Sección 5

Variables bidimensionales

En lo estudiado anteriormente hemos podido aprender cómo a partir de la gran cantidad de datos que describen una muestra mediante una variable, X, se representan gráficamente los mismos de modo que resulta más intuitivo hacerse una idea de como se distribuyen las observaciones.

Otros conceptos que según hemos visto, también nos ayudan en el análisis, son los estadísticos de tendencia central, que nos indican hacia donde tienden a agruparse los datos (en el caso en que lo hagan), y los estadísticos de dispersión, que nos indican si las diferentes modalidades que presenta la variable están muy agrupadas alrededor de cierto valor central, o si por el contrario las variaciones que presentan las modalidades con respecto al valor central son grandes.

También sabemos determinar ya si los datos se distribuyen de forma simétrica a un lado y a otro de un valor central.

En este capítulo pretendemos estudiar una situación muy usual y por tanto de gran interés en la práctica:

> Si Y es otra variable definida sobre la misma población que X, ¿será posible determinar si existe alguna relación entre las modalidades de X y de Y?

Un ejemplo trivial consiste en considerar una población formada por alumnos de primero de Medicina y definir sobre ella las variables

$$\begin{aligned} X &\equiv \text{altura medida en centímetros,} \\ Y &\equiv \text{altura medida en metros,} \end{aligned}$$

ya que la relación es determinista y clara: $Y = X/100$. Obsérvese que aunque la variable Y, como tal puede tener cierta dispersión, vista *como función* de X, su dispersión es nula.

Un ejemplo más parecido a lo que nos interesa realmente lo tenemos cuando sobre la misma población definimos las variables

$$\begin{aligned} X &\equiv \text{altura medida en centímetros,} \\ Y &\equiv \text{peso medida en kilogramos.} \end{aligned}$$

Intuitivamente esperamos que exista cierta relación entre ambas variables, por ejemplo,

$$Y = X - 110 \pm \textbf{dispersión}$$

Tablas de doble entrada

Consideramos una población de n individuos, donde cada uno de ellos presenta dos caracteres que representamos mediante las variables X e Y. Representamos mediante

$$X \rightsquigarrow x_1, x_2, \ldots, x_i, \ldots, x_k$$

las k modalidades que presenta la variable X, y mediante

$$Y \rightsquigarrow y_1, y_2, \ldots, y_j, \ldots, y_p$$

las p modalidades de Y.

Con la intención de reunir en una sóla estructura toda la información disponible, creamos una tabla formada por $k \cdot p$ casillas, organizadas de forma que se tengan k filas y p columnas. La casilla denotada de forma general mediante el subíndice$_{ij}$ hará referencia a los elementos de la muestra que presentan simultáneamente las modalidades x_i e y_j.

Y / X	y_1	y_2	$\dots$	y_j	$\dots$	y_p	
x_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1p}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2p}	$n_{2\bullet}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{ip}	$n_{i\bullet}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_k	n_{k1}	n_{k2}	$\dots$	n_{kj}	$\dots$	n_{kp}	$n_{k\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet j}$	$\dots$	$n_{\bullet p}$	$n_{\bullet\bullet}$

De este modo, para $i = 1, \dots, k$, $j = 1, \dots, p$, se tiene que n_{ij} es el número de individuos o **frecuencia absoluta**, que presentan a la vez las modalidades x_i e y_j.

El número de individuos que presentan la modalidad x_i, es lo que llamamos **frecuencia absoluta marginal** de x_i y se representa como $n_{i\bullet}$. Es evidente la igualdad

$$n_{i\bullet} = n_{i1} + n_{i2} + \cdots + n_{ip} = \sum_{j=1}^{p} n_{ij}$$

Obsérvese que hemos escrito un símbolo "$\bullet$" en la *"parte de las jotas"* que simboliza que estamos considerando los elemento que presentan la modalidad x_i, independientemente de las modalidades que presente la variable Y. De forma análoga se define la frecuencia absoluta marginal de la modalidad y_j como

$$n_{\bullet j} = n_{1j} + n_{2j} + \cdots + n_{kj} = \sum_{i=1}^{k} n_{ij}$$

Estas dos distribuciones de frecuencias $n_{i\bullet}$ para $i = 1, \ldots, k$, y $n_{\bullet j}$ para $j = 1, \ldots, p$ reciben el nombre de **distribuciones marginales** de X e Y respectivamente.

El número total de elementos de la población (o de la muestra), n lo obtenemos de cualquiera de las siguientes formas, que son equivalentes:

$$n = n_{\bullet\bullet} = \sum_{i=1}^{k} n_{i\bullet} = \sum_{j=1}^{p} n_{\bullet j} = \sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij}$$

Distribuciones condicionadas

De todos los elementos de la población, n, podemos estar interesados, en un momento dado, en un conjunto más pequeño y que está formado por aquellos elementos que han presentado la modalidad y_j, para algún $j = 1, \ldots, p$. El número de elementos de este conjunto sabemos que es $n_{\bullet j}$. La variable X definida sobre este conjunto se denomina **variable condicionada** y se suele denotar mediante $X_{|y_j}$ o bien $X_{|Y=y_j}$. La distribución de frecuencias absolutas de esta nueva variable es exactamente la columna j de la tabla.

