Introducción a la Estadística y Probabilidad

Tema 3. Análisis bivariante de datos

MANUEL MONGE, Ph.D.

Departamento de Economía Aplicada y Métodos Cuantitativos
Facultad de Derecho, Economía y Gobierno
Universidad Francisco de Vitoria

Contenido

1. Introducción

1. Introducción

- En el tema anterior se han estudiado medidas y gráficos para describir una única variable.
- Sin embargo, cuando se llevan a cabo estudios empresariales y económicos, se dispone de información relativa a distintas variables.
- En este caso no sólo es necesario describir cada una de las variables por separado sino detectar posibles relaciones entre las variables aprovechando la información conjunta disponible.
- Este tema se centra en la descripción de las relaciones entre dos variables.

1. Introducción

Ejemplos:

- ¿Cómo varía la mortalidad infantil en los países en vías de desarrollo cuando aumenta la renta per cápita?
- ¿Cómo aumenta la publicidad las ventas?
- ¿Cuánto varía la cantidad vendida cuando varía el precio?
- Etc.

2. Distribución conjunta de frecuencias

2. Distribución conjunta de frecuencias

- Para resumir la información de dos variables X e Y, vamos a utilizar lo que se denomina en la literatura **tablas de doble entrada** o **tablas de contingencia**.
- La tabla tendrá tantas filas (r) como respuestas posibles tenga X y tantas columnas (c) como respuestas posibles tenga Y.
- En el caso de que alguna de las variables sea continua, los datos se agruparán en intervalos, según se vio en temas anteriores.
- La combinación de respuestas posibles de las dos variables determina las celdas de la tabla y en ellas se enumera el número de observaciones correspondientes a las dos variables conjuntamente:
    - $f_{i,j}$ representará la frecuencia absoluta de las respuestas x_i e y_j de las variables X e Y, respectivamente

2. Distribución conjunta de frecuencias

Representación de la tabla de doble entrada

X/Y	y_1	y_2	...	y_c
x_1	$f_{1,1}$	$f_{1,2}$		$f_{1,c}$
x_2	$f_{2,1}$	$f_{2,2}$		$f_{2,c}$
x_r	$f_{r,1}$	$f_{r,2}$		$f_{r,c}$

2. Distribución conjunta de frecuencias

Ejemplo - X e Y variables cualitativas

El medallero de los Juegos Olímpicos de Barcelona 1992 de los países que obtuvieron más de 30 medallas fue el siguiente:

Países / Medallas	Oro	Plata	Bronce
Estados Unificados	45	35	29
Estados Unidos	37	34	37
Alemania	33	21	28
China	16	22	16
Cuba	14	6	11

2. Distribución conjunta de frecuencias

Ejemplo - X variable cuantitativa e Y variable cualitativa

En la siguiente tabla se indica la edad (en años) de 20 niños junto con su conducta agresiva (evaluada en una escala de 0 a 5):

Edad	C.A.
6	1
6,4	0
5	3
4,3	4
6,5	2
7	3
3	5
7	0
7,4	0
7,8	3
10	2
8,2	1
8,5	3
9,1	2
8,9	2
5,3	4
4	3
6	1
7,2	1
9,3	1

2. Distribución conjunta de frecuencias

La tabla de doble entrada que resume estos datos sería:

Edad/C. Agresiva	0	1	2	3	4	5
[2, 4[						1
[4, 6[				2	2	
[6, 8[	3	3	1	2		
[8, 10]		2	3	1		

2. Distribución conjunta de frecuencias

Ejemplo - X e Y variables cuantitativas

Un estudio nutricional sobre una muestra de 20 jóvenes revela los siguientes datos sobre su estatura y su peso.

Estatura	164	175	165	170	178	157	167	172	177	160
Peso	53	62	48	60	52	63	54	60	55	70

Estatura	168	160	164	174	170	182	161	171	173	193
Peso	63	51	50	80	65	63	60	62	63	86

2. Distribución conjunta de frecuencias

La tabla de doble entrada que resume estos datos sería:

Estatura/Peso	[45, 55[	[55, 65[	[65, 75[	[75, 85[	[85, 95[
[150, 160[		1			
[160, 170[	5	4	1	1	
[170, 180[	1	4	1		
[180, 190[		1			
[190, 200[					1

3. Distribución marginal

3. Distribución marginal

- Como hemos visto anteriormente, cada una de las variables tiene su propia distribución.
- En el caso de dos variables, cada una con su distribución, pueden calcularse a partir de la distribución conjunta sumando por filas y por columnas para obtener las **distribuciones marginales**.
- Es decir, $f_i = \sum_{j=1}^{c} f_{i,j}$ representa la distribución marginal de la variable X y $f_j = \sum_{j=1}^{r} f_{i,j}$ representa la distribución marginal de la variable Y.

