CORRELACIÓN Y REGRESIÓN LINEAL

En ocasiones queremos comparar dos variables entre sí, para contestar preguntas tales como:

- ¿El nivel de carencias de tipo social[1] de una región está relacionado con el nivel de desempleo en dicha región?

- ¿La cantidad de unidades de alcohol que beben los estudiantes en promedio a la semana está relacionada con la frecuencia que faltan a su clase de métodos estadísticos debido a una cruda?

Asimismo, podríamos estar interesados en predecir el valor de una variable a partir de otra, por ejemplo:

- Si una región tiene una taza de desempleo del 13%, ¿qué tantas carencias de tipo social esperamos que haya ahí?
- Si un estudiante bebe 25 unidades de alcohol en promedio a la semana, ¿cuántas clases es probable que falte durante el semestre?

Para resolver preguntas como las arriba planteadas podemos usar la correlación y la regresión.

La **correlación** mide la asociación entre dos variables, esto es la fuerza de la relación entre los valores de dos variables.

La **regresión** es la predicción de los valores de una variable a partir de los valores de otra variable.

Correlación

La correlación mide la asociación entre dos variables *continuas*. Podríamos encontrar que dos variables están estrechamente asociadas, sin embargo, eso no significa que exista una relación de causalidad entre ellas. Si el valor de una variable se incrementa en la medida en que el valor de la otra variable se incrementa (o si decrece), esto no significa necesariamente que una variable explica a la otra.

Diagramas de dispersión
(*scatter graph/scatter plot/scattergram*)

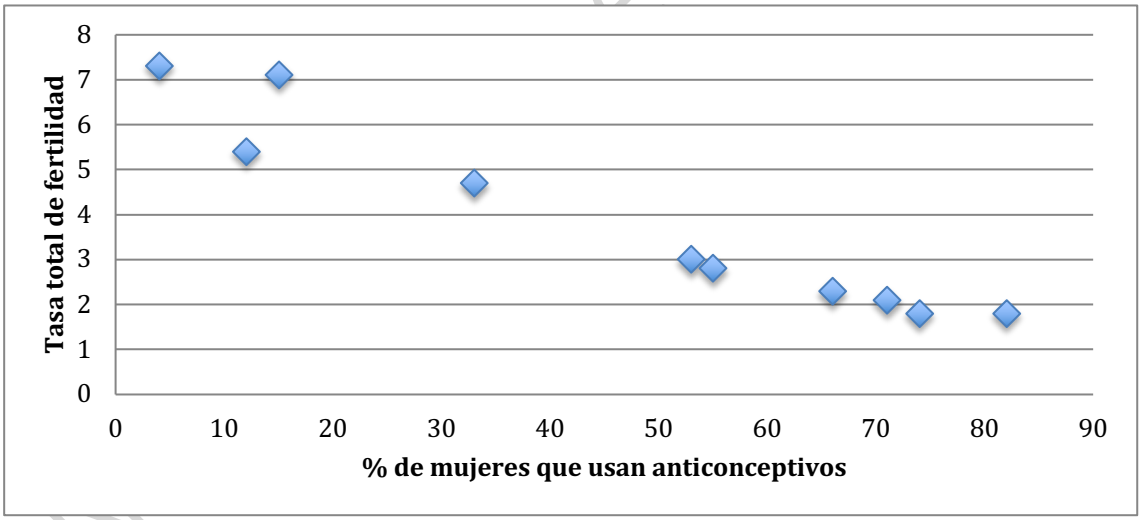
Al tratar de examinar la asociación entre dos variables, hay que graficar usando un diagrama de dispersión antes de hacer otra cosa.

[1] En México, la carencia social se mide por los siguientes índices: (1) rezago educativo, (2) acceso a los servicios de salud, (3) acceso a la seguridad social, (4) calidad y espacios de la vivienda, (5) acceso a los servicios básicos de la vivienda, y (6) acceso a la alimentación.