De la misma forma, es posible dividir la población inicial en k subconjuntos, cada uno de ellos caracterizados por la propiedad de que el i–ésimo conjunto todos los elementos verifican la propiedad de presentar la modalidad x_i. Sobre cada uno de estos conjuntos tenemos la variable condicionada $Y_{|x_i} \equiv Y_{|X=x_i}$, cuya distribución de frecuencias relativas condicionadas es:

$$f_j^i = \frac{n_{ij}}{n_{i\bullet}} \qquad \forall j = 1, \ldots, p$$

Dependencia funcional e independencia

La relación entre las variables X e Y, parte del objetivo de este capítulo y en general de un número importante de los estudios de las Ciencias Sociales, puede ser más o menos acentuada, pudiendo llegar ésta desde la dependencia total o *dependencia funcional* hasta la *independencia*.

Dependencia funcional

La dependencia funcional, que nos refleja cualquier fórmula matemática o física, es a la que estamos normalmente más habituados. Al principio del capítulo consideramos un ejemplo en el que sobre una población de alumnos definíamos las variables

$$\begin{aligned} X &\equiv \text{altura medida en centímetros,} \\ Y &\equiv \text{altura medida en metros,} \end{aligned}$$

Al tomar a uno de los alumnos, hasta que no se realice una medida sobre el mismo, no tendremos claro cual será su altura. Podemos tener cierta intuición sobre qué valor es más probable que tome (alrededor de la media, con cierta dispersión). Sin embargo, si la medida X ha sido realizada, no es necesario practicar la de Y, pues la relación entre ambas es exacta (dependencia funcional):

$$Y = X/100$$

Independencia

Existe un concepto que es radicalmente opuesto a la dependencia funcional, que es el de *independencia*. Se dice que dos variables X e Y son **independientes** si la distibución marginal de una de ellas es la misma que la condicionada por cualquier valor de la otra.

Esta es una de entre muchas maneras de expresar el concepto de independencia, y va a implicar una estructura muy particular de la tabla bidimensional, en el que todas las filas y todas las columnas van a ser proporcionales entre sí.

Covarianza

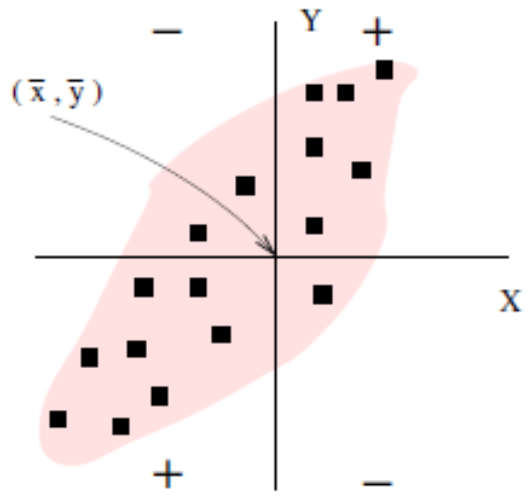
La **covarianza** $\mathcal{S}_{XY}$, es una medida que nos hablará de la variabilidad conjunta de dos variables numéricas (cuantitativas). Se define como:

$$\mathcal{S}_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

Una interpretación geométrica de la covarianza

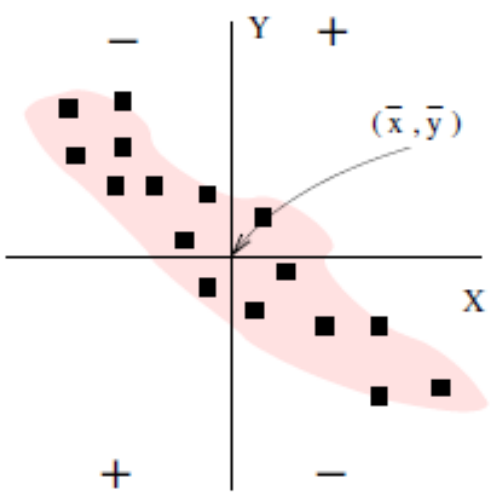
Consideremos la *nube de puntos* formadas por las n parejas de datos (x_i, y_i). El centro de gravedad de esta nube de puntos es $(\overline{\overline{x}}, \overline{\overline{y}})$, o bien podemos escribir simplemente $(\overline{x}, \overline{y})$ si los datos no están ordenados en una tabla de doble entrada. Trasladamos los ejes XY al nuevo centro de coordenadas $(\overline{x}, \overline{y})$. Queda así dividida la nube de puntos en cuatro cuadrantes como se observa en la figura 3.1. Los puntos que se encuentran en el primer y tercer cuadrante contribuyen positivamente al valor de $\mathcal{S}_{XY}$, y los que se encuentran en el segundo y el cuarto lo hacen negativamente.