3. Distribución marginal

Ejemplo

Calculemos las distribuciones marginales de las variables del ejemplo anterior:

Estatura/Peso	[45, 55[	[55, 65[	[65, 75[	[75, 85[	[85, 95[	f_i
[150, 160[		1				1
[160, 170[	5	4	1	1		11
[170, 180[	1	4	1			6
[180, 190[		1				1
[190, 200[					1	1
f_j	6	10	2	1	1	20

4. Distribución condicionadas

4. Distribución condicionadas

- Una **distribución condicionada** nos permite conocer el tanto por ciento de los valores de una variable condicionada a una respuesta concreta de la otra variable.
- La expresión de la frecuencia condicionada en tantos por ciento es:

$$f_{i|j} = \frac{f_{i,j}}{f_j}$$

4. Distribución condicionadas

Ejemplo

Consideremos la clasificación de 20 alumnos según su color de pelo y su sexo:

Sexo/Pelo	Moreno	Castaño	Rubio	f_i
Chica	5	2	4	11
Chico	4	5		9
f_j	9	7	4	20

4. Distribución condicionadas

- La pregunta, ¿qué tanto por ciento de los alumnos de pelo moreno son chicas? corresponde a una pregunta sobre una distribución condicionada.
- La condición en este caso es ser moreno, de tal forma que nos restringimos a la primera columna de la tabla en la que participan 9 alumnos y calculamos:

$$f_{chica|morena} = \frac{5}{9} = 0{,}56$$

- Es decir, del total de alumnos morenos, el 56 % son chicas.

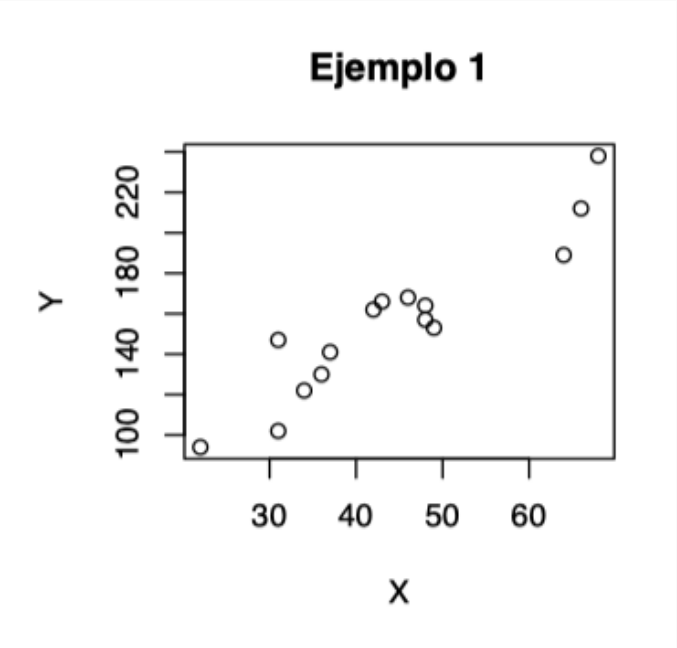
5. Diagrama de dispersión

5. Diagrama de dispersión

- La distribución conjunta de dos variables puede expresarse gráficamente mediante un **diagrama de dispersión**.
- Este diagrama se construye representando cada elemento observado por un punto en el plano de manera que sus coordenadas sobre los dos ejes cartesianos sean los valores que toman las dos variables en ese elemento.
- Este gráfico proporciona una buena descripción de la relación entre las dos variables.
- También es capaz de caracterizar varios aspectos de las variables que lo componen, como el rango en que varía cada una de las variables, la posible asociación de los datos y una indicación de los casos atípicos.
- En este tipo de gráficos puede ocurrir que dos o más individuos tengan valores idénticos. En este caso, se podría mover mínimamente alguno de los dos datos para que aparezcan ligeramente separados los distintos puntos que realmente existen, ya que en este caso se solaparían y el ordenador los representaría como si fuesen solo uno.

