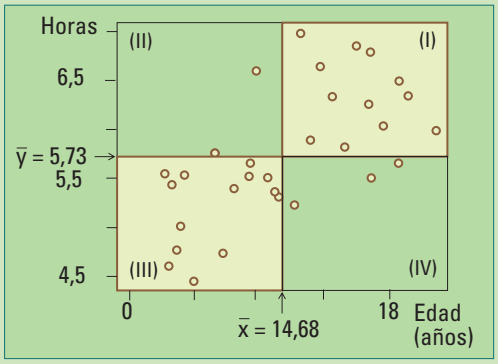


Figura 22.5. *Diagrama de dispersión de las horas en función de la edad para los varones y los cuatro cuadrantes determinados por las medias muestrales de las edades y las horas.*

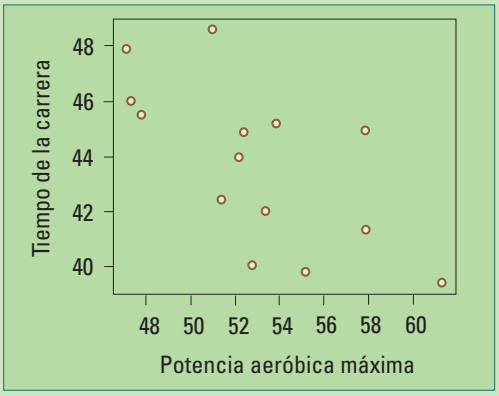


Figura 22.6. *Diagrama de dispersión de Tiempo (Y, min) en 10 km y la Potencia aeróbica máxima (X; ml kg^{-1} min^{-1}) de 14 atletas entrenadas. A medida que aumenta la potencia aeróbica máxima alcanzada disminuye el tiempo.*

Ejemplo 2: Los datos de la tabla 22.3 (Atletas) corresponden a un estudio sobre la relación entre el grado de entrenamiento y el desempeño posterior en una carrera de 10 km. Se evaluaron 14 mujeres entrenadas. El grado de entrenamiento se mide mediante la "Potencia aeróbica máxima" (ml/(kg min) alcanzada y el desempeño posterior mediante el tiempo empleado en completar 10 km durante una competencia.

TIEMPO (Y, min) EN 10 km Y LA POTENCIA AERÓBICA MÁXIMA (X, ml kg^{-1} min^{-1}). TABLA 22.3

Atleta	x	y
1	47,170	47,830
2	47,410	46,030
3	47,880	45,600
4	51,050	48,550
5	51,320	42,370
6	52,180	43,930
7	52,370	44,900
8	52,830	40,030
9	53,310	42,030
10	53,930	45,120
11	55,290	39,800
12	57,910	44,900
13	57,940	41,320
14	61,320	39,370

A medida que **aumenta la Potencia** aeróbica máxima (X) alcanzada por las atletas, durante el entrenamiento previo a la competencia, mejora su rendimiento en la carrera de 10 km, **disminuyendo el tiempo** (Y) (figura 22.6). Por lo tanto, X e Y están asociadas negativamente.

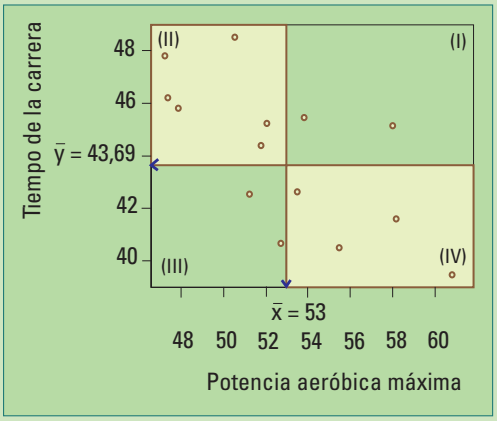


Figura 22.7. *Tiempo (Y, min) en 10 km y la Potencia aeróbica máxima (X; ml kg^{-1} min^{-1}) de 14 atletas entrenadas. Los cuadrantes II y IV contienen 10 de los 14 puntos del diagrama de dispersión, mostrando la asociación negativa entre X e Y.*

¿Cuántos puntos hay en cada uno de los cuadrantes de la figura 22.7?
2 puntos en el Cuadrante I, 6 puntos en el Cuadrante II, 2 puntos en el Cuadrante III y 4 puntos en el Cuadrante IV.

Esta vez, los cuadrantes II y IV, determinados por las medias muestrales de cada una de las variables, con 10 de los 14, contienen la mayoría de los puntos. Reafirmamos que el tiempo en realizar la carrera de 10 km y el nivel de entrenamiento medido por la potencia aeróbica máxima tienen **asociación negativa.**

En general, un diagrama de dispersión muestra la forma, la dirección y el grado de la asociación entre los valores de dos variables cuantitativas. Cuanto más cerca se encuentren los puntos del diagrama de una recta tanto más fuerte es la asociación lineal entre los valores de las variables graficadas. Pero nuestra percepción visual del grado de asociación puede estar equivocada debido a la escala. La figura 22.8. muestra el diagrama de dispersión del mismo conjunto de datos (Atletas) en diferentes escalas. En el diagrama de la izquierda la asociación lineal parece más fuerte en comparación con el de la derecha.

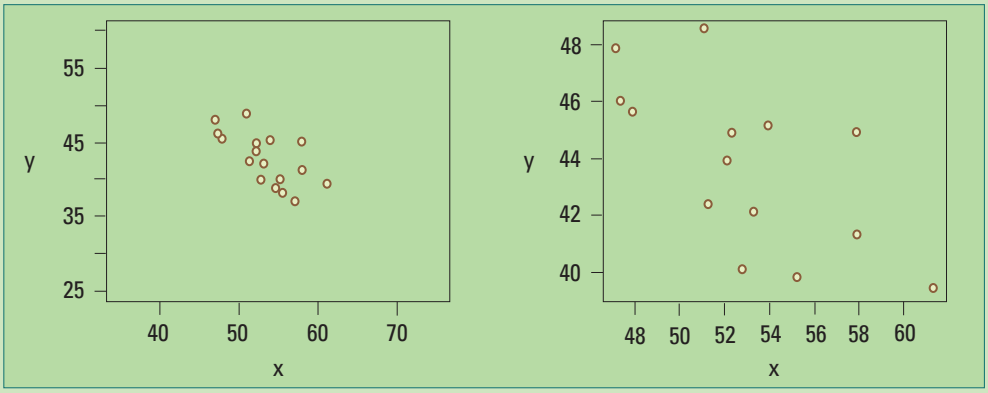


Figura 22.8. *Dos diagramas de dispersión de los mismos datos (tabla 22.3). El de la izquierda sugiere una asociación más fuerte entre las variables que el de la derecha.*

□ 22.2. Coeficiente de correlación

Necesitamos un número que no dependa de las escalas del gráfico de dispersión y represente el grado de asociación lineal entre los pares de valores de dos variables continuas.

¿Qué propiedades debería tener ese número?

- Ser positivo si la asociación lineal es positiva.
- Ser negativo si la asociación lineal es negativa.
- Ser más grande, en valor absoluto, cuánto más cerca se encuentren de alguna recta los pares de valores.
- No depender de las unidades en las que se expresan las variables.

Un número con todas las propiedades anteriores es el Coeficiente de Correlación de Pearson (r):

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

donde $(x_1, y_1)...(x_n, y_n)$, es un conjunto de **pares de datos** de tamaño **n**, correspondiente a observaciones de **dos variables** continuas **X** e **Y.**

Como otras de las fórmulas de cálculo de estadística, **¡asusta!,** pero no es problema para las calculadoras y menos aún para las computadoras.

¿Coeficiente de correlación de Pearson?
Su fórmula me recuerda a la del desvío estándar muestral.

Otra forma de escribir elcoeficiente de correlación es: $$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

donde

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} \quad \textbf{y} \quad s_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}$$

son los **desvíos estándar muestrales de X e Y respectivamente.**

¿Cuánto vale r en el ejemplo 1, de las atletas?

r = -0,659

Como esperábamos es negativo. Su valor absoluto (0,659) es menor a 1.

Más propiedades del coeficiente de correlación muestral r:

- No **depende de las unidades** en que se miden las variables y su valor está siempre entre -1 y 1.
- No distingue entre variable explicativa (**X**) y variable respuesta (**Y**): el coeficiente de correlación entre **X** e **Y** es igual al coeficiente de correlación entre **Y** y **X.**
- A mayor valor absoluto de r, mayor el grado de **asociación lineal.**
- Cuando r = 0 no hay una tendencia lineal creciente o decreciente en la relación entre los valores x´s e y´s.
- Los valores extremos, r = 1 y r = -1, ocurren únicamente cuando **los puntos en un diagrama de dispersión caen exactamente sobre una recta.** Corresponde a asociaciones positivas ó negativas perfectas.
- Valores de **r cercanos a 1** ó **-1** indican que los puntos yacen **cerca** de una recta.