En el cuadro siguiente tenemos información de la tasa total de fertilidad total y los porcentajes de mujeres que usaron algún método anticonceptivo en diez países. La tasa total de fertilidad es el número de niños que una mujer esperaría tener a lo largo de su vida si los niveles mostrados se mantuvieran. Estos datos se pueden presentar en un diagrama de dispersión, como se muestra después del cuadro.

Cuadro 1. Tasa total de fertilidad y prevalencia de métodos anticonceptivos en 10 países

País	Tasa total de fertilidad, 1994	% de mujeres usando métodos anticonceptivos (1987-1994)
Gran Bretaña	1.8	82
Estados Unidos	2.1	71
Gambia	5.4	12
Indonesia	2.8	55
México	3.0	53
Brasil	2.3	66
Uganda	7.1	15
Eslovaquia	1.8	74
Nigeria	7.3	4
Botsuana	4.7	33

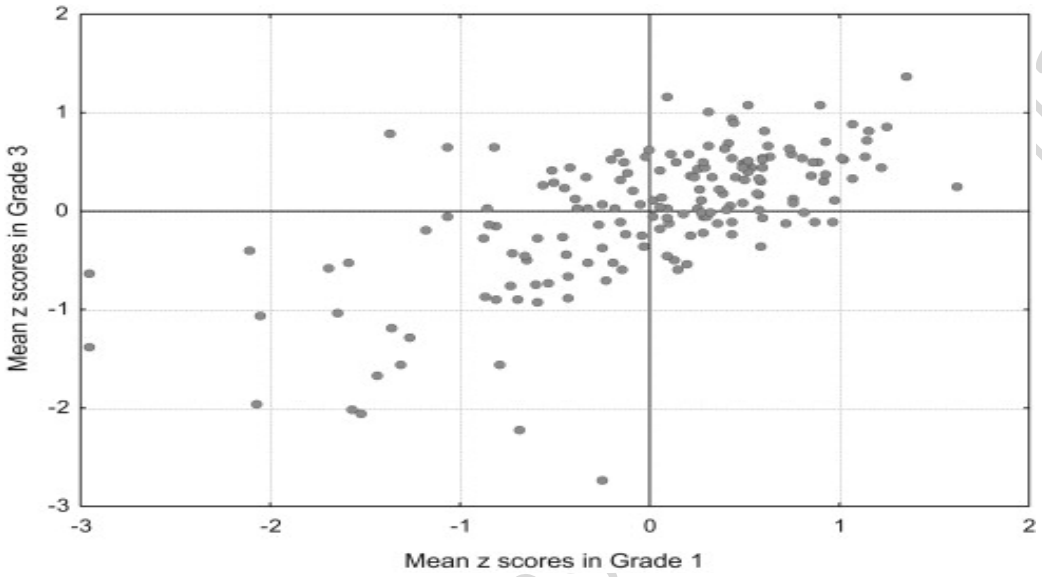


Gráfica 1. Tasa total de fertilidad y % de mujeres que usan anticonceptivos en 10 países

En este ejemplo la variable de respuesta es el número de hijos porque depende del uso de anticonceptivos, que es la variable independiente o explicativa. En algunos casos no es claro cuál es cuál o no importa. En el caso de las gráficas de dispersión sí es importante puesto que la variable dependiente siempre debe ir en el eje Y y la variable independiente en el eje X.

La gráfica 1 muestra que, en general, los países donde hay un alto porcentaje de mujeres que usan algún método anticonceptivo tienen tasas totales de fertilidad bajas, mientras que los países donde la prevalencia de anticonceptivos es baja tienden a tener una tasa alta de fertilidad.

Si se hace una gráfica de variables que han sido estandarizados (puntajes Z) y debido a que los puntajes z son positivos y negativos, es necesario una gráfica de forma cruzada.



La interpretación de las gráficas de dispersión

A continuación, se muestran varias gráficas de dispersión que muestran diferentes tipos de correlación. Puede haber: (1) correlación positiva, (2) correlación negativa y (3) sin correlación.