5. Diagrama de dispersión

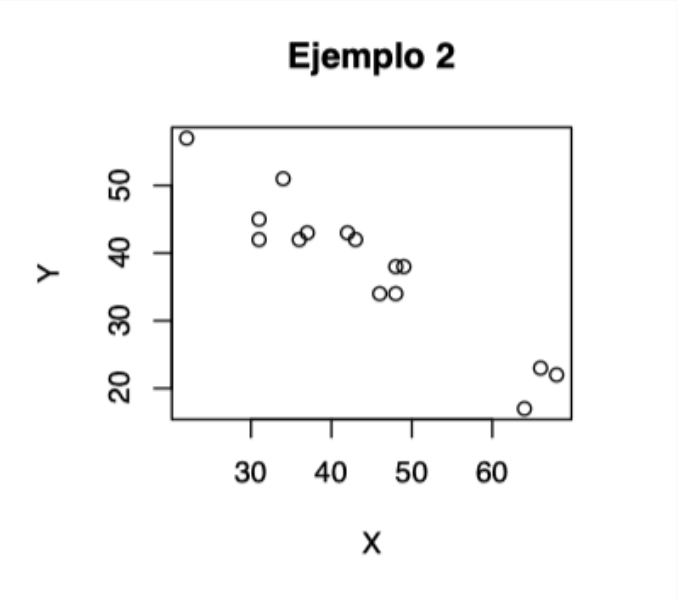
Asociación lineal positiva



Como vemos, cuando aumenta el valor de la variable X, también aumenta el valor de la variable Y. Además, los puntos tienden a colocarse siguiendo una línea, por lo que se dice que ambas variables tienen una relación lineal positiva.

5. Diagrama de dispersión

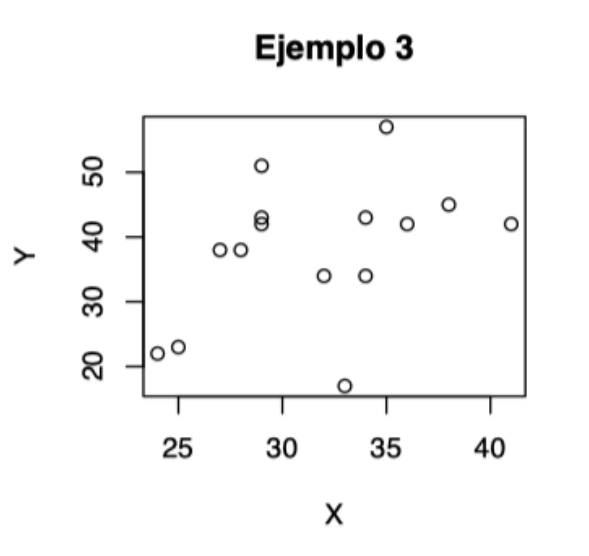
Asociación lineal negativa



En este ejemplo los puntos también se disponen de forma lineal, pero, esta vez, cuando X aumenta, Y disminuye, por lo que se dice que las variables tienen una relación lineal negativa.

5. Diagrama de dispersión

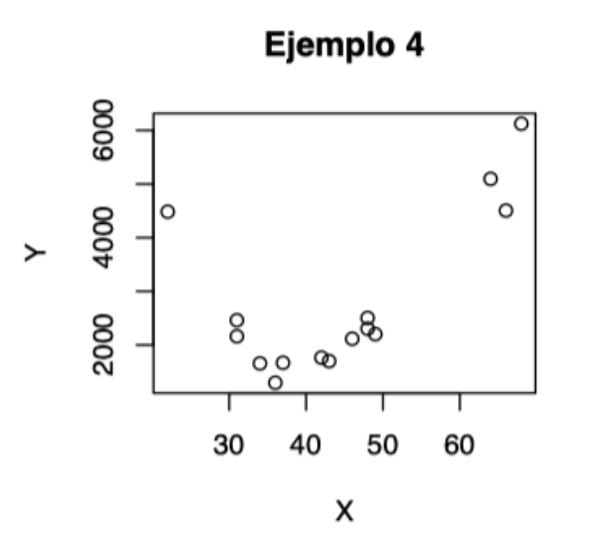
Ausencia de asociación



- La nube de puntos del tercer ejemplo no muestra ninguna tendencia entre las dos variables.
- Decimos, entonces, que las variables son independientes y que el conocimiento de una de ellas no proporciona información sobre el valor de la otra.

5. Diagrama de dispersión

Asociación no lineal



En este último ejemplo se observa que las variables están relacionadas pero esta relación no es lineal.

