CORRELACIÓN Y REGRESIÓN

CR I

Correlación Lineal

a. Considérese el problema de tratar de hallar la relación funcional existente entre dos variables aleatorias X e Y.
La investigación de dicha interrelación, basada en *n* experimentos en que dichas variables asumieron pares de valores $(x_1 , y_1) \ldots (x_n , y_n)$, generalmente se encara graficando dichos pares de valores sobre un sistema de coordenadas ortogonales. Dicho gráfico, llamado diagrama de dispersión a menudo permite discernir si existe alguna tendencia hacia algún tipo de interrelación entre ambas variables, y, si posible, la naturaleza de dicho tipo de interrelación.
A título ilustrativo considérese la tabla de la figura CR I.a correspondientes a las notas obtenidas por 30 alumnos en Matemáticas y Física, siendo 50 la nota máxima posible. Arbitrariamente, se asignará la variable X a la nota obtenida en Matemáticas e Y a la nota obtenida en Física, resultando así el diagrama de dispersión indicado en la figura CR I.b.

x_i	y_i	x_i	y_i	x_i	y_i
34	37	28	30	39	36
37	37	30	34	33	29
36	34	32	30	30	29
32	34	41	37	33	40
32	33	38	40	43	42
36	40	36	42	31	29
35	39	37	40	38	40
34	37	33	36	34	31
29	36	32	31	36	38
35	35	33	31	34	32

Fig. CR I.a

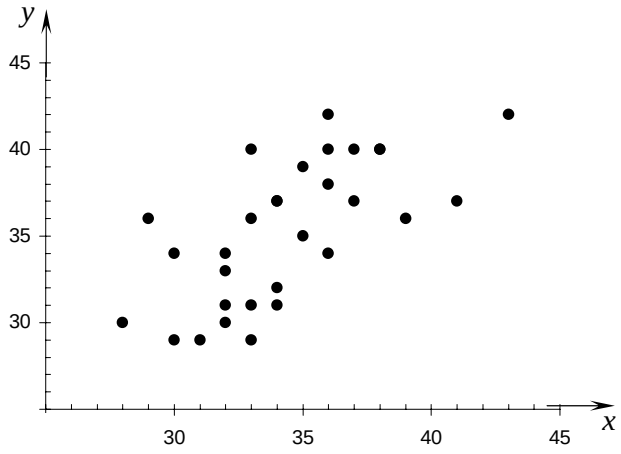

Fig. CR I.b

A primera vista, del diagrama de dispersión resulta que existe una tendencia a que valores altos de X estén asociados con valores altos de Y, y a que valores bajos de X estén asociados con valores bajos de Y, resultando así que la tendencia de la dispersión es hacia una línea recta, y sería deseable medir de alguna manera el grado en que las antedichas variables X e Y están linealmente interrelacionadas.

b. Con el fin de efectuar dicha medición se empezará por ver que condiciones sería deseable que tuviera.
Para empezar dicha medición debería ser independiente de la elección del origen de las variables. El hecho de que el diagrama de dispersión de la figura CR I.b tenga por origen el punto (25 , 25) implica que se admitió que la interrelación entre X e Y es independiente del origen.
Esta propiedad se obtiene usando valores $x_i - \bar{x}$ e $y_i - \bar{y}$ en vez de los valores primitivos x_i e y_i.
En segundo lugar, la medición de la interrelación debe ser independiente de las unidades en que vienen medidos los valores X e Y.

Así, si los valores x_i e y_i indicados en la figura CR I.a fueran ambos duplicados, no debería quedar afectada la medida de la interrelación entre X e Y. Esta propiedad es obtenida dividiendo los valores x_i por una constante que tenga su misma dimensión, y haciendo lo propio con los valores y_i. Por razones que serán evidentes mas adelante se tomarán a s_x y a s_y como dichas constantes.
Es decir que las dos propiedades recién indicada serán obtenidas usando valores:

$$u_i = \frac{x_i - \bar{x}}{s_x} \qquad \text{y} \qquad v_i = \frac{y_i - \bar{y}}{s_y}$$

en vez de los x_i e y_i.
El diagrama de dispersión de los puntos (u_i , v_i) correspondiente a los datos de la figura CR I.a es el indicado en la figura CR I.c.