Como el **denominador de r es siempre positivo** $\left(\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}\right)$, para comprender de donde se obtiene su signo, sólo es necesario estudiar el signo del numerador $\left(\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})\right)$

- Cuando la mayoría de los sumandos son positivos: $(x_i-\bar{x})\ (y_i-\bar{y}) > 0$ la suma $\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

  es positiva y por lo tanto **r es positivo.** Ocurre cuando la mayoría de los puntos (x_i, y_i) se encuentran en los cuadrantes (I) y (III). En esos cuadrantes los desvíos $\mathbf{x_i-\bar{x}}$ e $\mathbf{y_i-\bar{y}}$ tienen **el mismo signo** y su producto es positivo.
- Cuando la mayoría de los sumandos son negativos: $(x_i-\bar{x})\ (y_i-\bar{y}) < 0$, o sea cuando los puntos (x_i, y_i) se encuentran en su mayoría en los cuadrantes (II) y (IV), allí los desvíos $\mathbf{x_i-\bar{x}}$ e $\mathbf{y_i-\bar{y}}$ tienen **signos opuestos** y su producto es negativo. La suma $\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ resulta negativa y por lo tanto **r es negativo.**

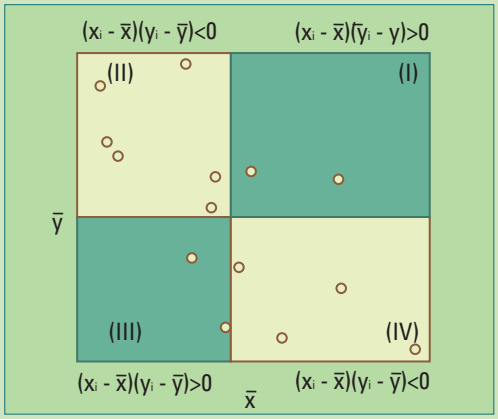


Figura 22.9. *El signo del producto* $(x_i - \bar{x})(y_i - \bar{y})$ *es positivo para los puntos de los cuadrantes (I) y (III); negativos en los otros dos cuadrantes.*

Cuantos más puntos caigan en los cuadrantes (I) y (III) habrá **más sumandos** de

$$\sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})$$

positivos contribuyendo al total positivo **r > 0**

Cuantos más puntos caigan en los cuadrantes (II) y (IV) habrá **más sumandos** de

$$\sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})$$

negativos contribuyendo al total positivo **r < 0**

La figura 22.10 muestra cómo los valores de r decrecen en valor absoluto, se alejan del 1 ó -1, a medida que decrece el grado de asociación lineal entre las variables.

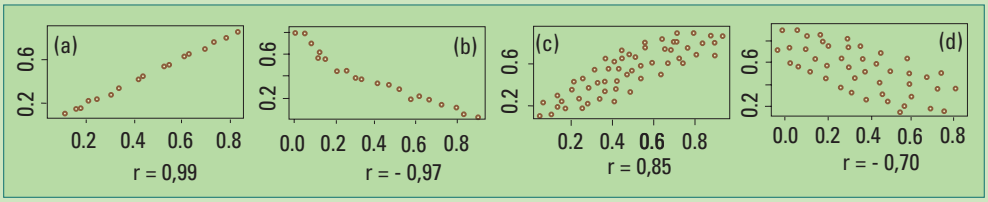


Figura 22.10. *En (a) y (b) se muestran datos con un alto grado de asociación lineal, en el primero la asociación es positiva y en el segundo negativa. En (c) y (d) los datos están menos concentrados sobre una recta pero sigue habiendo claras tendencias: creciente (c) y decreciente (d).*

Karl Pearson (1837 -1936) Estadístico, historiador y pensador británico. Realizó una intensa investigación sobre desarrollo y aplicación de métodos estadísticos a problemas provenientes de la biología. En 1911 fundó el primer departamento de estadística en la Universidad de Londres.

El coeficiente de correlación de Pearson r mide el **grado de asociación lineal** entre los pares de valores de dos variables continuas. Es un número entre -1 y 1. Su signo refleja si la asociación es positiva o negativa y su valor está más acerca de 1 (ó -1) a medida que los puntos del diagrama se aproximan a una recta. Los valores extremos 1 y -1 se obtienen cuando los puntos del diagrama de dispersión están perfectamente alineados.

A veces el coeficiente de correlación no refleja lo esperado, como vemos en (b) y (c) de la figura 22.11. Ante una relación curvilínea (b) el coeficiente de correlación (r = -0.01) indica que no hay asociación entre las variables a pesar de existir una **asociación no lineal** muy fuerte. En (c) el coeficiente de correlación (r = -0.75) indicaría una asociación negativa cuando en realidad se trata de dos grupos de datos, uno con asociación positiva y el otro con asociación nula.

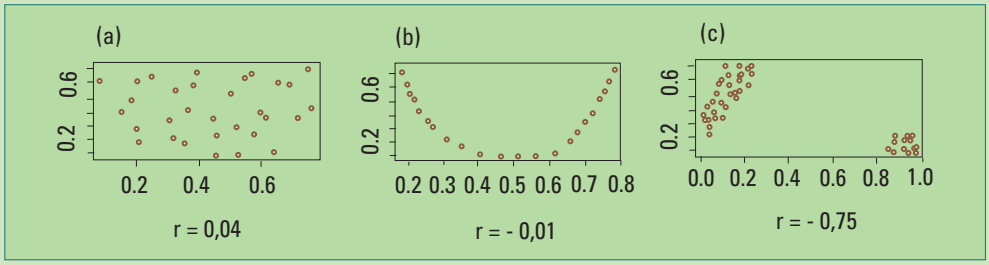


Figura 22.11. *El coeficiente de correlación es cercano a cero cuando los pares de datos no están asociados (a), pero también puede ser nulo o casi nulo ante una relación no lineal entre los valores de X e Y (b).*

□ 22.3. Recta de regresión lineal simple

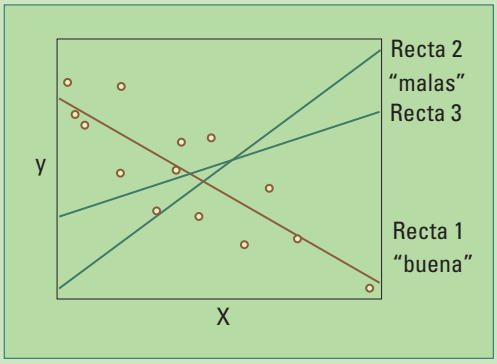


Fig. 22.12. *Tres rectas en un diagrama de dispersión.*

Cuando un diagrama de dispersión muestra un patrón lineal es deseable resumir ese patrón mediante la ecuación de una recta. Esa recta debe representar a la mayoría de los puntos del diagrama, aunque ningún punto esté sobre ella.

La recta 1 de la figura 2.12. representa bien la dirección y el sentido de la asociación entre los valores de **X** e **Y**, pasa "cerca" de la mayoría de los puntos del diagrama de dispersión; decimos que es una recta "buena". Esto no ocurre con las rectas 2 y 3. Pero, ¿cómo podemos elegir la mejor de las rectas posibles?

22.3.1. Cuadrados mínimos

El **método de cuadrados mínimos** propone elegir la recta que **minimiza** la suma de los cuadrados de las **distancias verticales** de cada punto a la recta.

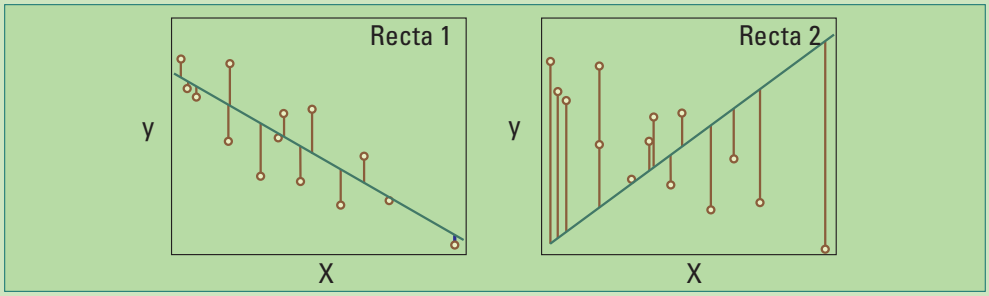


Figura 22.13. *Dos diagramas de dispersión del mismo conjunto de pares de datos y las distancias verticales a dos rectas.*

Los diagramas de dispersión de la figura 22.13 muestran **los mismos puntos.** Las dis**tancias verticales** de los puntos a la recta 1 son, en su mayoría, menores que a la recta 2. Por lo tanto la suma de los cuadrados de esas distancias será menor para la recta 1 que para la recta 2.

Consideremos la ecuación de una recta cualquiera **y=a+bx.** Sean:

- (x_i, y_i) las coordenadas de un punto del plano representando al dato i;
- $(x_i, \hat{y}_i)$ las coordenadas de un **punto sobre la recta** con $x = x_i$

donde $\hat{y}$ (se lee y sombrero sub i) se obtiene reemplazando x_i en la ecuación de la recta ($\hat{y}_i = a+bx_i$, figura 22.14).