6. Covarianza

6. Covarianza

- La **covarianza** es una medida de asociación lineal que resume la información existente en un gráfico de dispersión.
- Mide el sentido de la relación lineal.
- Un valor positivo de la covarianza indica una relación lineal directa o creciente, es decir cuando los valores de una variable crecen los valores de la otra variable también crecen y viceversa.
- Un valor negativo indica una relación lineal inversa o decreciente, es decir si una variable crece la otra decrece y viceversa.
- Un valor nulo (covarianza nula) indica que las variables son estadísticamente independientes linealmente.

6. Covarianza

La covarianza se formula mediante la siguiente expresión:

$$cov(X, Y) = S_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n}$$

En la fórmula anterior, x_i e y_i son los valores observados de las variables X e Y, $\overline{x}$ e $\overline{y}$ son las medias y n es el tamaño de la muestra.

7. Coeficiente de correlación

7. Coeficiente de correlación

- El **coeficiente de correlación (r) o de Pearson** mide la dirección y la magnitud de la asociación entre dos variables cuantitativas.
- Nos da una medida estandarizada de la relación lineal entre dos variables.
- Indica tanto el sentido como el grado de relación.
- Es el más utilizado[1].
- Se trata de un índice que mide lo bien que se ajustan los puntos a una línea recta ideal (se dice que existen relación lineal si $|r| \geq \frac{2}{\sqrt{n}}$).
- r solo detecta asociaciones lineales.
- Es un método estadístico paramétrico, ya que utiliza la media, la varianza, etc., de modo que requiere criterios de normalidad para las variables analizadas.
- No varía ante los cambios de origen y/o escala.
- El coeficiente de correlación es adimensional y varía entre -1 y $+1$.

[1] Existe el coeficiente de correlación no paramétrico de Spearman (rho), que se utiliza en aquellos casos en los que las variables examinadas no cumplen necesariamente criterios de normalidad o bien cuando las variables son ordinales.

7. Coeficiente de correlación

El coeficiente de correlación se formula mediante la siguiente expresión:

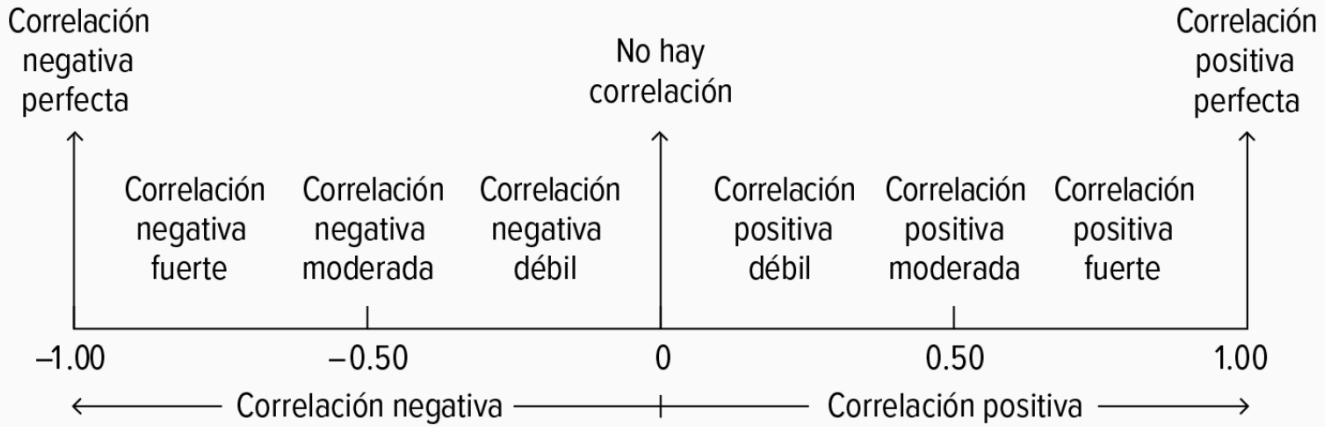
$$r = \frac{S_{XY}}{S_X S_Y}$$

En la expresión anterior, S_{XY} es la covarianza entre las dos variables y $S_X S_Y$ son las desviaciones típicas muestrales de X e Y, respectivamente.