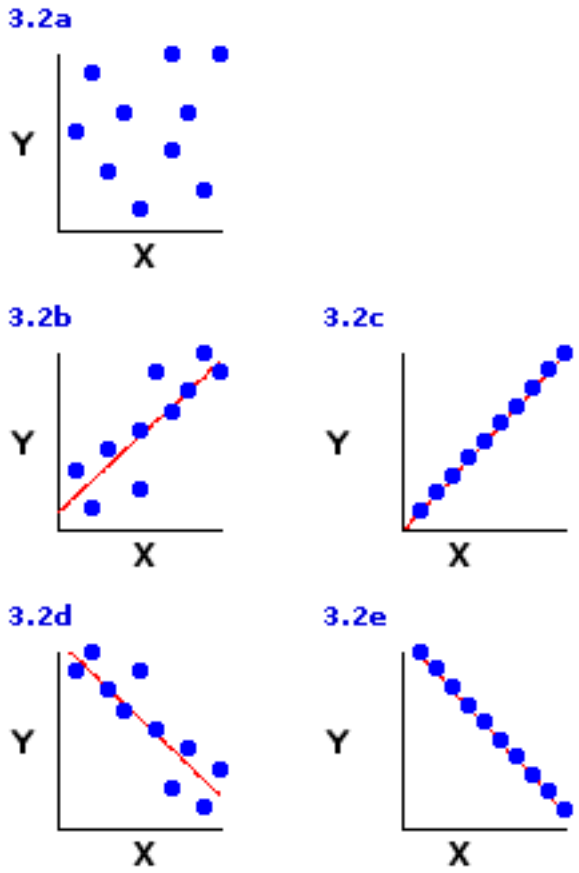
Cuando ***no hay asociación o correlación*** la gráfica se vería como 3.2a. Esto es, no hay una tendencia sistemática que muestre que X e Y estén asociados en ninguna dirección. Esto es los puntos se encuentran esparcidos de forma azarosa en toda la gráfica.

En caso de ***correlación positiva***, se mostrará una tendencia a valores altos de X que se asociarán con valores altos de Y o viceversa. En este caso los puntos tenderán a alinearse hacia arriba en una diagonal inclinada como se observa en la gráfica 3.2b. Un ejemplo sería el PIB y el número de televisiones por X número de población. Se esperaría que un país más rico tendría más bienes por habitante.

La gráfica para ***correlación negativa*** mostrará una tendencia opuesta para valores altos de X estén asociados con valores bajos de Y y viceversa. Entonces, los puntos tenderán a alinearse a lo largo de una diagonal inclinada hacia abajo, como se observa

en la gráfica 3.2d. Un ejemplo de correlación negativa sería el valor del carro y la edad: entre más viejos su valor disminuirá.

Existen, además dos casos mostrados en las gráficas 3.2c y 3.2e donde los puntos se alinean a lo largo de la diagonal como en fila. Esta forma se conoce con el nombre de ***correlación perfecta***, que representaría el grado máximo de correlación lineal ya sea positiva o negativa



EL COEFICIENTE DE CORRELACIÓN

Para medir la fuerza de la asociación podemos calcular el coeficiente de correlación de Pearson, también conocido como el coeficiente de correlación *r*.

$$r = \frac{\sum [(X_i - \bar{X})(Y_i - \bar{Y})]}{\sqrt{[\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2]}}$$

PASOS PARA CALCULAR EL COEFICIENTE DE CORRELACIÓN

Todos los pasos corresponden a los datos de la tasa total de fertilidad y el uso de anticonceptivos en 10 países. El procedimiento es el siguiente:

Paso 1

Escribir todas las observaciones para las dos variables X e Y en dos columnas. Recuerde que X corresponde a la prevalencia de uso de anticonceptivos y que Y es la tasa total de fertilidad.

Paso 2

Para cada variable, agregar los todos los valores de las observaciones y dividirlos por el total de número para obtener las medias X, Y al final de cada columna.

Paso 3

En las columnas C y D, calcular los residuales $(X - \bar{X})$ y $(Y - \bar{Y})$ restando la media a cada uno de los valores.