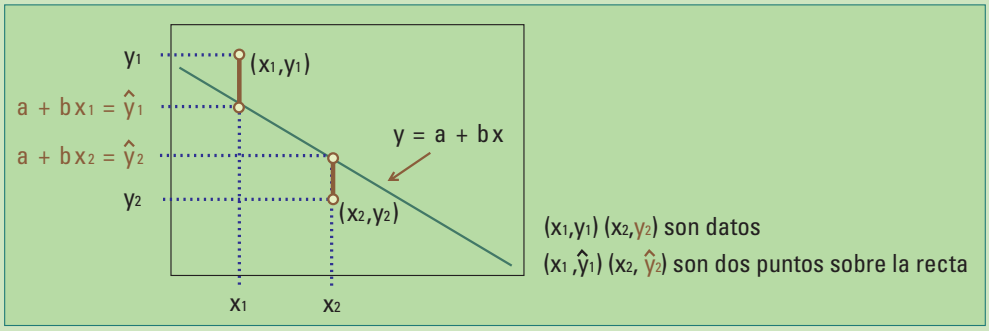


Fig. 22.14. *Dos puntos en el plano y una recta. El primero está por encima de la recta, tiene residuo positivo ($y_1 - \hat{y}_1 > 0$). El segundo está por debajo de la recta, tiene residuo negativo ($y_2 - \hat{y}_2 < 0$).*

La distancia vertical de un punto (x_i, y_i) a la recta es llamada residuo y se obtiene de la siguiente manera:

$$\text{residuo}_i = y_i - \hat{y}_i$$
$$= y_i - (a+bx_i)$$

Algunos **residuos** son **positivos,** la respuesta observada está por encima de la recta, y otros son **negativos,** la respuesta observada está por debajo de la recta. La figura 22.15 muestra esta situación:

- El primer punto se encuentra por encima de la recta $(y_1 > \hat{y}_1)$ luego $y_1 - \hat{y}_1 > 0$.
- El segundo punto se encuentra por debajo de la recta $(y_2 > \hat{y}_2)$ luego $y_2 - \hat{y}_2 < 0$.

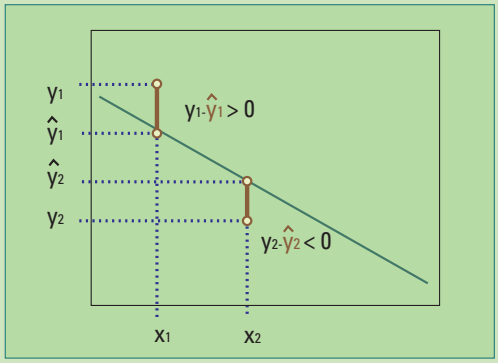


Fig. 22.15. *Dos residuos uno positivo y otro negativo.*

Pero ¿cómo hallamos los coeficientes ***a y b*** de la recta que minimiza la suma de los cuadrados de los residuos?

Mediante el **método de cuadrados mínimos** (CM) ***a y b*** se eligen de manera que **la suma de los cuadrados de los residuos**

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - (a + bx_i) \right)^2$$

sea mínima.

Los coeficientes de la recta estimada por CM se calculan, a partir de los datos, mediante las siguientes ecuaciones:

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

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

La deducción está fuera del alcance de este texto, sin embargo es interesante lo que estas ecuaciones nos muestran:

- La primera ecuación dice $b = r \frac{s_y}{s_x}$, donde **r** es el coeficiente de correlación y $\mathbf{s}_x$ y $\mathbf{s}_y$ son los desvíos estándar muestrales de las x's y las y's respectivamente. Por lo tanto si $s_x = s_x = 1$ resulta que la pendiente de la recta ajustada es igual al coeficiente de correlación.
- La segunda ecuación nos dice que **la recta de cuadrados mínimos pasa por** el punto $(\bar{x}, \square)$ pues sus coordenadas satisfacen la ecuación de la recta ajustada $(\square = a + b\bar{x})$.

- La suma de los residuos es 0 y equivalentemente su promedio: Si sumamos los residuos, $y_1 - (a+bx_1)$, y los dividimos por n entonces,

$$\frac{\sum_{i=1}^{n} y_i}{n} - \frac{\sum_{i=1}^{n} (a + bx_i)}{n} = \bar{y} - (a + b\bar{x}) \qquad \text{pues } \mathbf{a = \hat{y} - b\bar{x}}$$

$$= 0$$

En la práctica, la recta de CM se obtiene utilizando una calculadora o, mejor aún, una computadora.

Ninguna otra recta tendrá, para el mismo conjunto de datos, una suma de cuadrados de los residuos tan baja como la obtenida por CM. En este sentido, el método de mínimos cuadrados brinda la solución que mejor ajusta a un conjunto de datos.

Ejemplo 3: Retomemos los datos de la tabla 22.3 (Atletas) considerando al Tiempo (Y) como variable respuesta y a la Potencia aeróbica máxima (X) como variable explicativa. Pensamos que el grado de entrenamiento medido por la variable X puede explicar, aunque sea parcialmente, el tiempo realizado en una carrera de 10 km.

Calcularemos la pendiente b y la ordenada al origen a mediante las fórmulas:

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

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

Obtenemos los siguientes resultados utilizando una calculadora o una computadora: $\bar{x}=52,99 \quad \bar{y}=43,70$

$$\sum_{i=1}^{14} (x_i - \bar{x})(y_i - \bar{y}) = -104,39 \qquad \sum_{i=1}^{14} (x_i - \bar{x})^2 = 223,11$$

Por lo tanto

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \frac{-104,39}{223,11}$$

$$= -0,468$$

$$a = 43{,}70 - (-0{,}4678)\ (52{,}99) = 68{,}49$$

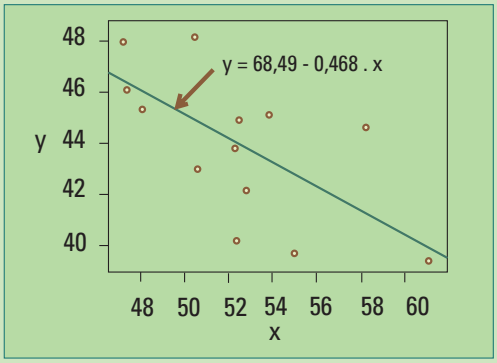


Fig. 22.16. *Diagrama de dispersión de los datos Atletas y la recta de regresión simple ajustada por el método de cuadrados mínimos.*

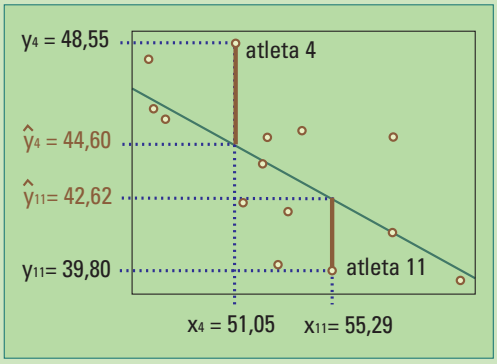


Fig. 22.17. *El punto correspondiente a la atleta 4 se encuentra por encima de la recta ajustada por cuadrados mínimos (su residuo es positivo). El de la atleta 11 por debajo, su residuo es negativo .*

La ecuación de la recta ajustada a los datos de las atletas (figura 22.19) es:

$$y = 68{,}49 - 0{,}468\ x$$

Algunos residuos son positivos y otros negativos (figura 22.17) :

Para la atleta 4, el punto en el diagrama de dispersión se encuentra **por encima de la recta.** El **residuo** es **positivo,**

$$y_4 - \hat{y}_4 = 48{,}55 - 44{,}60$$
$$= 3{,}95 > 0$$

$$\hat{y}_4 = 48{,}55$$
$$\hat{y}_4 = 68{,}49 - 0{,}468 \times 51{,}05$$
$$= 44{,}60$$

Para la atleta 11, el punto en el diagrama de dispersión se encuentra **por debajo de la recta.** El **residuo** es **negativo** y

$$y_{11} - \hat{y}_{11} = 39{,}80 - 42{,}62$$
$$= -2{,}82 < 0$$

$$\hat{y}_{11} = 39{,}80$$
$$\hat{y}_{11} = 68{,}49 - 0{,}468 \times 55{,}29$$
$$= 42{,}62$$

En las secciones siguientes veremos como la recta ajustada en un diagrama de dispersión de dos variables X e Y resume la relación entre las mismas

22.3.2. Significado de la recta

Veamos qué significa que dos variables **X** e **Y** tengan una relación "perfectamente lineal". Esto es, los pares de (x, y) de los valores de las variables **siempre** se encuentren sobre una recta.

La ecuación de una recta con pendiente b y ordenada al origen a es: y = a + b x. Cuando "x" aumenta una unidad "y" crece (o decrece) b unidades; a es el valor donde la recta corta al eje vertical (figura 22.18).