7. Coeficiente de correlación

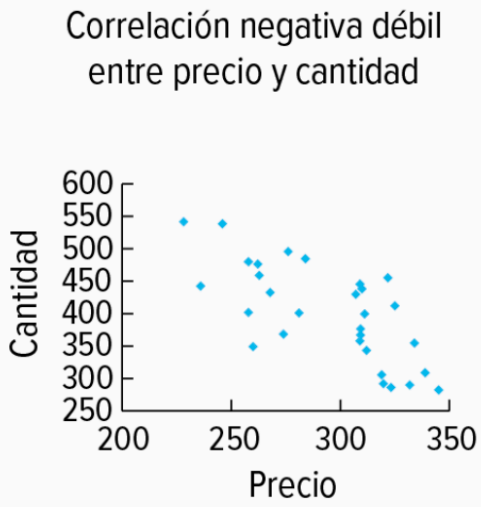
Interpretación del resultado:

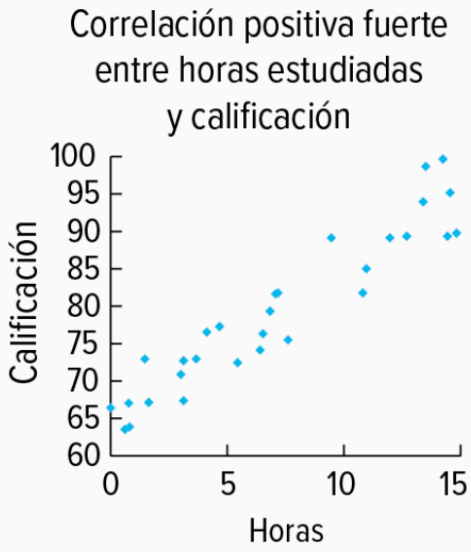
- Cuanto más cerca se encuentra r de $+1$, más cerca se encuentran los datos de alinearse sobre una recta ascendente que indica una **relación lineal positiva**.
- Cuanto más cerca se encuentra r de -1 más cerca se encuentran los datos de alinearse sobre una recta descendente que indica una **relación lineal negativa**.
- Cuando $r = 0$, no existe ninguna relación lineal entre X e Y, pero eso no quiere decir necesariamente que no exista ningún tipo de relación entre las variables.



7. Coeficiente de correlación

Ejemplos gráficos





7. Ejemplo

Ejemplo

Foxconn, empresa que manufactura productos tecnológicos, desea estudiar la relación entre el número de trabajadores (X) y el número de productos producidos (Y) en su planta de Taiwan. La empresa ha tomado una muestra aleatoria de 10 horas de producción. En la tabla se muestran las observaciones recogidas. Analiza brevemente la relación entre el número de trabajadores y el número de mesas producidas por hora:

x_i	y_i	$(x_i - \overline{x})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})$	$(y_i - \overline{y})^2$	$(x_i - \overline{x})(y_i - \overline{y})$
12	20	-9.3	86.49	-21.2	449.44	197.16
30	60	8.7	75.69	18.8	353.44	163.56
15	27	-6.3	39.69	-14.2	201.64	89.46
24	50	2.7	7.29	8.8	77.44	23.76
14	21	-7.3	53.29	-20.2	408.04	147.46
18	30	-3.3	10.89	-11.2	125.44	36.96
28	61	6.7	44.89	19.8	392.04	132.66
26	54	4.7	22.09	12.8	163.84	60.16
19	32	-2.3	5.29	-9.2	84.64	21.16
27	57	5.7	32.49	15.8	249.64	90.06
213	412		378.1		2505.6	962.4

7. Ejemplo

$$S_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n} = \frac{962{,}4}{10} = 96{,}24 \text{ (asociación lineal positiva)}$$

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{106{,}93}{\sqrt{42{,}01}\sqrt{278{,}4}} = 0{,}989$$

$$|0{,}989| \geq \frac{2}{\sqrt{10}} \cong 0{,}64$$

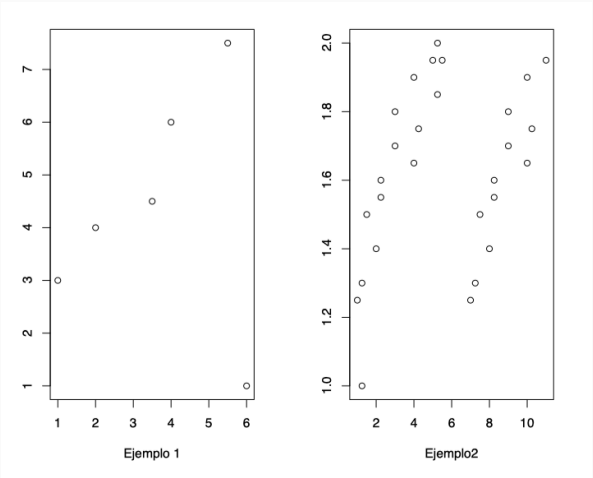
Llegamos a la conclusión de que existe una estrecha relación positiva entre el número de trabajadores y el número de mesas producidas por hora.