- Si hay mayoría de puntos en el tercer y primer cuadrante, ocurrirá que $\mathcal{S}_{XY} \geq 0$, lo que se puede interpretar como que la variable Y tiende a aumentar cuando lo hace X;



Cuando X crece, Y crece

Casi todos los puntos pertenecen a los cuadrantes primero y tercero

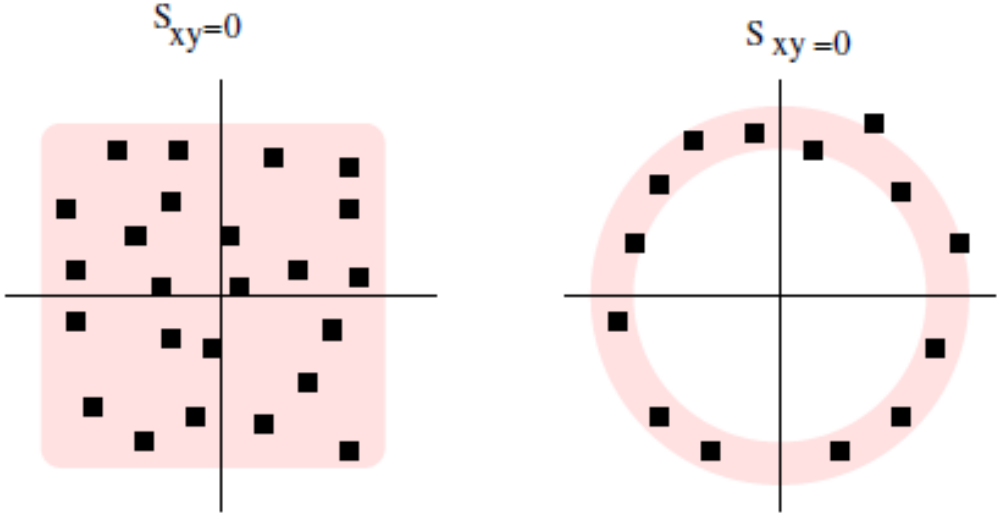


Cuando X crece, Y decrece

Casi todos los puntos pertenecen a los cuadrantes segundo y cuarto

- Si la mayoría de puntos están repartidos entre el segundo y cuarto cuadrante entonces $\mathcal{S}_{XY} \leq 0$, es decir, las observaciones Y tienen tendencia a disminuir cuando las de X aumentan;

- Si los puntos se reparten con igual intensidad alrededor de $(\overline{\overline{x}}, \overline{\overline{y}})$, entonces se tendrá que $\mathcal{S}_{XY} = 0$. Véase la figura como ilustración.



Las dos variables son independientes.

Hay dependencia entre las dos variables, aunque la covarianza sea nula.

Cuando los puntos se reparte de modo más o menos homogéneo entre los cuadrantes primero y tercero, y segundo y cuarto, se tiene que $\mathcal{S}_{XY} \approx 0$. Eso no quiere decir de ningún modo que no pueda existir ninguna relación entre las dos variables, ya que ésta puede existir como se aprecia en la figura de la derecha.

La Covarianza

- Si $\mathcal{S}_{XY} > 0$ las dos variables crecen o decrecen a la vez (nube de puntos creciente).

- Si $\mathcal{S}_{XY} < 0$ cuando una variable crece, la otra tiene tendencia a decrecer (nube de puntos decreciente).

- Si los puntos se reparten con igual intensidad alrededor de $(\overline{x}, \overline{y})$, $\mathcal{S}_{XY} = 0$ (no hay relación lineal).

Coeficiente de correlación lineal de Pearson

La covarianza es una medida de la variabilidad común de dos variables (crecimiento de ambas al tiempo o crecimiento de una y decremimiento de la otra), pero está afectada por las unidades en las que cada variable se mide. Así pues, es necesario definir una medida de la relación entre dos variables, y que no esté afectada por los cambios de unidad de medida. Una forma de conseguir este objetivo es dividir la covarianza por el producto de las desviaciones típicas de cada variable, ya que así se obtiene un coeficiente adimensional, r, que se denomina **coeficiente de correlación lineal de Pearson**