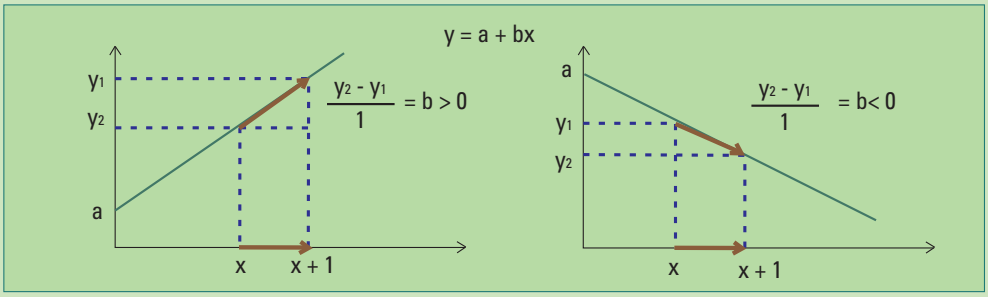


Fig. 22.18. *Dos rectas. La de la izquierda tiene pendiente positiva y la de la derecha pendiente negativa.*

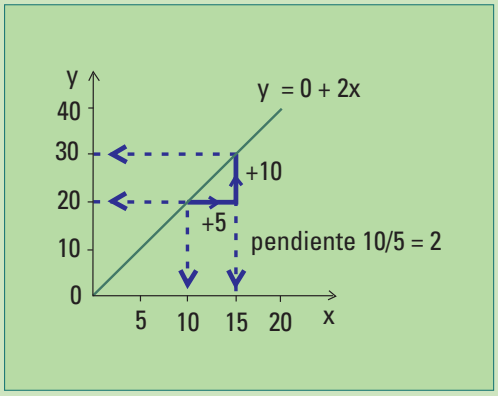


Figura 22.19. *Relación, hipotética, lineal entre la variable "vocabulario en miles de palabras" (y) y* ***la variable edad*** *(x).*

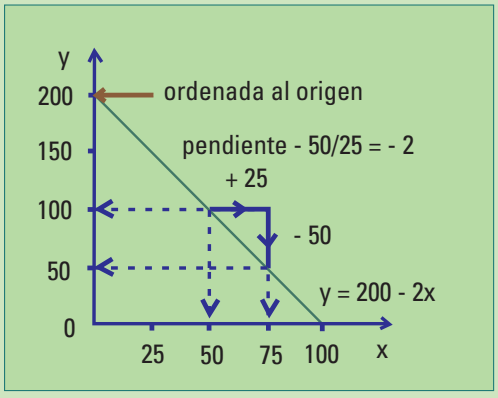


Figura 22.20. *Relación lineal entre la cantidad de palabras recordadas en una prueba* ***(y)*** *y la edad* ***(x).***

Ejemplo 4: Supongamos que la relación entre la variable "vocabulario en miles de palabras" (**Y, variable respuesta**) y la **variable edad** (X, variable **explicativa**) satisface la ecuación de la recta con **ordenada al origen a** = 0 y **pendiente b** = 2:

Esta recta corta al eje **Y** en el punto (0,0) y aumenta 2 unidades verticalmente (**Y**) con cada aumento horizontal (**X**) de 1 unidad. En términos de la relación entre la variable edad y el vocabulario esto significa que 0 es la cantidad inicial de palabras (a los 0 años) y que hay un **aumento de 2000** palabras **por año,** aproximadamente (7 por día). La pendiente indica la tasa de cambio. Como **la pendiente es positiva, la relación** entre el vocabulario y la edad **es creciente** (figura 22.19).

Ejemplo 5: En la figura 22.20 **y** representa la cantidad de palabras recordadas en una prueba y nuevamente **x** es la edad. La recta tiene ordenada al origen a = 200 y pendiente ***b = -2:***

y = 200 - 2 x ;

corta el eje Y en el punto (0,200) y cae 2 unidades verticalmente con cada aumento de 1 unidad en X. A medida que **aumenta la edad** en un año la **cantidad de palabras** que se recuerda en la prueba **disminuye** en 2.

Como **la pendiente es negativa,** la relación entre la cantidad de palabras recordadas y la edad **es decreciente.** La ordenada al origen, de 200 palabras para edad = 0, no tiene ningún significado biológico como ocurre muchas veces.

En general, **si la relación entre X e Y es perfectamente lineal** y conocemos los valores a y b, la ecuación y = a + b x permite predecir qué valor de Y corresponde a cualquier valor de X. Se trata de una **asociación determinística.** Más aún, dos pares de datos son suficientes para hallar los parámetros a y b, de la misma manera que **dos puntos y una regla alcanzan para dibujar una línea recta.**

22.3.3. Modelo

La relación entre datos reales rara vez es tan simple como la expresada por la ecuación de la recta.

Un modelo más realista plantea que **la media poblacional** de **Y**, más que los valores individuales, **cambia linealmente con X**

$$\mu_Y(x) = \alpha + ßx$$

En el caso las atletas (tabla 22.3), el modelo de regresión lineal dice que para **cada valor** de **la potencia aeróbica máxima** (**x**), **la media de los tiempos** (**Y**) en una carrera de 10 km, en la población de todas las atletas entrenadas, es:

Tiempo medio (depende de x) = **α** + **ß** . (Potencia aeróbica máxima)

Otras cosas, además de X, hacen que los valores individuales Y varíen alrededor de la media ($\boldsymbol{\mu_Y(x)}$) de todos los valores de Y cuando X toma el valor **x.** Por ejemplo, además del grado de entrenamiento (medido por la Potencia aeróbica máxima), **otros factores** causan que los **tiempos individuales varíen** alrededor de su media poblacional: **la edad, la longitud de piernas y brazos, la fuerza de piernas y brazos, la motivación por ganar,** etc.

Representamos esas "otras cosas" con **un término de error,** □ (epsilon). Es la diferencia entre un valor individual y la media de la variable **Y** en la población, para un valor fijo **x** de la variable explicativa:

$$□ = Y - \mu_Y(x)$$
$$□ = Y - (\alpha + ßx)$$

Despejando Y de la primera y de la segunda de las igualdades anteriores resultan: $Y = \mu_Y(x) + □$

y equivalentemente $Y = \alpha + ßx + □$

El valor de Y es igual a la media más un error, distinto para cada valor de la variable.

El **modelo de regresión lineal** permite que los **valores individuales** de la variable respuesta se encuentren alrededor de la recta de regresión y no necesariamente sobre ella.

Modelo de regresión lineal simple:

α y ß son parámetros fijos
$\square$ es aleatorio

recta
$\Uparrow$
$Y = \alpha + ßx + \square \Rightarrow$ otras cosas además de x

En la sección 15.5 vimos el siguiente **modelo de medición** (sin sesgo):

$$Y = \mu_Y + \text{error aleatorio},$$

donde Y representa un valor cualquiera de una variable definida en una población y μ es la media de todos los valores posibles de esa variable en la población considerada. Por ejemplo Y = tiempo que tarda un atleta cualquiera en una carrera de 10 km, μ_Y = media de los tiempos que tardan todas las atletas de una población. En este modelo no se tiene en cuenta la edad y se establece que el **tiempo de una atleta** es igual a la **media de los tiempos de toda la población** más un término llamado de **error.**

Observemos que el **modelo de regresión lineal:**

$$Y = \mu_Y(x) + \square$$

tiene una estructura similar al modelo de medición, pero ahora **la media poblacional depende linealmente de una variable explicativa** (x).

El término de error, en cualquiera de los dos modelos permite describir la variabilidad de las observaciones alrededor de la media poblacional (μ_Y o $\mu_Y(x)$). Muchas veces se utilizan curvas Normales para describir esa variabilidad.

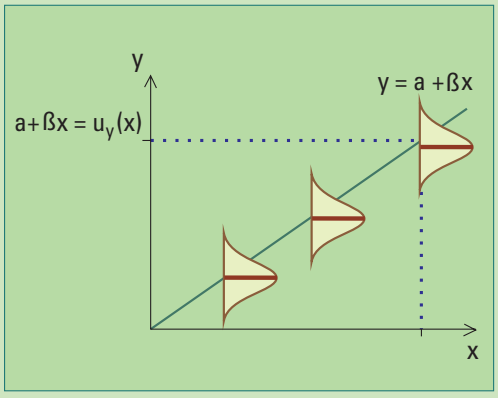


Figura 22.21. *Modelo de regresión lineal simple. Para cada valor de x, los valores de la variable Y se distribuyen alrededor de la media* $\mu_Y(x) = \alpha + ßx$.

Veamos qué ocurre con el modelo de regresión lineal:

$$y = \mu_y(x) + \square$$

Utilizamos la curva Normal para describir la variabilidad de las observaciones alrededor de la media (figura 22.21) y tenemos una media diferente para cada valor de la variable explicativa. Sobre la recta de regresión, el valor de y esta determinado por x; $y = \alpha + ßx$.

Se trata de medias poblacionales, pero no conocemos ni α ni ß.

La recta de regresión verdadera, $\mu_Y(x) = \alpha + ßx$, vincula la media poblacional de la variable respuesta con la variable explicativa. La ordenada al origen (α) y la pendiente (ß) de esta recta son desconocidos.

22.3.3.1 Interpretación de los coeficientes estimados

Los coeficientes a y b obtenidos por el método de cuadrados mínimos (sección 22.3.1) estiman los parámetros α y ß.

Ejemplo 6: Sigamos con los datos de la tabla 22.3 (Atletas) considerando al Tiempo (**Y**) como variable respuesta y a la Potencia aeróbica máxima (**X**) como variable explicativa. Vimos (ejemplo 3) que la ecuación de la recta ajustada a los datos de las atletas (figura 22.16) resulta y = 68,49 - 0,468 x.