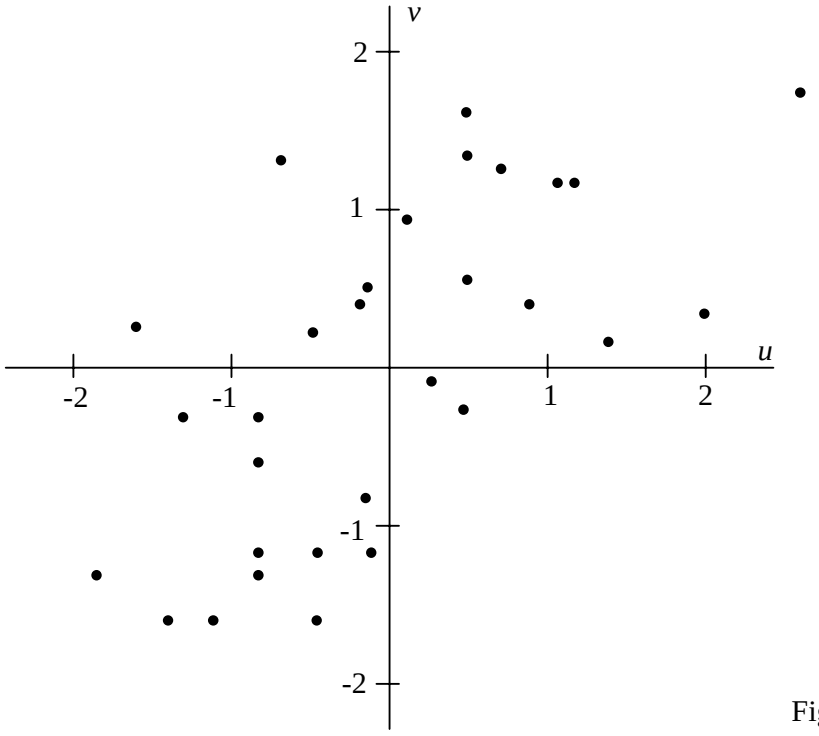

Fig. CR I.c

Se puede observar en este diagrama que la mayoría de los puntos están ubicados en los cuadrantes primero y tercero y que dichos puntos tienden a tener coordenadas cuyo valor absoluto es mayor que la de los puntos ubicados en los cuadrantes segundo y cuarto.

Una sencilla medida de esta tendencia es $\sum_{i=1}^{n} u_i v_i$. Los puntos de los cuadrantes primero y tercero contribuirán valores positivos a esta sumatoria, mientras que los puntos de los cuadrantes segundo y cuarto contribuirán valores negativos. Por lo tanto un valor positivo grande de dicha sumatoria parece indicar una tendencia lineal positiva en el diagrama de dispersión.
A la recíproca, un valor negativo grande indicaría una tendencia lineal negativa en el diagrama de dispersión.

Todo lo antedicho no es estrictamente cierto ya que si la cantidad de puntos fuera duplicada sin cambiar la naturaleza de la dispersión, el valor de la sumatoria se vería aproximadamente duplicado. Por lo tanto es necesario dividir el valor de la sumatoria por la cantidad de puntos, *n*, antes de que se la pueda usar como medida de la interrelación lineal entre las variables.
Por lo tanto, la deseada medida de la interrelación lineal será:

$$r = \text{Coeficiente de correlación lineal} = \frac{\sum_{i=1}^{n} u_i v_i}{n} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n s_x s_y} \quad \textbf{[1]}$$

c. Efectuando cálculos puede hallarse que para los datos indicados en la figura CR I.a corresponde un coeficiente de correlación lineal $r = 0{,}66$.
Para interpretar este resultado y para ver que valores de *r* se obtienen para distintos tipos de diagramas de dispersión, considérese los diagramas de dispersión de la figura CR I.d.
Los primeros cuatro diagramas corresponden a valores crecientes de *r*, es decir a interrelaciones lineales crecientes entre X e Y.
Si estos diagramas fueran rotados 180º alrededor del eje *y*, las dispersiones aparecerían como teniendo una dispersión lineal negativa, y los correspondientes valores de *r* serían los negativos de los indicados en dichos diagramas.
Por lo tanto el valor absoluto de *r* determina la magnitud de la interrelación lineal mientras que su signo indica si los valores y_i tienden a aumentar o disminuir a medida que los x_i crecen.