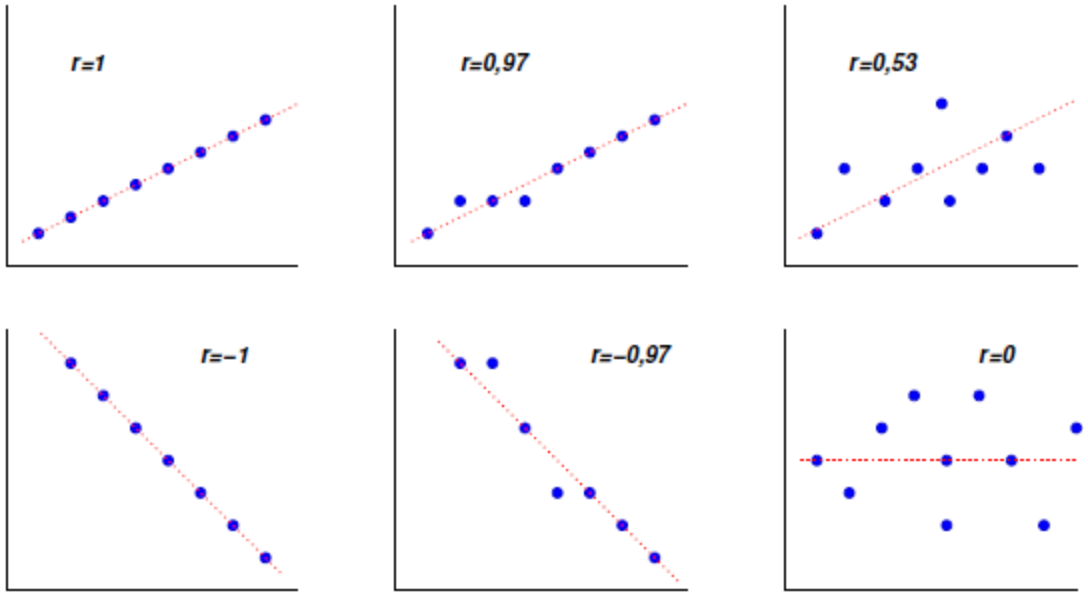
$$r = \frac{S_{XY}}{S_X S_Y}$$

Propiedades del coeficiente de correlación lineal

- Carece de unidades de medida (adimensional).
- Es invariante para transformaciones lineales (cambio de origen y escala) de las variables.
- Sólo toma valores comprendidos entre -1 y 1,
- Cuando $|r|$ esté próximo a uno, se tiene que existe una *relación lineal* muy fuerte entre las variables.
- Cuando $r \approx 0$, puede afirmarse que no existe relación lineal entre ambas variables. Se dice en este caso que las variables son **incorreladas.**

Regresión

Las técnicas de regresión permiten hacer predicciones sobre los valores de cierta variable Y *(dependiente)*, a partir de los de otra X *(independiente)*, entre las que intuimos que existe una relación.



$r = \pm 1$ es lo mismo que decir que las observaciones de ambas variables están perfectamente alineadas. El signo de r, es el mismo que el de $\mathcal{S}_{XY}$, por tanto nos indica el crecimiento o decrecimiento de la recta. La relación lineal es tanto más perfecta cuanto r está cercano a ± 1.

$$\begin{aligned} X &\equiv \text{altura medida en centímetros,} \\ Y &\equiv \text{altura medida en metros,} \end{aligned}$$

no es necesario hacer grandes esfuerzos para *intuir* que la relación que hay entre ambas es:

$$Y = \frac{X}{100}.$$

Obtener esta relación es menos evidente cuando lo que medimos sobre el mismo grupo de personas es

$$\begin{aligned} X &\equiv \text{altura medida en centímetros,} \\ Y &\equiv \text{peso en kilogramos.} \end{aligned}$$

La razón es que no es cierto que conocida la altura x_i de un individuo, podamos determinar de modo exacto su peso y_i (v.g. dos personas que miden $1,70m$ pueden tener pesos de 60 y 65 kilos). Sin embargo, alguna relación entre ellas debe existir, pues parece mucho más probable que un individuo de $2m$ pese más que otro que mida $1,20m$. Es más, nos puede parecer más o menos aproximada una relación entre ambas variables como la siguiente

$$Y = X - 110 \pm \textbf{ error.}$$

A la deducción, a partir de una serie de datos, de este tipo de relaciones entre variables, es lo que denominamos **regresión**.

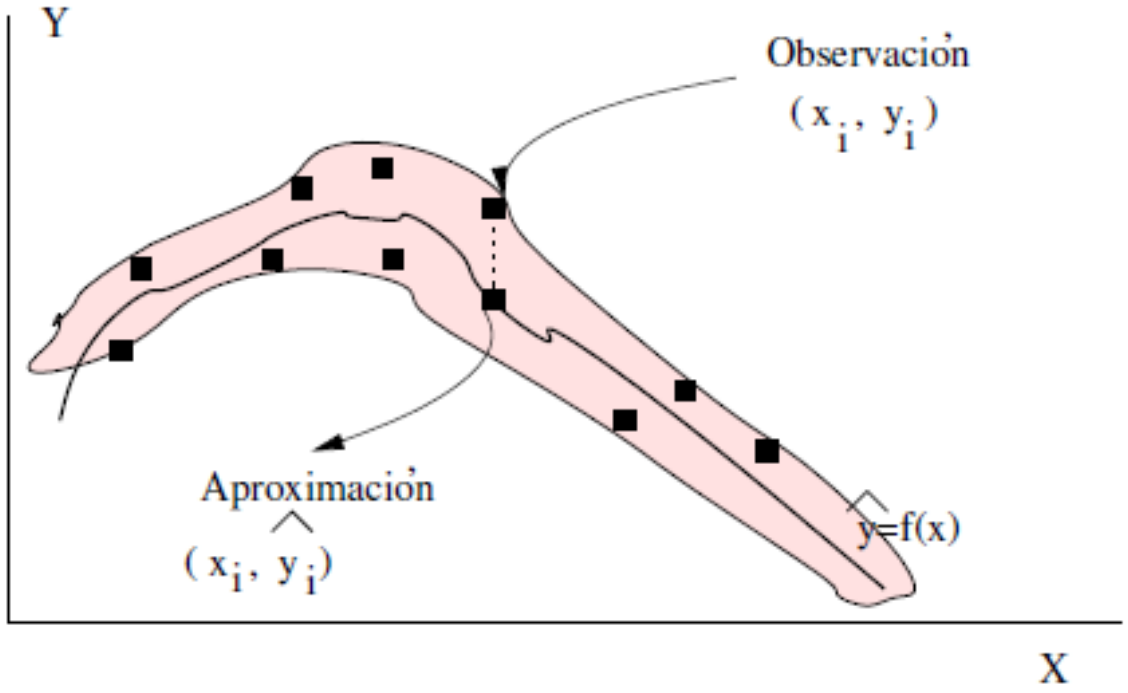
Mediante las técnicas de regresión inventamos una variable $\hat{Y}$ como función de otra variable X (o viceversa),

$$\hat{Y} = f(X).$$

Esto es lo que denominamos **relación funcional**. El criterio para construir $\hat{Y}$, tal como citamos anteriormente, es que la diferencia entre Y e $\hat{Y}$ sea pequeña.

$$\hat{Y} = f(X), \qquad Y - \hat{Y} = \textbf{error},$$

El término que hemos denominado **error** debe ser tan pequeño como sea posible . El objetivo será buscar la función (también denominada **modelo de regresión**) $\hat{Y} = f(X)$ que lo minimice.