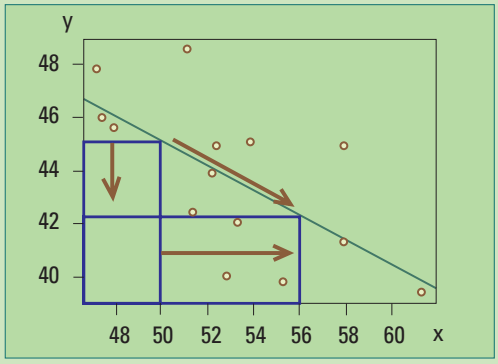
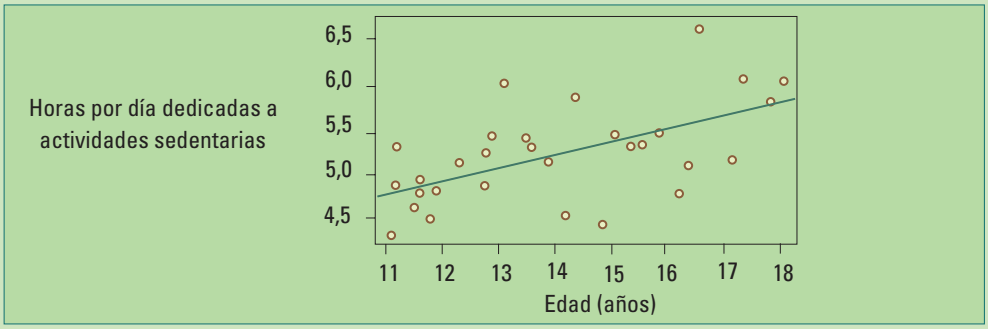


Fig. 22.22. *Sobre la recta de regresión ajustada a los datos "Atleta", un aumento de 6 unidades en x resulta en un reducción de 2,81 unidades en y.*

La pendiente ajustada es b = -0,468, esto significa **que sobre la recta ajustada,** un aumento de una unidad en x produce un descenso 0,468 unidades en y. Estimamos una reducción de la **media del tiempo** en una carrera de 10 km en 0,468 minutos, cada vez que la **Potencia aeróbica máxima aumenta una unidad.** Esta reducción del tiempo podría no tener importancia. Quizás sea más interesante la reducción del tiempo cuando la Potencia aeróbica máxima aumenta 6 unidades, esto es una **reducción** de 6 . 0,468 = 2,81 minutos en el tiempo de la carrera (figura 22.22).

a=68,49 y b=-0,468 son valores de estadísticos, es decir, estimaciones de los parámetros poblacionales α y ß.



22.23. *Diagrama de dispersión de las horas dedicadas a mirar televisión, estudiar o utilizar la computadora en función de la edad para 30 mujeres adolescentes, junto con la recta de regresión ajustada por cuadrados mínimos.*

Ejemplo 7: Consideremos nuevamente los datos de la tabla 22.1 horas por día dedicadas a mirar televisión, estudiar o utilizar la computadora, esta vez para las mujeres.

Tomando como variable respuesta Y = "horas por día dedicadas a mirar televisión, estudiar o utilizar la computadora" y variable explicativa **X** = edad, obtenemos la siguiente recta ajustada por cuadrados mínimos a los datos de la tabla 22.1 (mujeres):

$$Y = 3{,}24 + 0{,}13.\text{edad}$$

El signo de la pendiente de la recta ajustada **es positiva,** mostrando una asociación positiva entre las horas y la edad.

El valor de la pendiente es el **cambio de y cuando la variable explicativa aumenta una unidad.** En este caso se trata del aumento de las horas por cada año de aumento en la edad.

Recordemos que utilizamos la notación $\hat{y}(x)$ para indicar una ordenada obtenida a partir de la ecuación de la recta.

De la ecuación de la recta, estimamos una **media** poblacional **de 5,19 horas** dedicadas a actividades sedentarias entre todas las jóvenes de 15 años:

$$\begin{aligned}\hat{y}\ (\text{edad} = 15) &= 3{,}24 + 0{,}13 \times 15 \\ &= 3{,}24 + 1{,}95 \\ &= 5{,}19\end{aligned}$$

Para 16 años:

$$\begin{aligned}\hat{y}\ (\text{edad} = 16) &= 3{,}24 + 0{,}13 \times 16 \\ &= 3{,}24 + 2{,}08 \\ &= 5{,}32\end{aligned}$$

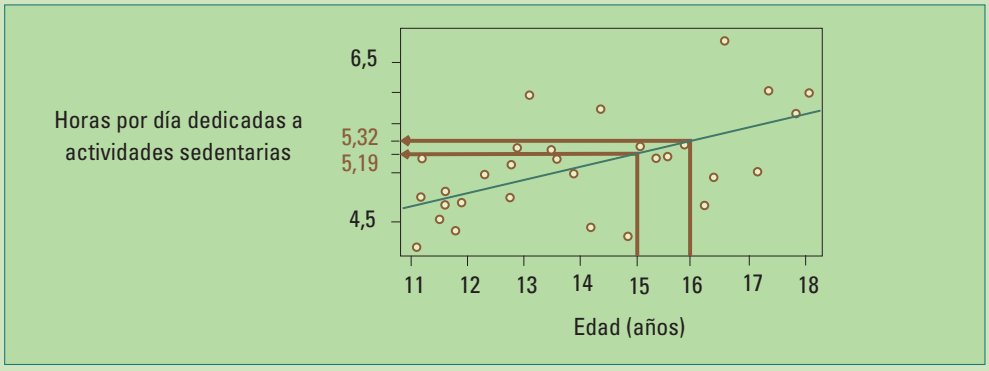


Figura 22.24. *Un aumento de un año corresponde a un aumento de 0,13 horas dedicadas a actividades sedentarias.*

Un aumento de un año en la variable explicativa corresponde a un aumento de 0,13 horas:

$$\begin{array}{r} \hat{y}\ (\text{edad} = 16) = 5{,}32 \\ \hat{y}\ (\text{edad} = 15) = 5{,}19 \\ \hline \hat{y}\ (\text{edad} = 16) - \hat{y}\ (\text{edad} = 15) = 5{,}32 - 5{,}19 \\ = 0{,}13 \end{array}$$

Obtendríamos el mismo resultado al comparar 16 con 17 años, 12 con 13, etc. O sea, **por cada año de aumento en la edad** se espera **un aumento** de 7,8 minutos (= 0,13 x 60) en el tiempo dedicado a actividades sedentarias en las mujeres.

Este aumento (0,13) es la pendiente de la recta ajustada.

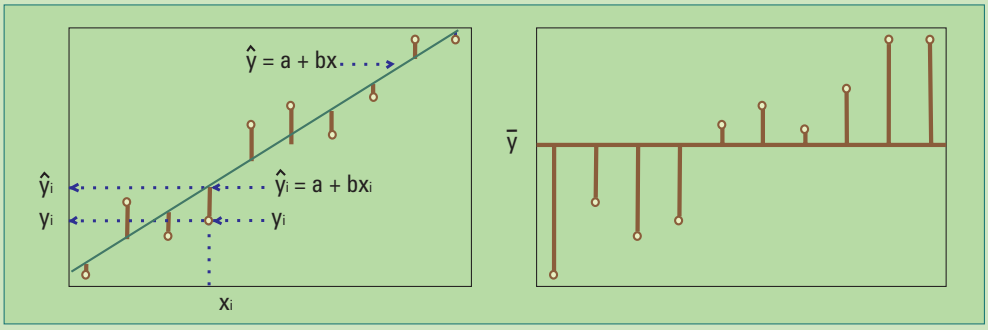
22.3.3.2. Interpretación de la recta ajustada. Coeficiente de determinación.

La recta de regresión ajustada en un diagrama de dispersión, y=a+bx , provee una estimación de la media poblacional de la variable respuesta en función de la variable explicativa.

> ¡ Recordemos que los coeficientes a y b de la recta estimada son números conocidos.
> En el ejemplo 6: y= 3,24 + 0,13 edad, por lo tanto a= 3,24 y b= 0,13

¿Qué significa $\hat{y}$ calculado como **a + bx** , para un valor determinado **x**? Es el **valor estimado** de la media poblacional de la variable respuesta para ese valor fijo de la variable explicativa.

¿Entonces $\hat{y}$ es la media muestral? No, pero juega un papel similar a la media muestral en el sentido de estimar una media poblacional, esta vez con un valor para cada valor de **x**.



22.25. *En el gráfico de la izquierda los residuos a la recta ajustada y=a+bx; a la derecha las distancias entre las respuestas observadas y su medida muestral ($\bar{y}$).*

¿Por qué no tomamos directamente el promedio de los valores de la variable **Y** , y los promediamos? Porque la información de la variable explicativa contenida en los coeficientes **a** y **b** provee una estimación más precisa que la media muestral $\bar{y}$, en cuyo cálculo sólo se utilizan los valores de la variable **Y**.

Los valores observados (y_i) están, en general más cerca de $\hat{y}_i$ que de □. La figura 22.25 ilustra esta situación. Las líneas verticales rojas muestran cómo se desvían los valores observados de la variable respuesta (y_i) respecto a la media muestral (□). Sus longitudes (y_i - □) al cuadrado son en promedio mayores que la de las líneas azules (□$_i$-**(a+bx$_i$)**)que miden las distancias de cada y_i a su correspondiente valor sobre la recta ajustada, $\hat{y}_i$ Recordemos que cada cada y_i se obtiene un valor $\hat{y}_i$ reemplazando x_i en la ecuación de la recta $\hat{y} = a+bx$; o sea $\hat{y}_i = a+bx_i$.