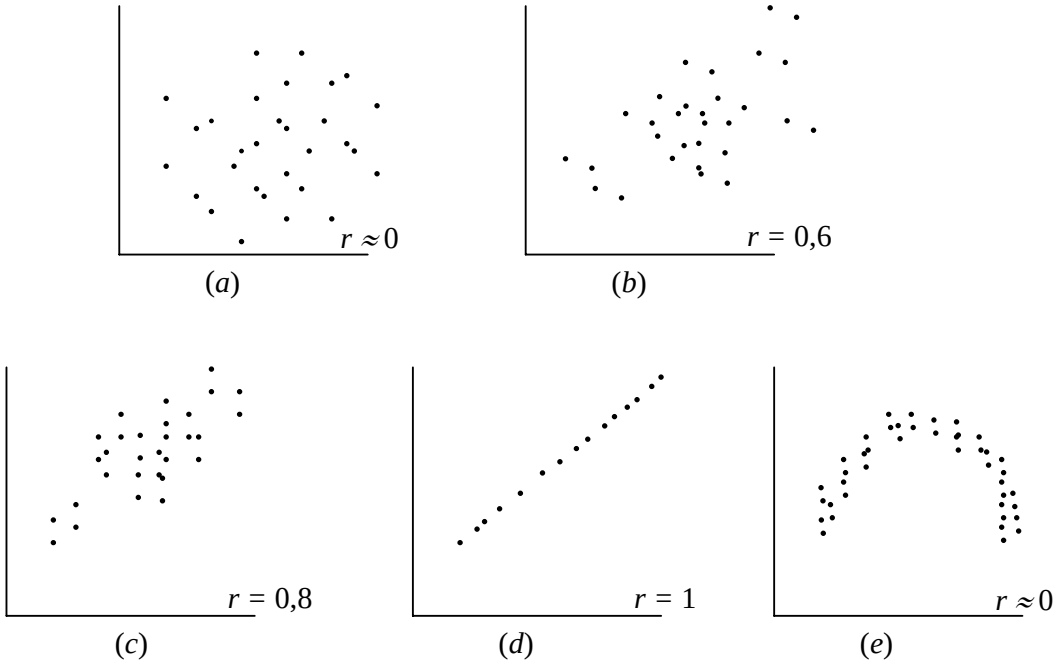

Fig. CR I.d

El quinto diagrama ilustra un caso en el cual los x_i e y_i están muy interrelacionados pero que dicha interrelación no es lineal.
Esto ilustra que el coeficiente de correlación r es una medida útil de la interrelación entre variables solo cuando dicha interrelación tiende a ser lineal.
Los diagramas de la figura CR I.d y sus valores asociados de r hacen plausible dos propiedades de dicho coeficiente:

1º. $-1 \leq r \leq 1$

2º. $r = \pm 1$ cuando y solo cuando los puntos del diagrama de dispersión caen todos sobre una misma línea recta.

La demostración de estas propiedades es mas bien larga y tediosa, por lo que será omitida acá.

d. La interpretación del coeficiente de correlación lineal como medida de la interrelación lineal entre dos variables es en esencia una interpretación puramente matemática, y está desprovista de toda connotación causa – efecto. Así por ejemplo, la cantidad de llamadas telefónicas que se inician en Bs. As. entre las 11 y 12 de la mañana y la cantidad de huevos que ponen las gallinas en el campo en dicho período, tienen una fuerte correlación lineal positiva a pesar de que uno de estos hechos no tiene ninguna influencia sobre el otro.

e. En cualquier problema que concierna a la correlación lineal, el valor r puede ser considerado como una muestra tomada de una población. Por ejemplo si los datos indicados en la figura CR I.a corresponden a 30 estudiantes tomados al azar dieron un coeficiente de correlación lineal igual a r_1, otras muestras darán valores r_2, r_3, etc.
La población así muestreada tiene una distribución de probabilidad intrínseca $D_{E^{XY}}$, y supóngase que de dicha distribución surja un parámetro ρ que indique el grado de correlación lineal verdadero entre las variables X e Y.
Puede demostrarse que el valor r indicado en **[1]** es una estimación máxima verosímil de dicho parámetro ρ.