Mediante las técnicas de regresión de una variable Y sobre una variable X, buscamos una función que sea una buena aproximación de una nube de puntos (x_i, y_i), mediante una curva del tipo $\hat{Y} = f(X)$. Para ello hemos de asegurarnos de que la diferencia entre los valores y_i e $\hat{y}_i$ sea tan pequeña como sea posible.

Bondad de un ajuste

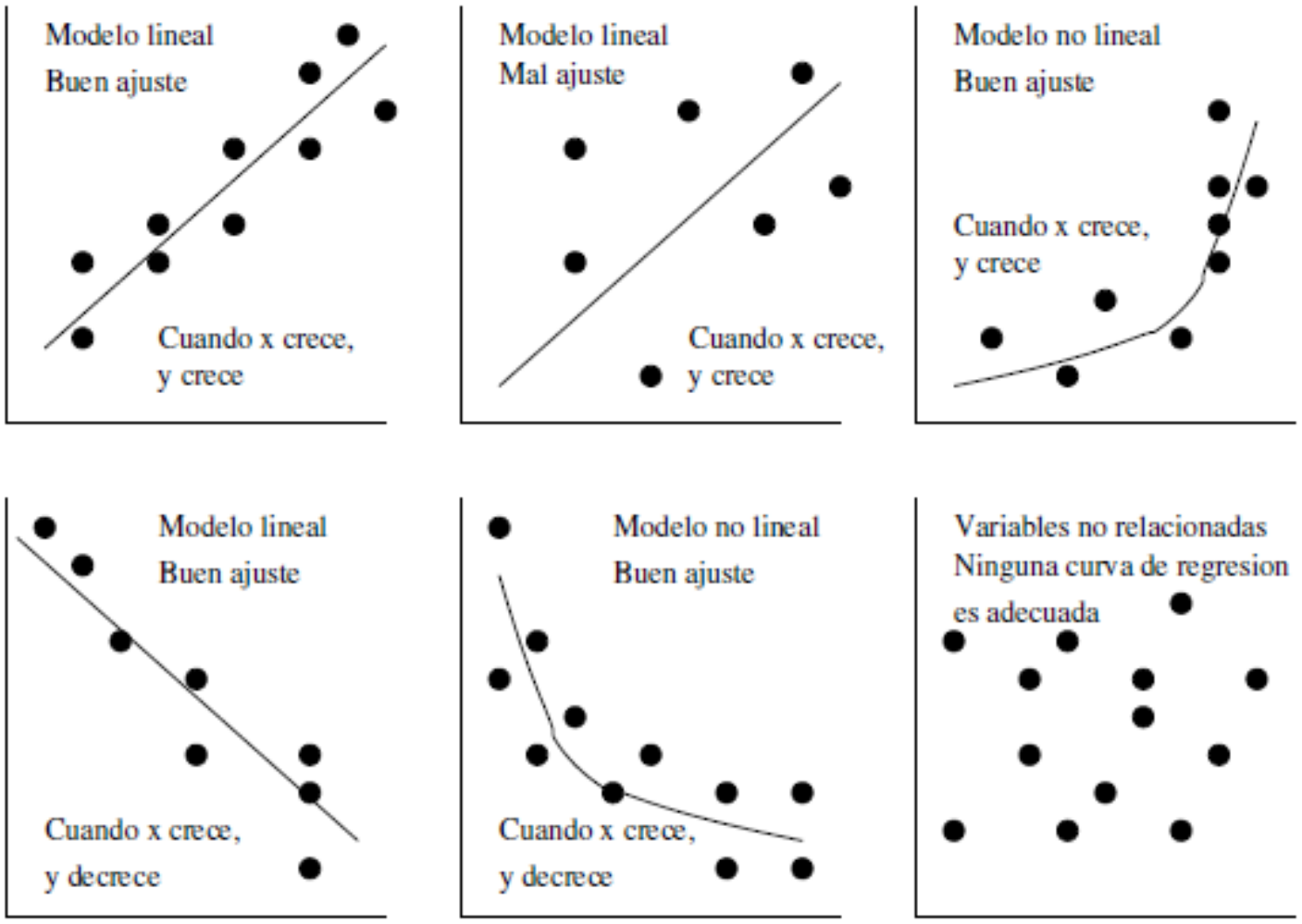
Consideremos un conjunto de observaciones sobre n individuos de una población, en los que se miden ciertas variables X e Y:

$$\begin{aligned} X &\rightsquigarrow x_1, x_2, \ldots, x_n \\ Y &\rightsquigarrow y_1, y_2, \ldots, y_n \end{aligned}$$