Paso 4

En las columnas E y F, sacar el cuadrado de los resultados obtenidos en las columnas C y D para obtener $(X - X)^2$ y $(Y - Y)^2$.

Paso 5

En la columna G, multiplicar los valores de la columna C por los valores de la columna D para obtener $(X - X)(Y - Y)$.

Paso 6

Para las columnas E, F y G sumar todos los valores en cada columna para obtener un total al final de cada columna. Estos son los valores que nos servirán para la ecuación.

OJO: un error muy común es saltarse el paso representado en la columna G y obtener una respuesta de 0 para la parte de arriba de la ecuación al pensar que la suma de los residuales X es de 0 y la suma de los residuales de Y es de 0 y por ende 0 x 0 es igual a 0. De hecho, necesitamos multiplicar cada residual X por su correspondiente residual Y para posteriormente sumar estas respuestas (como en la columna G). La respuesta está muy lejos de ser de 0.

A	B	C	D	E	F	G
X	**Y**	**$(X - \bar{X})$**	**$(Y - \bar{Y})$**	**$(X - \bar{X})^2$**	**$(Y - \bar{Y})^2$**	**$(X - \bar{X})(Y - \bar{Y})$**
82	1.8	35.5	-2.03	1260.25	4.12	-72.07
71	2.1	24.5	-1.73	600.25	2.99	-42.39
12	5.4	-34.5	1.57	1190.25	2.46	-54.17
55	2.8	8.5	-1.03	72.25	1.06	-8.76
53	3.0	6.5	-0.83	42.25	0.69	-5.40
66	2.3	19.5	-1.53	380.25	2.34	-29.84
15	7.1	-31.5	3.27	992.25	10.69	-103.01
74	1.8	27.5	-2.03	756.25	4.12	-55.83
4	7.3	-42.5	3.47	1806.25	12.04	-147.48
33	4.7	-13.5	0.87	182.25	0.76	-11-75
$\bar{X}$ =46.5	$\bar{Y}$=3.83		Totales	7282.50	41.27	-530.70

Ahora, con los datos obtenidos:

$$r = \frac{-530.70}{\sqrt{(7282.50 \times 41.27)}} = -0.97$$

¿Qué significa un valor de *r* = -0.97?

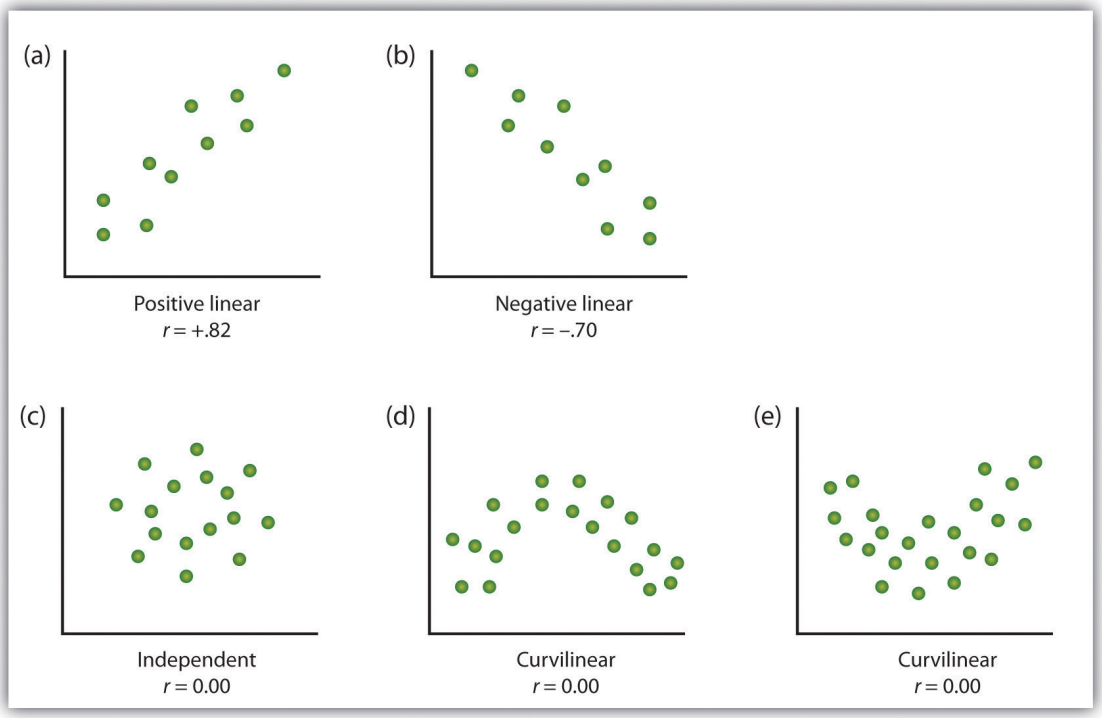
El coeficiente de correlación r toma valores entre +1 y -1. Los diferentes valores r significan lo siguiente:

r = +1: correlación positiva perfecta (todos los puntos en línea derecha).
r entre 0 y 1: correlación positiva (pero no perfecta).
r = 0: no hay asociación entre las variables
r entre -1 y 0: correlación negativa (pero no perfecta).
r = -1: correlación negativa perfecta (todos los puntos en línea derecha)

Entre más se acerque ***r*** a +1 o -1 más fuerte será la relación entre las dos variables. Los valores más cercanos a 0 indican una relación muy débil.

En el ejemplo anterior el resultado de *r* fue de -0.97, lo cual indica una correlación negativa muy fuerte entre la prevalencia en el uso de anticonceptivos y la tasa total de fertilidad (como se esperaba a partir de la gráfica de dispersión). En las ciencias sociales un valor tan cercano a ±1 es extremadamente raro.

Por último, el coeficiente de correlación mide una ***asociación lineal***: qué tan cerca se ubican los puntos con respecto de una línea derecha. Sin embargo, con ciertas relaciones, los datos siguen un *patrón curvado* (como se aprecia en las gráficas (d) y (e)). Esto sería una relación curvilínea. En tales casos la *r* no es una buena medida de la fuerza de la asociación entre las variables. Por eso es importante hacer la gráfica de dispersión antes de hacer los cálculos.



REGRESIÓN

Predecir Y a partir de X

Cuando describimos sólo una variable, podemos resumir los datos usando medias, medianas, etc. Pero si lo que tenemos son dos variables, podemos resumir su relación usando la ecuación de una línea recta.

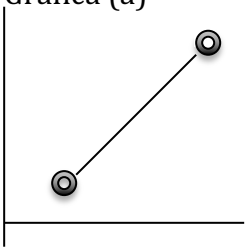
Si X e Y son dos variables, la ecuación para una línea recta es:

$$Y = a + bX$$

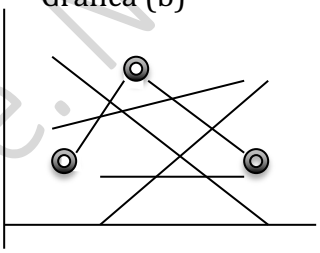
Donde *a* es la intersección de Y y *b* es el gradiente.

Si calculamos los valores de a y b, está ecuación puede usarse para predecir el valor de la variable Y a partir de un valor dado de la variable X. El problema es, ¿cómo podemos decidir cuál es la mejor línea recta con respecto de una serie de datos? Si tenemos dos puntos, la respuesta es fácil (como se aprecia en la gráfica (a)), pero con 3 o más puntos hay varias posibilidades (como se muestra en la gráfica (b)).