f. Sean dos variables X e Y entre las cuales exista una cierta interrelación lineal, la cual es medida por el coeficiente de correlación lineal r.
Supóngase que además del grado de interrelación lineal interese conocer el valor que asumirá una de las variables conocido el valor que asumió la otra.
Por ejemplo, supóngase que en la tabla de la figura CR I.a los valores x_i correspondan al rendimiento de un hombre como alumno y los valores y_i correspondan a su posterior rendimiento como profesional.
A un posible empleador del recién graduado le interesaría mucho poder predecir el eventual rendimiento del candidato en base a su rendimiento como alumno.
El coeficiente r "pelado" es incapaz de efectuar dicha predicción, para efectuar la cual será necesario usar las técnicas de regresión a ser consideradas en los próximos párrafos.

CR II

Curvas de aproximación

a. Dado un diagrama de dispersión, a menudo interesa conocer una curva que se aproxime lo mejor posible a los puntos del mismo.
A dicha curva se la llamará curva de aproximación.
A menudo es posible visualizar por inspección cuál es el tipo de curva que mejor se aproximará (por ejemplo la recta de la figura CR II.a y la parábola de la figura CR II.b), y a veces el diagrama de dispersión tomará un aspecto de "perdigonada" para la cual aparentemente no hay ninguna curva de aproximación que tenga sentido (ver la figura CR II.c).

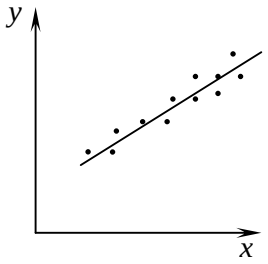

Fig. CR II.a

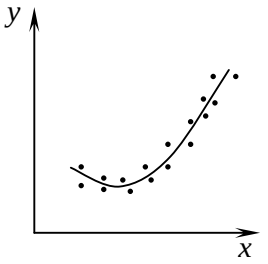

Fig. CR II.b

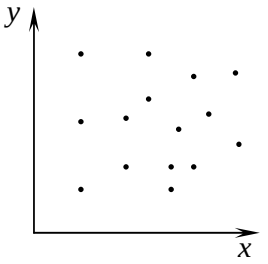

Fig. CR II.c

b. A veces el tipo de curva de aproximación adecuada al caso es elegida "a ojo", pero a menudo dicha elección proviene de algún conocimiento previo sobre una probable relación existente entre las variables.
Las ecuaciones de las curvas de aproximación mas comunes son:

Línea recta : $y^* = a_0 + a_1 \cdot x$
Parábola : $y^* = a_0 + a_1 \cdot x + a_2 \cdot x^2$
Polinomio de grado n : $y^* = a_0 + a_1 \cdot x + a_2 \cdot x^2 + \ldots + a_n \cdot x^n$
Exponencial : $y^* = k\, e^x$

c. El caso en que la curva de aproximación sea una recta es el que ulteriormente será mas fácil de tratar.
Existen curvas de aproximación que no son rectas pero que mediante adecuados cambios de variables se transforman en rectas.
A estas curvas de aproximación no lineales susceptibles de ser linealizadas se las llamará intrínsecamente lineales.
Evidentemente, en el caso de que los resultados experimentales conduzcan a un diagrama de dispersión al cual corresponde una curva de aproximación intrínsecamente lineal, aplicando los cambios de variables del caso a los resultados experimentales se obtendrá un diagrama de dispersión al cual corresponderá una curva de aproximación rectilínea.

d. En el caso de que nada sugiera el tipo de curva de aproximación a adoptar, por ejemplo el caso de la "perdigonada" de la figura CR II.c, la política que generalmente se adopta es suponer una línea recta como curva de aproximación.

CR III

Método de los mínimos cuadrados

a. Una vez decidido el tipo de curva que se adoptará como curva de aproximación, quedan por definir los coeficientes de la ecuación correspondiente que den el mejor ajuste posible de la curva al diagrama de dispersión.
Considérese el caso de la figura CR III.a, en la cual los puntos del diagrama de dispersión son (x_1 , y_1), …, (x_n , y_n). Sea C una curva de aproximación del tipo elegido, pero de la cual todavía no se conocen los valores de sus coeficientes.