La figura 22.25 muestra como los valores observados (y_i) tienen una distancia vertical a la recta ajustada ($\hat{y} = a+bx$) menor que la recta horizontal de ecuación $y =$ □. El coeficiente de determinación, definido a continuación, cuantifica esa reducción.

El coeficiente de determinación (R^2) mide la proporción en que se reduce la suma de las distancias verticales al cuadrado de los valores observados (x_i, y_i) alrededor de la recta de cuadrados mínimos en comparación con esas distancias alrededor de la recta de ecuación $y =$ □:

$$R^2 = \frac{SCT - SCR}{SCT}$$

donde

- Suma de cuadrados total SCT. Mide las distancias verticales de los pares observados (x_i, y_i) a la recta de ecuación y = □

$$SCT = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- SCR es la suma de cuadrados de los residuos:

$$SCR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

El coeficiente de correlación al cuadrado coincide con el coeficiente de determinación:

$$r^2 = R^2$$

Ejemplo 8: Datos atletas. Cálculo del coeficiente de determinación.

DESARROLLO DEL CÁLCULO DEL COEFICIENTE DE DETERMINACIÓN. DATOS ATLETAS. TABLA 22.4

i	x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	y_i - □	$(y_i - \hat{y}_i)^2$	$(y_i$ - □$_i)^2$
1	47,17	47,83	46,42	1,41	4,13	1,99	17,06
2	47,41	46,03	46,31	-0,28	2,33	0,08	5,43
3	47,88	45,6	46,09	-0,49	1,9	0,24	3,61

i	x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$y_i -$ □	$(y_i - \hat{y}_i)^2$	$(y_i -$ □$_i)^2$
4	51,05	48,55	44,61	3,94	4,85	15,52	23,52
5	51,32	42,37	44,48	-2,11	-1,33	4,45	1,77
6	52,18	43,93	44,08	-0,15	0,23	0,02	0,05
7	52,37	44,9	43,99	0,91	1,2	0,83	1,44
8	52,83	40,03	43,78	-3,75	-3,67	14,06	13,47
9	53,31	42,03	43,55	-1,52	-1,67	2,31	2,79
10	53,93	45,12	43,26	1,86	1,42	3,46	2,02
11	55,29	39,8	42,62	-2,82	-3,9	7,95	15,21
12	57,91	44,9	41,4	3,5	1,2	12,25	1,44
13	57,94	41,32	41,38	-0,06	-2,38	0,00	5,66
14	61,32	39,37	39,8	-0,43	-4,33	0,18	18,75
						63,36	**112,22**

$R^2 = 0,4354$

□ = 43,699

$$SCT = \sum_{i=1}^{14} (y_i - \bar{y})^2 \quad \text{y} \quad SCR = \sum_{i=1}^{14} (y_i - \hat{y}_i)^2$$

$$= 112,22 \qquad = 63,36$$

$$R^2 = \frac{112,22 - 63,36}{112,22}$$

$$= 0,4354$$

La variabilidad de los datos de las atletas alrededor de la recta de regresión ajustada por cuadrados mínimos, y= 68,49 - 0,468 x, **se redujo en un 43,54%** en comparación con la variabilidad de la respuesta alrededor de su media muestral (recta y = 43,699).

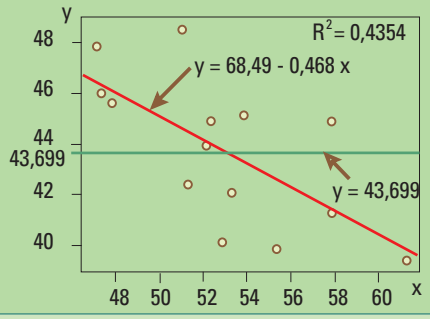


Figura 22.26. *Diagrama de dispersión datos atletas. En promedio las distancias a la recta de regresión son menores que a la recta y = 43,699.*

Recordemos que r = -0,659, su cuadrado es 0,4342. La diferencia con R^2 se debe únicamente a errores de redondeo de los que no nos preocupamos porque es habitual que estos cálculos se realicen con muchos más decimales utilizando computadoras.

100 x R^2 es la **reducción de la variabilidad** de los valores (y_i) de la **variable respuesta** a una recta de regresión lineal y = a+bx en comparación con la variabilidad de los valores (y_i) respecto de la recta y = $\bar{y}$.

Decimos que **la recta de regresión explica** el 100 x R^2 de la variabilidad de la variable respuesta.

R^2 = 0 cuando la regresión no explica nada; en ese caso, la suma de cuadrados total es igual a la suma de cuadrados de los residuos.

R^2 = 1 cuando **todos los puntos están sobre la recta, la variabilidad** observada de la respuesta **es explicada totalmente por la regresión** y la suma de los cuadrados de los residuos es cero.

22.3.4 El método de cuadrados mínimos puede fallar

Alcanza con un solo dato "malo" para **arruinar completamente el resultado** de la **media muestral** y el **desvío estándar muestral.** Consideremos por ejemplo el siguiente conjunto de datos: 0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16. Tiene media = 8 y desvío estándar = 5,05. Si el 16 es reemplazado por 1.600 (puede tratarse de un error o un valor de otra población) entonces el nuevo conjunto de datos tiene media = 101,18 y desvío estándar = 386,27.

En forma similar, **alcanza con un dato** "malo" para **arruinar completamente** a la recta de regresión al coeficiente de determinación y al coeficiente de correlación.

La figura 22.27 muestra dos diagramas de dispersión, los datos son los mismos en ambos diagramas salvo uno.

Para los datos del diagrama de la izquierda tenemos:

- Recta ajustada: **y = 0,99 + 1,99 x**
- Coeficiente de determinación: **R^2 = 0,9993**
- Coeficiente de correlación: **r = 0,99965**

La recta ajustada tiene pendiente positiva y representa a la mayoría de los datos. El coeficiente de determinación es R^2 = 0,9993. Vimos su significado general en la sección 22.3.3.2., en este ejemplo dice que la recta explica el 99,93 % de la variabilidad de los datos. Un resultado similar se obtiene mediante el coeficiente de correlación r: su valor (0,99965) tan cercano a 1 refleja la casi linealidad de los puntos en el diagrama de dispersión.

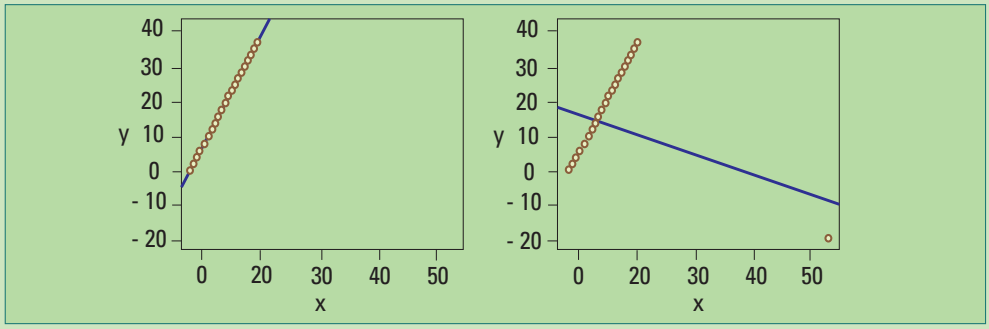
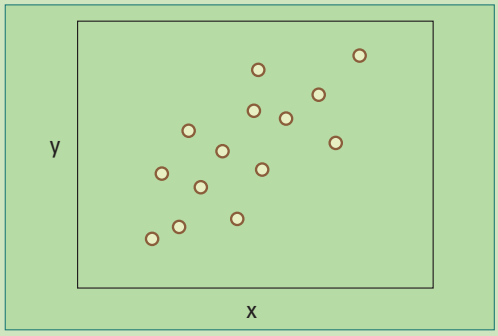


Figura 22.27. *Dos diagramas de dispersión y las rectas ajustadas. Un "punto palanca" distorsiona completamente el ajuste de la recta de cuadrados mínimos.*

Para los datos del diagrama de la derecha tenemos:

- Recta ajustada: **y = 17- 0,33 x**
- Coeficiente de determinación: **R^2 = 0,09226**
- Coeficiente de correlación: **r = -0,30374**

Agregando un "punto palanca" el ajuste por cuadrados mínimos cambia completamente. Se trata de un punto que se encuentra alejado de la mayoría de los valores de la variable respuesta. La recta ajustada tiene pendiente negativa y su dirección es casi perpendicular a la de la mayoría de los datos. El coeficiente de correlación dice que la asociación entre las variables es negativa cuando en realidad, salvo por un dato la asociación es positiva.



22.28. *Diagrama de dispersión con forma de nube.*

El ajuste de una recta y el cálculo del coeficiente de correlación serán buenas medidas resumen de la relación entre dos variables continuas, siempre que el diagrama de dispersión tenga una forma de nube, como el de la figura 22.28, más o menos concentrada alrededor de una recta.

□ 22.4. Dos rectas

Muchas veces interesa comparar la relación de dos variables continuas en dos grupos distintos definidos por una variable categórica.