Estamos interesamos en hacer regresión para determinar, de modo aproximado, los valores de Y conocidos los de X, debemos definir cierta variable $\hat{Y} = f(X)$, que debe tomar los valores

$$\hat{Y} \rightsquigarrow \hat{y}_1 = f(x_1), \hat{y}_2 = f(x_2), \ldots, \hat{y}_n = f(x_n)$$

de modo que:



Diferentes nubes de puntos y modelos de regresión para ellas.

$$Y - \hat{Y} \rightsquigarrow y_1 - \hat{y}_1 \approx 0, y_2 - \hat{y}_2 \approx 0, \ldots, y_n - \hat{y}_n \approx 0$$

Ello se puede expresar definiendo una nueva variable E que mida las diferencias entre los auténticos valores de Y y los teóricos suministrados por la regresión,

$$E = Y - \hat{Y} \rightsquigarrow e_1 = y_1 - \hat{y}_1, e_2 = y_2 - \hat{y}_2, \ldots, e_n = y_n - \hat{y}_n$$

y calculando $\hat{Y}$ de modo que E tome valores cercanos a 0. Dicho de otro modo, E debe ser una variable cuya media debe ser 0 , y cuya varianza $\mathcal{S}_E^2$ debe ser pequeña (en comparación con la de Y). Por ello se define el

coeficiente de determinación de la regresión de Y sobre X, $R^2_{Y_{|X}}$, como

$$R^2_{Y_{|X}} = 1 - \frac{\mathcal{S}^2_E}{\mathcal{S}^2_Y}$$

Si el ajuste de Y mediante la curva de regresión $\hat{Y} = f(X)$ es bueno, cabe esperar que la cantidad $R^2_{Y_{|X}}$ tome un valor próximo a 1.

La cantidad $R^2_{Y_{|X}}$ sirve entonces para medir de qué modo las diferencias entre los verdaderos valores de una variable y los de su aproximación mediante una curva de regresión son pequeños en relación con los de la variabilidad de la variable que intentamos aproximar. Por esta razón estas cantidades miden el **grado de bondad del ajuste**.

Regresión lineal

La **regresión lineal** consiste en encontrar aproximar los valores de una variable a partir de los de otra, usando una relación funcional de tipo lineal, es decir, buscamos cantidades a y b tales que se pueda escribir

$$\hat{Y} = a + b \cdot X$$

con el menor error posible entre $\hat{Y}$ e Y.

Las cantidades a y b que minimizan dicho error son los llamados *coeficientes de regresión*:

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

$$\boxed{b = \frac{\mathcal{S}_{XY}}{\mathcal{S}_X^2}}$$

La cantidad b se denomina *coeficiente de regresión de Y sobre X*.

En el modelo lineal de regresión la *bondad del ajuste* es simplemente r^2. Con lo cual el modelo lineal dará mejores predicciones cuando r sea próximo a 1 ó -1.

Interpretación de los coeficientes de regresión

Obsérvese que la relación $\hat{Y} = a + b \cdot X$ explica cosas como que si X varia en 1 unidad, Y varía la cantidad b. Por tanto:

- Si $b > 0$, las dos variables aumentan o disminuyen a la vez;
- Si $b < 0$, cuando una variable aumenta, la otra disminuye.

Ejemplo de cálculo con un modelo de regresión lineal

En una muestra de 1.500 individuos se recogen datos sobre dos medidas antropométricas X e Y. Los resultados se muestran resumidos en los siguientes estadísticos:

$$\overline{x} = 14 \quad \mathcal{S}_X = 2$$

$$\mathcal{S}_{XY} = 45$$

$$\overline{y} = 100 \quad \mathcal{S}_Y = 25$$

Obtener el modelo de regresión lineal que mejor aproxima Y en función de X. Utilizando este modelo, calcular de modo aproximado la cantidad Y esperada cuando $X = 15$.

Solución:

Lo que se busca es la recta, $\hat{Y} = a + b \cdot X$, que mejor aproxima los valores de Y (según el criterio de los mínimos cuadrados) en la nube de puntos que resulta de representar en un plano (X, Y) las 1.500 observaciones. Los coeficientes de esta recta son:

$$b = \frac{\mathcal{S}_{XY}}{\mathcal{S}_X^2} = \frac{45}{4} = 11,25$$

$$a = \overline{y} - b \cdot \overline{x} = 100 - 11,25 \times 14 = -57,5$$

Así, el modelo lineal consiste en:

$$\hat{Y} = -57,5 + 11,25 \cdot X$$

Por tanto, si $x = 15$, el modelo lineal predice un valor de Y de:

$$\hat{y} = -57,5 + 11,25 \cdot x = -57,5 + 11,25 \times 15 = 111,25$$

Propiedades de la regresión lineal

Una vez que ya tenemos perfectamente definida $\hat{Y}$, (o bien $\hat{X}$) nos preguntamos las relaciones que hay entre la media y la varianza de esta y la de Y (o la de X). La respuesta nos la ofrece la siguiente proposición:

Proposición

En los ajustes lineales se conservan las medias, es decir

$$\begin{aligned} \overline{\hat{y}} &= \overline{y} \\ \overline{\hat{x}} &= \overline{x} \end{aligned}$$

En cuanto a la varianza, no necesariamente son las mismas para los verdaderos valores de las variables X e Y y sus aproximaciones $\hat{X}$ y $\hat{Y}$, pues sólo se mantienen en un factor de r^2, es decir,

$$\begin{aligned} \mathcal{S}^2_{\hat{Y}} &= r^2\, \mathcal{S}^2_Y \\ \mathcal{S}^2_{\hat{X}} &= r^2\, \mathcal{S}^2_X \end{aligned}$$

Observación

Como consecuencia de este resultado, podemos decir que *la proporción de varianza explicada por la regresión lineal es del* $r^2 \cdot 100\,\%$.