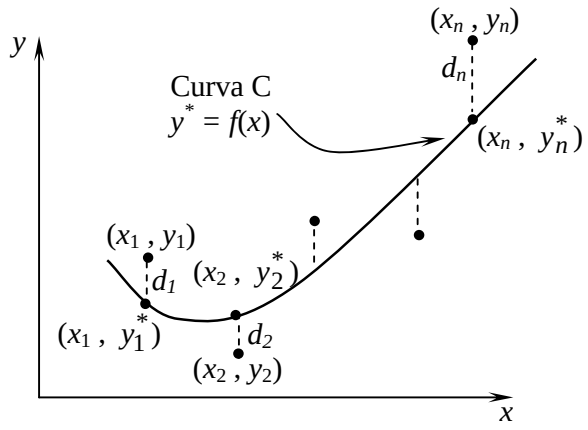

Fig. CR III.a

Para $x = x_1$ habrá una desviación d_1 entre y_1 y el valor y_1^* correspondiente a la curva C. Es decir que $d_1 = y_1 - y_1^*$. A este valor d_1 se lo llamará desviación, error, o valor residual en x_1. Este valor puede ser positivo, negativo o nulo.
Análogamente, asociadas a $x = x_2$, …, $x = x_n$, existirán desviaciones d_2, …, d_n.
Se define que la magnitud de la desviación del conjunto de los n valores y_1, …, y_n de la curva de aproximación está dada por la magnitud:

$$D = d_1^2 + ... + d_n^2 \qquad \textbf{[1]}$$

El problema queda ahora reducido a encontrar los coeficientes de un tipo de curva de la C que hagan mínimo el valor D. Una vez determinados estos valores, a la curva correspondiente se la llamará curva de regresión de *Y* sobre *X*.

b. Si en vez de considerar el procedimiento recién indicado se hubieran tomado desviaciones horizontales en vez de verticales (ver figura CR III.b), se encontraría una curva C´ de regresión de *X* sobre *Y*.

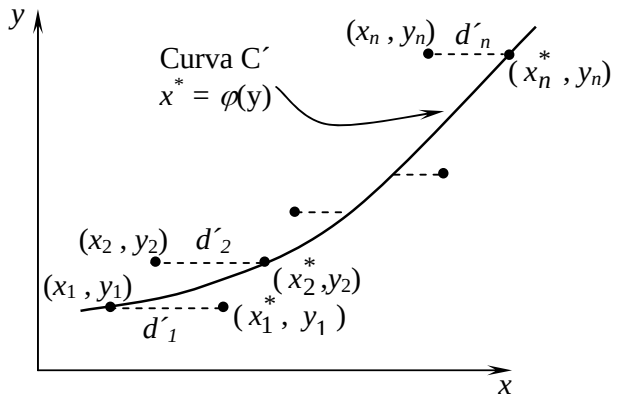

Fig. CR III.b

Notar que:
1º. La curva C de regresión de *Y* sobre *X* es tal que $y^* = f(x)$
La curva C´ de regresión de *X* sobre *Y* es tal que $x^* = \varphi(y)$
2º. Salvo casos excepcionales, estas dos curvas de regresión no coinciden.

c. Surge ahora una pregunta:
Porque en vez de considerar desviaciones "verticales" tal como usadas en **a.** o desviaciones "horizontales" tal como usadas en **b.** no se consideran desviaciones perpendiculares a los puntos a la curva.
La respuesta es muy sencilla: El trato ulterior del problema sería mucho mas complicado.