7. Reflexiones

Correlación y heterogeneidad

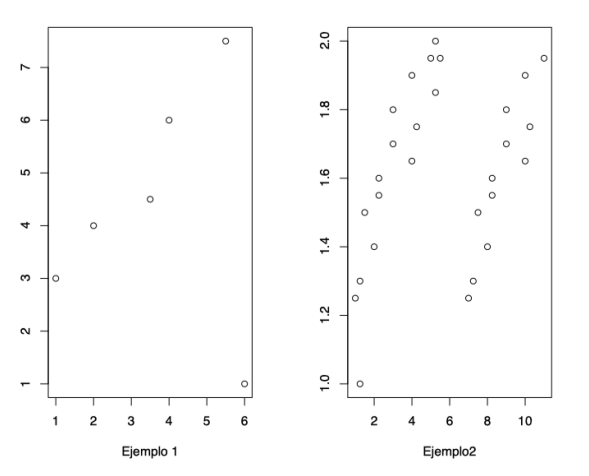
- Cuando se estudia la relación entre dos variables es importante asegurarse de que los elementos estudiados son homogéneos respecto a dichas variables.
- Por ejemplo, la figura que pongo más abajo presenta dos casos frecuentes de heterogeneidad.
- En el ejemplo 1 hay un **dato atípico** o discordante con el resto, que modifica el signo de correlación. Puede comprobarse que si el punto con coordenadas (6,1) del ejemplo 1 no existiese, el coeficiente de correlación sería positivo, mientras que su presencia hace la correlación negativa.

- Ante una situación de este estilo conviene asegurarse de que:
    1. No se ha cometido un error de medida o de transcripción del dato.
    2. El elemento de la población al que le corresponde el dato atípico es homogéneo con respecto a los demás.
- Si existe un error de medida conviene eliminar el dato y si el elemento es distinto de los demás por alguna razón objetiva conviene también suprimirlo del cálculo del coeficiente de correlación.



7. Reflexiones

- En el ejemplo 2 de la figura que pongo abajo se muestra otro caso de heterogeneidad.
- En este caso el gráfico indica que la relación entre las variables es distinta para los elementos de las dos zonas de puntos y si calculamos un coeficiente de correlación para todos los datos obtendremos un valor muy pequeño.
- Sin embargo, si obtenemos los coeficientes para los dos grupos de puntos separadamente, encontraremos que dentro de cada grupo hay una relación fuerte.



7. Reflexiones

- La conclusión fundamental de este análisis es que un coeficiente de correlación es el resumen de la relación presente en el gráfico de dispersión.
- Conviene, pues, asegurarse mirando este gráfico que el coeficiente es un buen resumen del mismo.
- Tratar de interpretar un coeficiente de correlación sin haber visto previamente el gráfico de las variables puede ser peligroso.

7. Reflexiones

Correlación y causalidad

- Hemos visto que un coeficiente de correlación alto entre dos variables indica que los elementos observados toman valores relacionados entre sí, pero no permite concluir la existencia de ninguna relación de causalidad de una variable respecto de otra.
- Por ejemplo, si relacionamos el número de matrimonios mensuales en una ciudad y la temperatura media mensual registrada, el coeficiente de correlación entre ambas es muy alto.
- Sin embargo, es obvio que no existe una relación causal entre ambas variables.
- Ni es probable que un aumento de los matrimonios eleve la temperatura media del mes, ni es esperable que una ola de calor cause una avalancha de matrimonios.
- La razón del alto coeficiente de correlación es que los matrimonios tienden a producirse en verano, ya que de esta manera las parejas pueden aprovechar sus vacaciones para tener más tiempo en el inicio de su vida en común.
- Este tipo de correlaciones se denomina **correlaciones espurias** y se deben al efecto de otra variable (las vacaciones veraniegas) que al tener una relación de dependencia con las variables que observamos (matrimonio y temperatura) crea la relación entre ellas.

7. Reflexiones

- Por otra parte, si no encontramos correlación entre dos variables tampoco podemos deducir que no exista relación lineal entre ellas.
- En primer lugar, si las observaciones tienen un rango de variación pequeño y existen otros factores que producen variabilidad es posible que no veamos causalidad aunque exista.
- Por ejemplo, consideremos la relación entre el tamaño de un apartamento y su precio de alquiler en un conjunto de pisos de una gran ciudad. Supongamos que tomamos una muestra de pisos entre 80 y $100m^2$. Como el tamaño de estos pisos es bastante homogéneo, su precio dependerá de su localización y de otros factores. En consecuencia, el coeficiente de correlación entre precio y tamaño será pequeño. Sin embargo no podemos concluir que el tamaño del piso no influye en su precio. Si tomásemos una muestra que incluyese pisos pequeños y grandes, la relación entre estas dos variables aparecería claramente.