Nos gustaría tener que $r = 1$, pues en ese caso ambas variables tendrían la misma varianza, pero esto no es cierto en general. Todo lo que se puede afirmar, como sabemos, es que

$$-1 \leq r \leq 1$$

y por tanto

$$0 \leq \mathcal{S}^2_{\hat{Y}} \leq \mathcal{S}^2_Y$$

La cantidad que le falta a la **varianza de regresión**, $\mathcal{S}^2{}_{\hat{Y}}$, para llegar hasta la varianza total de Y, $\mathcal{S}^2_Y$, es lo que se denomina **varianza residual**,

Proposición

La varianza residual del modelo de regresión es de Y sobre X es la varianza de la variable $E = Y - \hat{Y}$.

Obsérvese que entonces La *bondad del ajuste* es

$$R^2_{Y|X} = 1 - \frac{\mathcal{S}^2_E}{\mathcal{S}^2_Y} = 1 - (1 - r^2) = r^2$$

Para el ajuste contrario se define el error como $E = X - \hat{X}$, y análogamente su varianza residual es también proporcional a $1 - r^2$. Todo esto se puede resumir como sigue:

Proposición

Para los ajustes de tipo lineal se tiene que los dos coeficientes de determinación son iguales a r^2, y por tanto representan además la proporción de varianza explicada por la regresión lineal:

$$\boxed{R^2_{X_{|Y}} = r^2 = R^2_{Y_{|X}}}$$

Por ello:

- Si $\mid r \mid \approx 1$ el ajuste es bueno (Y se puede calcular de modo bastante aproximado a partir de X y viceversa).
- Si $\mid r \mid \approx 0$ las variables X e Y no están relacionadas (linealmente al menos), por tanto no tiene sentido hacer un ajuste lineal. Sin embargo no es seguro que las dos variables no posean ninguna relación en el caso $r = 0$, ya que si bien el ajuste lineal puede no ser procentente, tal vez otro tipo de ajuste sí lo sea.

Ejemplo

De una muestra de ocho observaciones conjuntas de valores de dos variables X e Y, se obtiene la siguiente información:

$$\sum x_i = 24; \quad \sum x_i y_i = 64; \quad \sum y_i = 40;$$

$$\mathcal{S}_Y^2 = 12; \quad \mathcal{S}_X^2 = 6.$$

Calcule:

1. La recta de regresión de Y sobre X. Explique el significado de los parámetros.

2. El coeficiente de determinación. Comente el resultado e indique el tanto por ciento de la variación de Y que no está explicada por el modelo lineal de regresión.

3. Si el modelo es adecuado, ¿cuál es la predicción $\hat{y}$ para $x = 4$.

Solución:

1. En primer lugar calculamos las medias y las covarianza entre ambas variables:

$$\begin{aligned} \overline{x} &= \sum x_i/n = 24/8 = 3 \\ \overline{x} &= \sum y_i/n = 40/8 = 5 \\ \mathcal{S}_{XY} &= (\sum x_i y_i)/n - \overline{xy} = 64/8 - 3 \times 5 = -7 \end{aligned}$$

Con estas cantidades podemos determinar los parámetros a y b de la recta. La pendiente de la misma es b, y mide la variación de Y cuando X aumenta en una unidad:

$$b = \frac{\mathcal{S}_{XY}}{\mathcal{S}_X^2} = \frac{-7}{6} = -1,667$$

Al ser esta cantidad negativa, tenemos que la pendiente de la recta es negativa, es decir, a medida que X aumenta, la tendencia es a la disminución de Y. En cuanto al valor de la ordenada en el origen, a, tenemos:

$$a = \overline{y} - b \cdot \overline{x} = 5 - (\frac{-7}{6}) \times 3 = 8,5$$

Así, la recta de regresión de Y como función de X es:

$$\hat{Y} = 8,5 - 1,1667 \cdot X$$

2. El grado de bondad del ajuste lo obtenemos a partir del coeficiente de determinación:

$$R^2_{Y/X} = r^2 = \left(\frac{\mathcal{S}_{XY}}{\mathcal{S}_X \cdot \mathcal{S}_Y}\right)^2 = \frac{(-7)^2}{6 \times 12} = 0,6805 = 68,05\,\%$$

Es decir, el modelo de regresión lineal explica el 68 % de la variabilidad de Y en función de la de X. Por tanto queda un 32 % de variabilidad no explicada.

3. La predicción que realiza el modelo lineal de regresión para $x = 4$ es:

$$\hat{y} = 8,5 - 1,1667 \cdot x = 8,5 - 1,6667 \times 4 = 3,833$$

la cual hay que considerar con ciertas reservas, pues como hemos visto en el apartado anterior,hay una razonable cantidad de variabilidad que no es explicada por el modelo.