Gráfica (a)

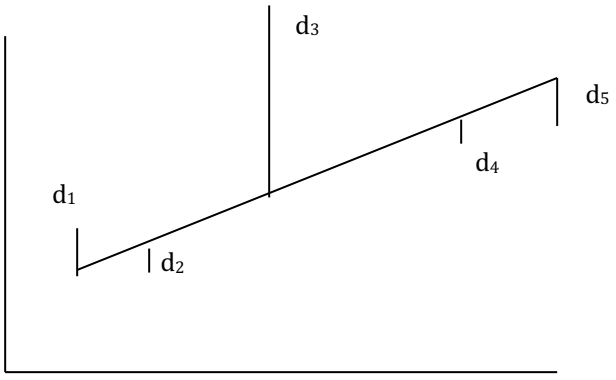


Gráfica (b)



Cuando la correlación es positiva o negativa pero débil es muy difícil juzgar dónde debería ir la línea. Así que tenemos que calcular la línea que sea la mejor o la más adecuada. Esta línea se conoce con el nombre de regresión lineal.

La línea que sea la mejor candidata será la línea que minimice la distancia vertical entre la línea y todos los puntos. En términos técnicos, la mejor línea que minimicé $\sum d_i^2$, donde d_i son las distancias verticales entre la línea y cada punto, como se muestra en la gráfica siguiente.



Hasta ahora ya tenemos calculado todos los números que necesitábamos para el coeficiente de correlación. Necesitamos un poco más.

El valor de *b* para la anticoncepción y la fertilidad se calcula así:

$$b = \frac{-530.70}{7282.50} = -0.07 \quad \left[b = \frac{\sum G}{\sum E} \right] = b = \frac{\sum (X_1 - \bar{X})(Y_1 - \bar{Y})}{\sum (X_1 - \bar{X})^2}$$

Este valor de b, junto con las dos medias $\bar{X}$ y $\bar{Y}$, es usado en la ecuación para encontrar el valor de a:

$$a = \bar{Y} - b\bar{X}$$

$$a = 3.83 - (-0.07 \times 46.5) = 7.09$$

Ahora conocemos los valores de *a* y *b*. Estos son conocidos como los coeficientes de regresión. La ecuación de una línea recta es Y = a + bX, así que la ecuación para la línea de regresión para nuestros datos será de:

$$Y = 7.09 + (-0.07\ X)$$

En esta ecuación, los signos [+] y [–] se cancelan mutuamente para dar como resultado un signo [-], así que podemos escribir la ecuación así:

$$Y = 7.09 - 0.07\ X$$

La variable X representa el porcentaje de uso de anticonceptivos y la variable Y representa la tasa total de fertilidad (TTF), así que:

TTF = 7.09 + (– 0.07) x porcentaje de uso de anticonceptivos
o
TTF = 7.09 – 0.07 x porcentaje de uso de anticonceptivos

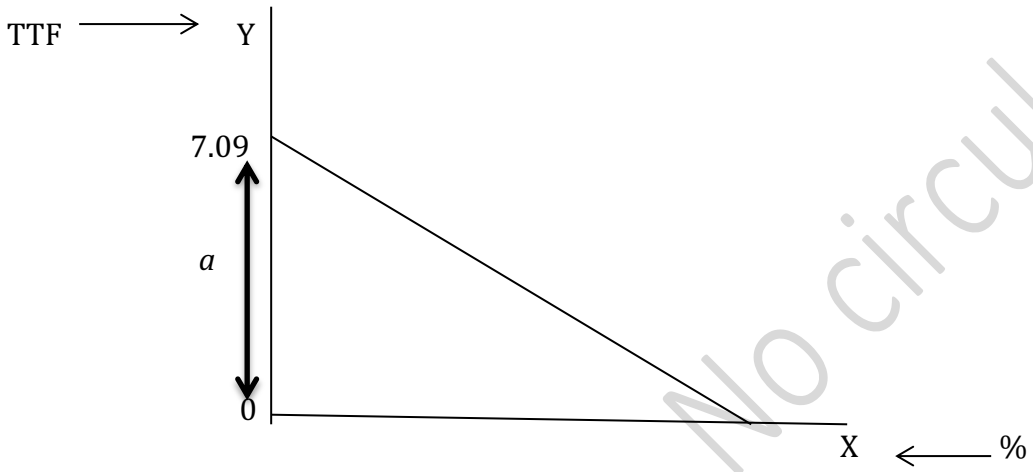
Interpretación de los coeficientes de correlación a y b

¿Cómo debemos interpretar los coeficientes de regresión en el contexto del ejemplo?