CR IV

Rectas de regresión

a. Sea un diagrama de dispersión. Se buscará la recta que mejor se adapte a dicha dispersión en el sentido de hacer mínima a la magnitud *D* definida en **[1]** de CR III.
Se deja constancia de que no es indispensable que los puntos del diagrama tengan una tendencia rectilínea.

b. Es conveniente, a fin de simplificar los cálculos, expresar la ecuación de dicha recta como:

$$y^* = a_0 + a_1(x - \bar{x}) \qquad \textbf{[1]}$$

donde los coeficientes a_0 y a_1 son por el momento desconocidos.
Los valores de a_0 y a_1 que hacen mínima a la magnitud:

$$D = d_1^2 + \ldots + d_n^2 = \sum_{i=1}^{n} (y_i - y_i^*)^2 = \sum_{i=1}^{n} \left[y_i - a_0 - a_1 (x_i - \bar{x}) \right]^2$$

Ver **[1]**

han de ser tales que:

$$\begin{cases} \dfrac{\partial D}{\partial a_0} = \displaystyle\sum_{i=1}^{n} 2\big[y_i - a_0 - a_1(x_i - \bar{x})\big](-1) = 0 \\ \dfrac{\partial D}{\partial a_1} = \displaystyle\sum_{i=1}^{n} 2\big[y_i - a_0 - a_1(x_i - \bar{x})\big]\big[-(x_i - \bar{x})\big] = 0 \end{cases}$$

Simplificando estas expresiones se obtiene:

$$\begin{cases} na_0 + a_1 \displaystyle\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} y_i = n\bar{y} \\ a_0 \displaystyle\sum_{i=1}^{n}(x_i - \bar{x}) + a_1 \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \bar{x})y_i \end{cases}$$

y teniendo en cuenta que $\displaystyle\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ resulta:

$$a_0 = \bar{y} \qquad ; \qquad a_1 = \frac{\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \textbf{[2]}$$

y por **[1]** resulta entonces que la recta de regresión óptima de *Y* sobre *X* será:

$$y^* = \bar{y} + \frac{\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2}(x - \bar{x}) \qquad \textbf{[3]}$$

c. Como:

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i + \bar{y}\underbrace{\sum_{i=1}^{n}(x_i - \bar{x})}_{=0} = \sum_{i=1}^{n}(x_i - \bar{x})y_i$$

resulta que la expresión **[3]** puede ser puesta bajo cualquiera de las siguientes formas:

$$y^* = \bar{y} + \frac{\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2}(x - \bar{x}) \qquad \textbf{[4]}$$

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

$= s_Y$ (arriba)

$= r$ (ver **[1]** de CR I) (abajo)

$= s_X$ (abajo)

$$y^* = \bar{y} + r\,\frac{s_Y}{s_X}(x-\bar{x}) \qquad \textbf{[5]}$$

$$\frac{(y^*-\bar{y})}{s_Y} = r\frac{(x-\bar{x})}{s_X} \qquad \textbf{[6]}$$

d. Por un procedimiento análogo al recién desarrollado se obtendría que la recta de regresión de X sobre Y es:

$$x^* = \bar{x} + r\,\frac{s_X}{s_Y}(y-\bar{y}) \qquad \textbf{[7]}$$

o, bajo otra forma:

$$\frac{(x^*-\bar{x})}{s_X} = r\frac{(y-\bar{y})}{s_Y} \qquad \textbf{[8]}$$

e. Observando las ecuaciones **[5]** y **[7]** resulta que:

Pendiente de la recta de regresión de Y sobre $X = r\,\frac{s_Y}{s_X}$

Pendiente de la recta de regresión de X sobre $Y = r\,\frac{s_X}{s_Y}$

f. Por **[1]** de CR III y por **[5]** se tiene que:

$$D = \sum_{i=1}^{n}(y_i - y_i^*)^2 = \sum_{i=1}^{n}\left[(y_i - \bar{y}) - r\frac{s_Y}{s_X}(x_i - \bar{x})\right]^2 =$$

$$= \sum_{i=1}^{n}\left[(y_i - \bar{y})^2 + r^2\frac{s_Y^2}{s_X^2}(x_i - \bar{x})^2 - 2r\frac{s_Y}{s_X}(x_i - \bar{x})(y_i - \bar{y})\right] =$$

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

(Ver **[1]** de CR I)

$$= n\,s_Y^2 + n\,r^2\,s_Y^2 - 2\,n\,s_Y^2\,r^2 = n\,s_Y^2(1 + r^2 - 2r^2) = n\,s_Y^2(1 - r^2)$$