Ejemplo de cálculo en regresión lineal

En un grupo de 8 pacientes se miden las cantidades antropométricas *peso* y *edad*, obteniéndose los siguientes resultados:

	Resultado de las mediciones							
$X \equiv$ edad	12	8	10	11	7	7	10	14
$Y \equiv$ peso	58	42	51	54	40	39	49	56

¿Existe una relación lineal importante entre ambas variables? Calcular la recta de regresión de la edad en función del peso y la del peso en función de la edad. Calcular la bondad del ajuste ¿En qué medida, por término medio, varía el peso cada año? ¿En cuánto aumenta la edad por cada kilo de peso?

Solución:

Para saber si existe una relación lineal entre ambas variables se calcula el coeficiente de correlación lineal, que vale:

$$r = \frac{\mathcal{S}_{XY}}{\mathcal{S}_X\,\mathcal{S}_Y} = \frac{15,2031}{2,3150 \times 6,9631} = 0,9431$$

$$\sum_{i=1}^{8} x_i = 79 \implies \overline{x} = \frac{79}{8} = 9,875 \text{ años}$$

$$\sum_{i=1}^{8} y_i = 389 \implies \overline{y} = \frac{389}{8} = 48,625 \text{ Kg}$$

$$\sum_{i=1}^{8} x_i^2 = 823 \implies \mathcal{S}_X^2 = \frac{823}{8} - 9,875^2 = 5,3594 \text{ años}^2$$

$$\implies \mathcal{S}_X = 2,3150 \text{ años}$$

$$\sum_{i=1}^{8} y_i^2 = 19{,}303 \implies \mathcal{S}_Y^2 = \frac{19{,}303}{8} - 48,625^2 = 48,4844 \text{ Kg}^2$$

$$\implies \mathcal{S}_Y = 6,9631 \text{ Kg}$$

$$\sum_{i=1}^{8} x_i y_i = 3{,}963 \implies \mathcal{S}_{XY} = \frac{3{,}963}{8} - 9,875 \times 48,625 = 15,2031 \text{ Kg} \cdot \text{año}$$

Por tanto el ajuste lineal es muy bueno. Se puede decir que el ángulo entre el vector formado por las desviaciones del peso con respecto a su valor medio y el de la edad con respecto a su valor medio, θ, es:

$$r = \cos\theta \qquad \Longrightarrow \qquad \theta = \arccos r \approx 19^\circ$$

es decir, entre esos vectores hay un buen grado de paralelismo (sólo unos 19 grados de desviación).

La recta de regresión del peso en función de la edad es

$$\begin{aligned}
\hat{Y} &= a_1 + b_1 X = 20,6126 + 2,8367 \cdot X \\
a_1 &= \overline{y} - b_1 \overline{x} = 20,6126 \text{ Kg} \\
b_1 &= \frac{\mathcal{S}_{XY}}{\mathcal{S}_X^2} = 2,8367 \text{ Kg/año}
\end{aligned}$$

La recta de regresión de la edad como función del peso es

$$\begin{aligned}
\hat{X} &= a_2 + b_2 Y = -5,3738 + 0,3136 \cdot Y \\
a_2 &= \overline{x} - b_2 \overline{y} = -5,3738 \text{ años} \\
b_2 &= \frac{\mathcal{S}_{XY}}{\mathcal{S}_Y^2} = 0,3136 \text{ años/Kg}
\end{aligned}$$

que como se puede comprobar, no resulta de despejar en la recta de regresión de Y sobre X.

La bondad del ajuste es

$$R^2_{X|Y} = R^2_{Y|X} = r^2 = 0,8894$$

por tanto podemos decir que el $88,94\,\%$ de la variabilidad del peso en función de la edad es explicada mediante la recta de regresión correspondiente. Lo mismo podemos decir en cuanto a la variabilidad de la edad en función del peso. Del mismo modo puede decirse que hay un $100 - 88,94\,\% = 11,06\,\%$ de varianza que no es explicada por las rectas

de regresión. Por tanto la varianza residual de la regresión del peso en función de la edad es

$$\mathcal{S}_E^2 = (1 - r^2) \cdot \mathcal{S}_Y^2 = 0,1106 \times 48,4844 = 5,33 \text{ Kg}^2$$

y la de la edad en función del peso:

$$\mathcal{S}_E^2 = (1 - r^2) \cdot \mathcal{S}_X^2 = 0,1106 \times 5,3594 = 0,59 \text{ años}^2$$

Por último la cantidad en que varía el peso de un paciente cada año es, según la recta de regresión del peso en función de la edad, la pendiente de esta recta, es decir, $b_1 = 2,8367$ Kg/año. Cuando dos personas difieren en peso, en promedio la diferencia de edad entre ambas se rige por la cantidad $b_2 = 0,3136$ años/Kg de diferencia.

Ejercicio 3.1. Se realiza un estudio para establecer una ecuación mediante la cual se pueda utilizar la *concentración de estrona en saliva*(X) para predecir la *concentración del esteroide en plasma libre* (Y). Se extrajeron los siguientes datos de 14 varones sanos:

X	1,4	7,5	8,5	9	9	11	13	14	14,5	16	17	18	20	23
Y	30	25	31,5	27,5	39,5	38	43	49	55	48,5	51	64,5	63	68

1. Estúdiese la posible relación lineal entre ambas variables.
2. Obtener la ecuación que se menciona en el enunciado del problema.
3. Determinar la variación de la concentración de estrona en plasma por unidad de estrona en saliva.