¿Cómo debe ser interpretada la *a*? En este caso a = 7.09. Si vemos la ecuación TTF = 7.09 + (–0.07) x porcentaje de uso de anticonceptivos, podemos ver que si el porcentaje de uso de anticonceptivos fuera de 0, entonces -0.07 x 0 sería igual a 0 y entonces la TTF tomaría el valor de 7.09. Por lo tanto, podemos predecir que, si ninguna mujer de un país está usando algún anticonceptivo, la tasa total de fertilidad será de 7.09. Por lo tanto:

a es el valor de Y cuando X = 0

En otras palabras, a es valor de Y cuando la línea de regresión cruza el eje Y (esto es cuando X = 0). Algunas veces se le denomina ***intersección Y***. En este caso la intersección Y es 7.09, así que la línea de regresión se vería más o menos así:



Evidentemente, X = 0 no siempre será una posibilidad. Por ejemplo, si la variable fuera el costo de los libros de texto en una librería, todos serían gratis (si x = 0). Hay que revisar el contexto de los datos que estamos analizando y preguntarnos si lo que afirmamos tiene sentido.

Ahora la interpretación de *b*:

b es una medida del 'gradiente' o empinado de la línea

Supongamos que necesitamos calcular la TTF para países con diferentes porcentajes de mujeres usando anticonceptivos. Esto se puede hacer poniendo diferentes porcentajes en la ecuación y calculando los TTF resultantes. Por ejemplo, si el 10% de las mujeres estuvieran usando anticonceptivos, la ecuación sería:

$$TTF = 7.09 + [- (0.07 \times 10)] = 6.39$$

Por lo tanto, esperaríamos que la TTF fuera de 6.39 en un país con una prevalencia del 10% de uso de anticonceptivos.

El cuadro siguiente muestra más valores de la TTF que estaríamos prediciendo, usando la ecuación.

Cuadro A. Valores de la TTF estimada del uso de anticonceptivos usando la ecuación de regresión

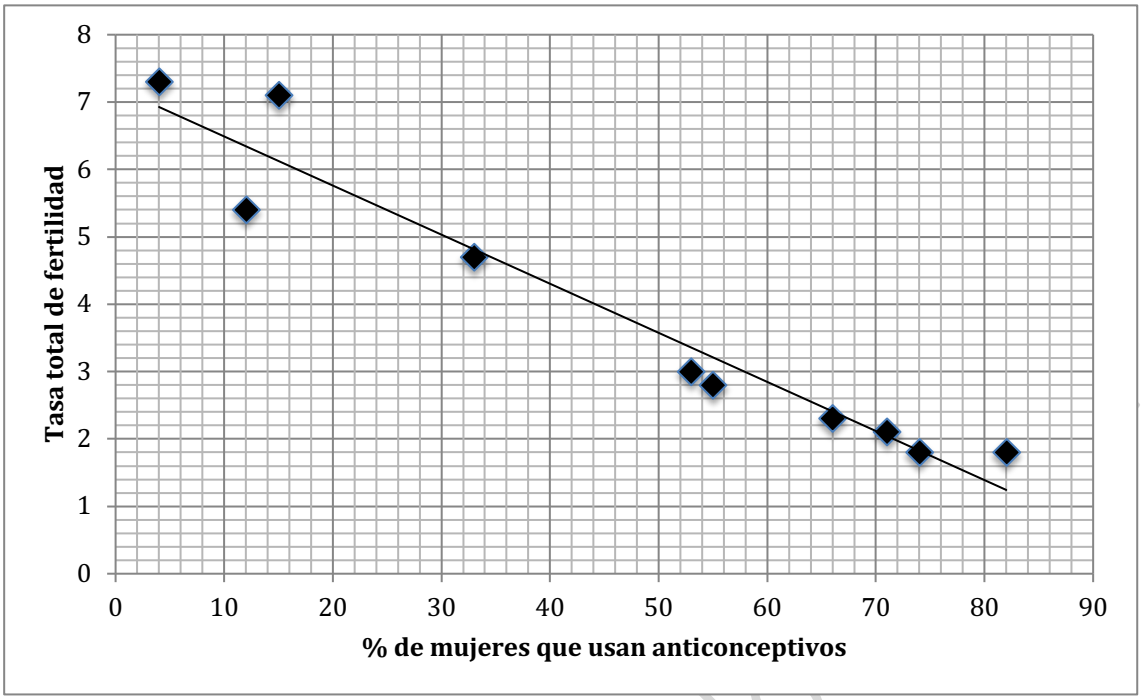
% de mujeres usando anticonceptivos	TTF estimada
0	7.09
1	7.02
2	6.95
25	5.34
40	4.29
80	1.49
100	0.09