Ejemplo 9: Consideremos nuevamente los datos de la tabla 22.1. Esta vez, tendremos en cuenta la edad al comparar las horas por día dedicadas a las actividades sedentarias (mirar televisión, estudiar o utilizar la computadora) entre varones y mujeres (tabla 22.1). Describiremos la relación entre la cantidad de horas y la edad en dos grupos; la variable categórica que los define es "género" y las categorías son: varón, mujer.

La figura 22.29 muestra el diagrama de dispersión de las horas por día dedicadas a actividades sedentarias para varones y mujeres y dos rectas de regresión. La recta azul fue estimada utilizando únicamente los datos de los varones y la roja los datos de las mujeres. Sus ecuaciones son:

- y = 2,84 + 0,197.edad, para varones
- y = 3,24 + 0,13.edad, para mujeres

Cada una de esas rectas estima la media en la población de la cantidad de horas que un varón (o una mujer) dedican a actividades sedentarias.

Sin tener en cuenta la edad, la diferencia de medias muestrales (5,73-5,06) de la cantidad de **horas** para varones y mujeres es **0,67 horas.**

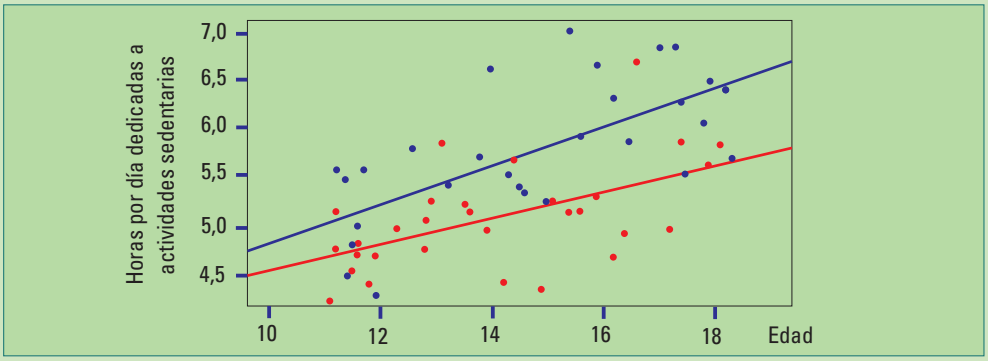


Figura 22.29. *Diagrama de dispersión de las horas por día dedicadas a actividades sedentarias en dos grupos: varones (en azul) y mujeres (en rojo) junto con sus respectivas rectas de regresión lineal simple.*

Teniendo en cuenta la edad. Las ecuaciones de las rectas de regresión simple, igual que la media muestral, proveen una estimación de la media poblacional, una para varones y otra para mujeres. Además esta vez se obtiene una estimación para cada valor de la variable explicativa (edad). Podemos estimar la misma diferencia que en el párrafo anterior para cada edad:

Media estimada de horas _ Varones = 2,84 + 0,197.edad
Media estimada de horas _ Mujeres = 3,24 + 0,13.edad

Diferencia de medias estimada = (2,84 + 0,197.edad) - (3,24 + 0.13 . edad)
= - 0,40 + 0,067.edad

Si edad = 12, resulta: Diferencia de medias estimada = -0,40 + 0,067 x 12
= 0,404

Si edad = 17, resulta: Diferencia de medias estimada = - 0,40 + 0,067 x 17
= 0,739

La figura 22.30 ilustra estas diferencias. Si edad = 12 obtenemos una estimación menor a 0,67 horas que se obtuvo simplemente restando las medias muestrales, si edad es 17 la diferencia es mayor.

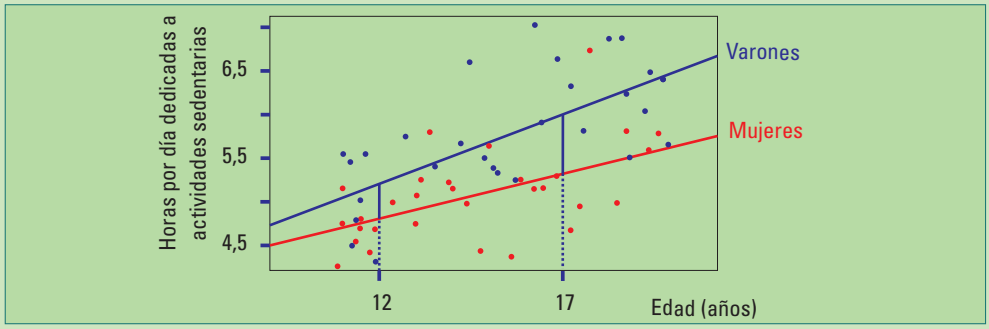


Figura 22.30. *A medida que aumenta la edad la diferencia de la cantidad de horas dedicadas a actividades sedentarias entre varones y mujeres también aumenta.*

□ 22.5. Cuantificación de la relación entre dos variables categóricas

Muchas veces interesa estudiar la relación entre dos variables categóricas. En este caso no se puede usar la palabra "correlación" para describirla. La correlación es un caso especial de asociación, mide la fuerza de la relación lineal entre las variables numéricas. El término adecuado para variables categóricas es simplemente "asociación".

Consideremos, en general, dos variables categóricas con dos categorías cada una:

- "grupo de tratamiento" (sí, no)
- "resultado" (sí, no)"

Estarán asociadas si el porcentaje de pacientes con resultado (sí) en un grupo (sí) es muy diferente del porcentaje de pacientes con ese mismo resultado en el otro grupo (no).

Ejemplo 10: La tabla 22.5 muestra los resultados de una encuesta hipotética para determinar si existe una asociación entre comer rápido y el sobrepeso. Se obtuvieron los siguientes resultados:

Resultados de una encuesta hipotética para determinar si existe una asociación entre comer rápido y el sobrepeso. TABLA 22.5

		Come rápido	
		si	no
Sobrepeso	si	75	22
	no	175	198
		250	220
% con sobrepeso		30 =100x75/250	10 = 100x22/220

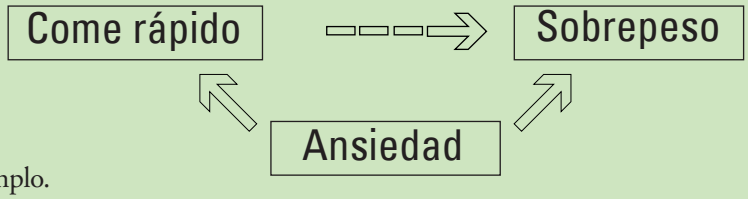
Las variables categóricas son:
Come rápido con categorías: si, no
Sobrepeso con categorías sí, no

En el grupo de individuos que come rápido el 30 % tiene sobrepeso; entre los que no comen rápido ese porcentaje se reduce al 10 %.

Los porcentajes de individuos con sobrepeso en las dos categorías de la variable come rápido son bastante diferentes, decimos que las dos variables están asociadas. En la sección 23.4.4 veremos si esa diferencia puede atribuirse a la variabilidad resultante del muestreo o es suficiente para establecer una asociación entre las variables.

□ 22.6. Causalidad

Es muy tentador considerar una evidencia sobre asociación como una evidencia sobre causalidad. En el ejemplo 10 podríamos pensar que comer rápido es una causa del sobrepeso porque están asociados. Sin embargo, podría ocurrir que la ansiedad sea causante de comer rápido y también de comer demás y por lo tanto tener sobrepeso. La asociación entre comer rápido y sobrepeso es consecuencia de la ansiedad (una causa común a ambas).



Veamos otro ejemplo.

Ejemplo 11: Se realizó una encuesta para estudiar la relación entre la preferencia por el dulce de leche y el alfajor de dulce de leche. Se seleccionaron al azar 10 personas y se les solicitó que asignaran un número entre 0 y 100 a su preferencia. Donde 0 indica que a la persona no le gusta el dulce de leche o el alfajor y 100 que le apasiona.

La figura 22.31 muestra el diagrama de dispersión de los datos de la tabla 22.6. Vemos una clara tendencia lineal. Las personas que asignaron un puntaje alto al dulce de leche, también lo hicieron para el alfajor de dulce de leche. La gente no asignó los mismos puntajes en ambas escalas pero la preferencia por el dulce de leche es similar en rasgos generales a la del alfajor, aunque para este último un poco menor.

La figura 22.32 muestra los mismos datos junto con la recta de regresión lineal ajustada por cuadrados mínimos y = -1,82 + 0,83 x. Cada vez que el puntaje del dulce de leche aumenta en 10 unidades, estimamos un aumento promedio de 8,3 unidades (10 x 0,83) para el puntaje del alfajor de dulce de leche.