8. Análisis de regresión

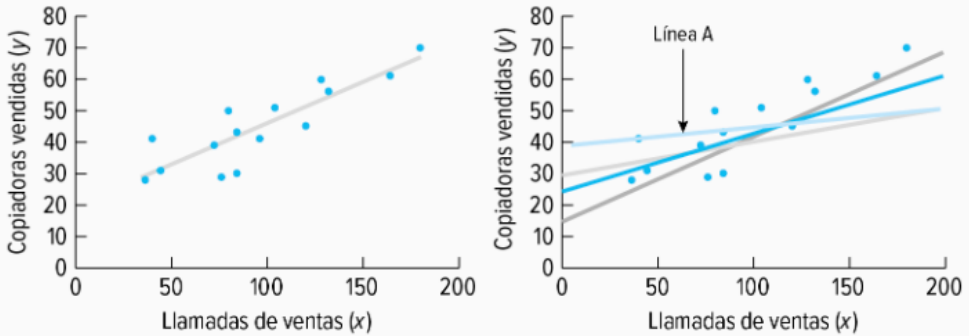
8. Análisis de regresión

- El **análisis de regresión** es otro método para examinar una relación lineal entre dos variables, es decir, una variable en función de otra variable.
- La variable que se estima es la **variable dependiente (o endógena)**. La llamaremos 'Y'.
- La variable utilizada para hacer la estimación o predecir el valor es la **variable independiente (o exógena)**. La llamaremos 'X'.
- La relación entre las variables es lineal.
- La variable dependiente ('Y') ocupa siempre el eje de ordenadas (eje vertical)
- La variable independiente ('X') ocupa el eje de abscisas (eje horizontal).
- Tanto la variable independiente como la dependiente deben ser escala de intervalo o razón.
- Proporciona mucha más información al expresar la relación lineal entre dos variables en forma de ecuación.

8. Análisis de regresión

Principio de los mínimos cuadrados

- Por definición, el **principio de los mínimos cuadrados** es un procedimiento matemático que emplea datos para ubicar una línea con la finalidad de minimizar la suma de los cuadrados de las distancias verticales entre los valores reales y Y y los que se pronostican.
- En otras palabras, nuestro objetivo es utilizar los datos para posicionar una línea que represente mejor la relación entre dos variables (comúnmente se conoce como recta del "mejor ajuste".
- El primer enfoque es usar un diagrama de dispersión para posicionar visualmente la línea.



8. Análisis de regresión

Forma general de la ecuación de regresión lineal

$$\hat{y} = a + bx$$

donde,

- $\hat{y}$ es el valor de la estimación de la variable y para el valor x seleccionado.
- a es la ordenada al origen. Es el valor estimado de Y cuando $x = 0$. En otras palabras, a es el valor estimado de y donde la recta de regresión cruza el eje Y cuando x es cero;
- b es la pendiente de la recta, o el cambio promedio en $\hat{y}$ por cada cambio de una unidad (ya sea aumento o reducción) de la variable independiente x.
- x es cualquier valor de la variable independiente que se seleccione.

El propósito de un análisis de regresión es calcular los valores de a y b para desarrollar una ecuación lineal que se ajuste mejor a los datos.

Hay que tener en cuenta otro concepto, que es **error de predicción** o **residuo de la recta**, que es el error al predecir cada uno de los valores observados de la variable dependiente y.

$$\textit{residuo} = \textit{error de predicción} = \textit{valor observado} - \textit{valor de la recta}$$

8. Análisis de regresión

La recta se calcula imponiendo la condición de que el **error promedio**, definido como la raíz cuadrada de la suma de los cuadrados de los errores al prever cada punto con la recta, sea mínimo. Este criterio se denomina **mínimos cuadrados**.

$$Min(\sqrt{\frac{\sum_{i=1}^{n}[y_i-(bx_i+a)]^2}{n}})$$

Como resultado de aplicar este criterio, se obtiene:

- la pendiente de la recta de regresión
- la ordenada al origen

Las fórmulas de a y b son:

Pendiente de la recta de regresión

$$b = r(\frac{S_y}{S_x})$$

donde,

- r es el coeficiente de correlación.
- S_y es la desviación estándar de y (variable dependiente);
- S_x es la desviación estándar de x (variable independiente).