¿Qué sucede cada vez que la prevalencia en el uso de anticonceptivos se incrementa en 1%?

La TTF cae un 0.07 en cada ocasión. Esto es verdadero para todos los valores de prevalencia de uso de anticonceptivos, independientemente de si el incremento es de 0% a 1% o de 78% a 79%.

Por lo tanto, podemos decir que la TTF incrementa por b cuando incrementamos la prevalencia de uso de anticonceptivos por 1%. Debido a que b es negativo en este caso, la TTF en realidad cae. De forma similar, si la prevalencia en el uso de anticonceptivos se incrementa por 10%, la TTF caerá 10 x 0.07 = 0.7.

¿Podemos observar qué está mal con la TTF cuando estimamos que el 100% de las mujeres estarían usando algún método anticonceptivo? Un valor de 0.09 para la TTF sería irrealísticamente bajo. Sería muy poco probable que el 100% de las mujeres estuvieran usando un método anticonceptivo al mismo tiempo. Esto nos sucedió porque hicimos un cálculo fuera del rango de nuestros datos originales. En estos la prevalencia más alta era de 82% y la más baja de 4%. Hacer predicciones fuera del rango (extrapolar) puede ser muy aventurado.



Tenemos datos en este rango | No datos

LA LÍNEA DE LA REGRESIÓN

Reglas generales aplicables a cualquier línea de regresión:

1. La ecuación de una línea de regresión es: $Y = a + bX$.
2. b es el cambio en Y por un cambio de una unidad en X.
3. Si b es positivo, Y incrementa en la medida en que X incrementa.
4. Si b es negativo, Y disminuye en la medida en que X disminuye.
5. Si $b = 0$, la línea será horizontal y no habrá relación entre las dos variables.

Ejemplo

En países no desarrollados, el agua contaminada frecuentemente está asociada con enfermedades. El cuadro siguiente muestra datos de la expectativa de vida en 1994 en 10 países en desarrollo, junto con el porcentaje de sus poblaciones con acceso a agua potable entre 1990-1996.

Esperanza de vida al nacer y % de la población con acceso a agua limpia en 10 países en desarrollo

Esperanza de vida al nacer, 1994	% de población con acceso a agua potable, 1990-1996
79.0	100
72.4	71
70.1	85
69.3	68
68.2	80
55.9	57
57.2	29
54.4	28
45.6	48
33.6	34

1. Haga una gráfica de dispersión. ¿Qué le dice acerca de la relación entre acceso a agua potable y esperanza de vida?
2. Calcule el coeficiente de correlación. ¿Qué significa su respuesta?
3. Calcule los coeficientes de regresión a y b.
4. Escriba, en palabras, la ecuación para la línea de regresión.
5. Si nadie en una población tuviera acceso a agua potable, ¿qué predicción haríamos en cuanto a la expectativa de vida? ¿Sería realista?
6. Haga una predicción de la expectativa de vida de un país con un porcentaje de población con acceso a agua potable de: (1) 40%, (2) 52% y (3) 90%.