PUNTAJE POR REFERENCIAS SOBRE EL DULCE DE LECHE Y EL ALFAJOR DE DULCE DE LECHE. TABLA 22.6

Persona	Dulce de leche	Alfajor de dulce de leche
1	100	85
2	90	80
3	60	44
4	60	50
5	20	25
6	95	70
7	90	80
8	70	50
9	50	30
10	10	5

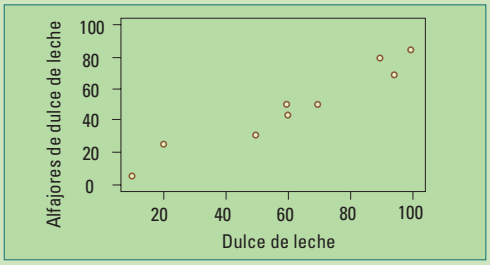


Figura 22.31. *Diagrama de dispersión del puntaje asignado por preferencia al dulce de leche y al alfajor de dulce de leche, 0 indica que a la persona no le gusta y 100 que le apasiona.*

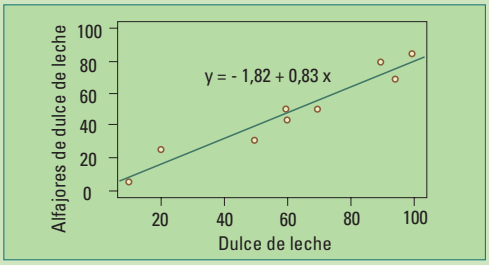


Figura 22.32. *Diagrama de dispersión del puntaje asignado por preferencia al dulce de leche y al alfajor de dulce de leche, junto con la recta de regresión lineal ajustada por cuadrados mínimos.*

El coeficiente de correlación **r** del puntaje (tabla 22.6) es 0,97. Este número es muy cercano a 1 (el valor máximo posible de **r**), mostrando que "la preferencia por el dulce de leche" y " la preferencia por el alfajor de dulce de leche" tienen un altísimo grado de asociación lineal. ¿Significa esto que la preferencia por el dulce de leche es la causa de la preferencia por el alfajor de dulce de leche? No necesariamente.

Las personas que le asignaron un puntaje alto al dulce de leche y al alfajor de dulce de leche pueden ser justamente las personas a las que les gusta la comida dulce y blanda. En este caso, la preferencia por ese tipo de comida sería la **causa común,** la que produce un alto grado de correlación entre las preferencias estudiadas.

En general, hallar una asociación entre variables es sólo un indicio. Si interesa establecer causalidad deberán realizarse estudios posteriores para confirmar o descartar las sospechas. Idealmente debería realizarse un experimento comparativo aleatorizado. La aleatorización produce grupos de sujetos, similares al comienzo de los tratamientos. Al comparar los grupos nos aseguramos que las diferencias observadas se deban a los efectos del tratamiento.

Pero muchas veces esto no es posible. Es difícil obligar a un sujeto a comer rápido o despacio. Sin embargo, sí es posible realizar un estudio observacional comparativo con grupos diferentes respecto del factor que se desea estimar y, lo más parecidos posibles en el resto. Si se sospecha que el nivel de ansiedad influye en el sobrepeso e interesa estudiar el efecto de comer rápido sobre el sobre peso, los grupos deberían ser similares en cuanto a nivel de ansiedad y también respecto de cualquier otra variable conocida que pueda afectar el resultado (edad, género, peso inicial, etc.).

□ 22.7. Más allá de un conjunto de datos

Cuando se ajusta una recta a pares de valores en un diagrama de dispersión el interés puede estar en, simplemente, obtener un resumen de la relación entre los puntos del diagrama, de la misma manera como puede interesar conocer un promedio o una proporción de un conjunto de datos. Este enfoque es llamado de **estadística descriptiva.** Se trata de obtener números que describen en forma resumida un conjunto de datos.

Pero si ese conjunto de datos proviene de un muestreo aleatorio simple de una población e **interesa** describir **el comportamiento de las variables en la población** (como ocurre la mayoría de las veces), la recta ajustada es una de las tantas rectas que se pueden obtener con diferentes muestras para el mismo problema. Se trata ahora de un problema de **inferencia estadística,** porque el interés ya no está centrado en el conjunto de datos que tenemos sino en toda la población de la cual provienen. En este caso **a, b** y **a+bx** son estimaciones de **α**, **ß**, y **α+ß** x respectivamente y **r** es una estimación de la correlación lineal de todos los valores las variables en la población. Como toda estimación tiene errores, por haber sido calculada a partir de una muestra en vez utilizar datos de toda la población. Se trata de **errores aleatorios debido al muestreo.**

Nos preguntamos qué pasaría si tomásemos muchas muestras de la misma población, ¿cómo cambiarían las rectas ajustadas?

Ya presentamos este enfoque en el capítulo 10 para proporciones cuando tratamos el "margen de error". En el capítulo 23, retomaremos el tema con más profundidad para medias muestrales y proporciones y daremos fórmulas de cálculo para los errores de estimación. Las fórmulas de cálculo de los errores de los coeficientes de la recta de regresión por cuadrados mínimos son matemáticamente más complejas, pero los conceptos estadísticos son similares a los que veremos en ese capítulo.

22.8. Actividades y ejercicios

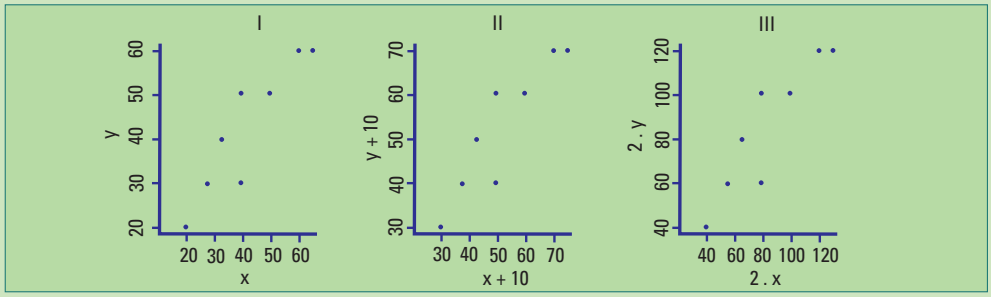
1.
a) ¿Cuáles son los todos los valores posibles del coeficiente de correlación?
b) ¿Cuáles son los todos los valores posibles del desvío estándar muestral **s**?

2. Muestre que

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

En los ejercicios 3 - 8 elija la respuesta correcta o la que completa la frase.

3. ¿Cuál de los siguientes tres diagramas de dispersión tiene mayor coeficiente de correlación?



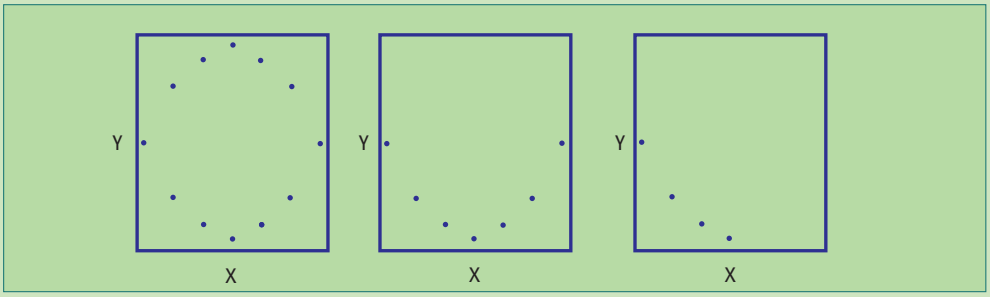
a) El I
b) El II
c) El III
d) Todos tienen el mismo coeficiente de correlación
e) No se puede responder a la pregunta porque falta información

4. Supongamos que el coeficiente de correlación es 0,7. Entonces, dados dos puntos del diagrama, ¿cuál de las siguientes situaciones es posible?

I) el punto que se encuentra a la izquierda (o sea con menor x) tiene un valor menor de y que el que se encuentra a la derecha.
II) el punto que se encuentra a la izquierda (o sea con menor x) tiene un valor mayor de y que el que se encuentra a la derecha.

a) Sólo I
b) Sólo II
c) I y II

5. ¿Cuál de las afirmaciones sobre los coeficientes de correlación de los puntos de los siguietes diagramas de dispersión son verdaderas?



a) Todos son cero.
b) Uno es cero, otro es positivo y otro es negativo.
c) Dos son cero y otro es cercano a -1.
d) Dos son cero y otro es -1.
d) Ninguno es cero.

6. Supongamos que la recta de regresión ajustada a un conjunto de datos $\mathbf{y} = 2+\mathbf{bx}$ pasa por el punto (3,11). Si $\bar{\mathbf{x}}$ e □ son las medias muestrales de los valores **x's** e **y's** respectivante, entonces □ =

a) $\bar{\mathbf{x}}$
b) $3\bar{\mathbf{x}}+2$
c) $\bar{\mathbf{x}}+2$
d) $2\bar{\mathbf{x}}-3$
e) $2\bar{\mathbf{x}}+3$

7. Un estudio determinó que el coeficiente de correlación entre el puntaje que los alumnos asignaron a sus profesores en una encuesta y el puntaje que la directora de la escuela les asignó a los mismos profesores es **r** = 1,25. Esto significa que

a) La directora y los alumnos coinciden en respecto a qué es un buen profesor.
b) La directora y los alumnos tienden a no estar de acuerdo respecto a qué es un buen profesor.
c) Hay poca relación entre los puntajes.
d) La asociación entre ambos puntajes es fuerte.
e) Hay un error de cálculo

8. El coeficiente de correlación satisface:

I. No es afectado por cambios en las unidades en que se miden las variables.

II. No es afectado por intercambiar las variables que se ponen en x e y.
III. No es afectado por la presencia de valores atípicos.

a) I y II
b) I y III
c) II y III
d) I, II y III
e) Ninguna de las afirmaciones es correcta.

9. ¿Le dedican los varones de su escuela más horas a realizar actividades sedentarias que las mujeres? Realice una encuesta para responder esta pregunta.