Ordenada al origen

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

donde,

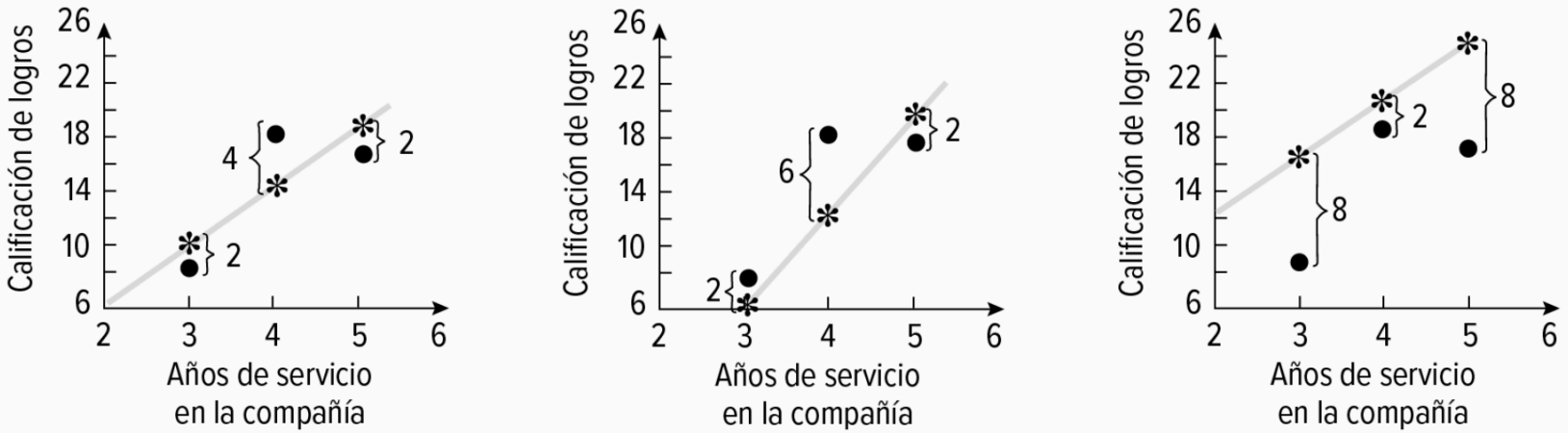
- $\overline{y}$ es la media de y (variable dependiente).
- $\overline{x}$ es la media de x (variable independiente).

Esta ecuación indica que la recta debe pasar por el punto $(\overline{x}, \overline{y})$, es decir, por el centro de los datos.

8. Análisis de regresión

Ejemplo

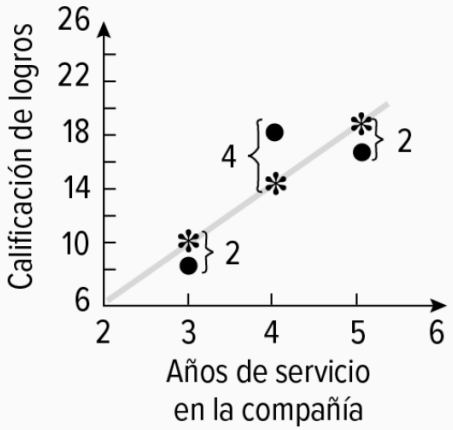
Para ilustrar este concepto, se trazan los mismos valores en las tres siguientes gráficas:



- Los puntos representan los valores reales de y.
- Los asteriscos, los valores predichos de y para un valor dado de x.
- La gráfica de la izquierda, es la recta de mínimos cuadrados.
- La gráfica del medio y de la derecha son diferentes rectas trazadas con una regla.

8. Análisis de regresión

Ejemplo



- Esta gráfica representa la recta de mínimos cuadrados, la cual es la recta de mejor ajuste porque la suma de los cuadrados de las desviaciones verticales respecto de si misma es mínima.
- En este gráfica, el primer punto que encontramos es $x = 3, y = 8$ que se desvía 2 unidades de la recta, calculada como $10 - 8$.
- El cuadrado de la desviación es 4.
- En este mismo gráfico, la desviación al cuadrado que se obtiene de la gráfica en $x = 4, y = 18$ es 16.
- La que se obtiene en $x = 5, y = 16$ es 4.
- La suma de las desviaciones al cuadrado es 24 $(4 + 16 + 4)$.
- Como podríamos apreciar en las otras dos gráficas, cuyas rectas se han trazado con regla, la suma de las desviaciones verticales al cuadrado son 44 y 132, respectivamente. Éstas son mayores que la recta que ha sido trazada por el método de mínimos cuadrados.