Resumiendo:

$$D = n\,s_Y^2(1 - r^2) \qquad \textbf{[9]}$$

Este valor ha sido hallado considerando la regresión de *Y* sobre *X*.
Si se hubiera considerado la regresión de *X* sobre *Y* se habría hallado:

$$D' = n\,s_X^2(1 - r^2) \qquad \textbf{[10]}$$

valor que diferirá de *D* si $s_X \neq s_Y$.
Observar que si fuera $r = 1$ (todos los puntos sobre una misma recta) resultaría $D = D' = 0$.

g. En todo lo antedicho se ha supuesto que tanto los valores x_i como los y_i son valores asumidos por variables aleatorias *X* e *Y*, pero los razonamientos hechos son asimismo válidos para el caso en que una de las variables asuma valores controlados por el operador.
Por ejemplo, supóngase que se forme un grupo de 1000 personas de 40 años de edad, otro de 1000 personas de 41 años, **…**, y otro de 1000 personas de 60 años, y que se halle la presión arterial promedio de cada grupo.
Haciendo corresponder la variable *X* a la edad y la variable *Y* a la presión arterial promedio, se tendrá que los valores asumidos por *X* son controlados y que los asumidos por *Y* son aleatorios.

h. El objetivo principal del análisis de la regresión consiste en obtener predicciones del valor que asumirá la variable *Y* para un valor determinado de *X* (o viceversa). Esto es válido no solo para el caso de una regresión lineal, sino también para cualquier otro tipo de regresión.
Considérese por el momento el caso de una regresión lineal. Sea **[5]** la recta de regresión lineal correspondiente al caso. Dado un valor *x*, se tiene que y^* es la estimación máximo verosímil del valor que asumirá *Y* cuando *X* asuma el valor *x*.
Por otra parte, como los valores (x_i , y_i) no se ajustan exactamente a la antedicha recta de regresión, la antedicha estimación máximo verosímil está sujeta a un cierto error.
El así llamado error típico de estimación de *Y* sobre *X* está definido por la expresión:

$$s_{YX} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_i^*)^2}{n}} = \sqrt{\frac{D}{n}} = \sqrt{s_Y^2(1-r^2)} = s_Y\sqrt{(1-r^2)} \qquad \textbf{[11]}$$

Ver **[9]**

Ver **[1]** de CR III

Este error típico de estimación tiene propiedades análogas a la de la desviación típica. Por ejemplo, si se trazan pares de rectas paralelas a la recta de regresión a distancias verticales s_{YX} (2 s_{YX}) para *n* grande se encontrará que entre dicho par de rectas estará aproximadamente el 68 % (el 95 %) de los puntos del diagrama de dispersión.

CR V

Aplicaciones

CRV 1

a. La tabla de la figura CR V.a da en Kg. los pesos *X* e *Y* de una muestra de 12 padres y sus hijos varones mayores, ambos a la edad de 18 años.

Peso *X* del padre	65	63	67	64	68	62	70	66	68	67	69	71
Peso *Y* del hijo	68	66	68	65	69	66	68	65	71	67	68	70

Fig. CR V.a

Se pide hallar:
- El coeficiente de correlación lineal entre *X* e *Y*.
- La recta de regresión de *Y* sobre *X*.
- La recta de regresión de *X* sobre *Y*.
- El error típico de estimación de *Y* sobre *X*.

b.

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
65	68	-1,7	0,4	2,89	0,16	-0,68
63	66	-3,7	-1,6	13,69	2,56	5,92
67	68	0,3	0,4	0,09	0,16	0,12
64	65	-2,7	-2,6	7,29	6,76	7,02
68	69	1,3	1,4	1,64	1,96	1,82
62	66	-4,7	-1,6	22,09	2,56	7,52
70	68	3,3	0,4	10,89	0,16	1,32
66	65	-0,7	-2,6	0,49	6,76	1,82
68	71	1,3	3,4	1,69	11,56	4,42
67	67	0,3	-0,6	0,09	0,36	-0,18
69	68	2,3	0,4	5,29	0,16	0,92
71	70	4,3	2,4	18,49	5,76	10,32
$\sum = 800$	$\sum = 811$			$\sum = 84{,}68$	$\sum = 38{,}92$	$\sum = 40{,}34$

Fig. CR V.b

$$\bar{x} = \frac{800}{12} = 66{,}7 \quad ; \quad \bar{y} = \frac{811}{12} = 67{,}6 \quad ; \quad s_X = \sqrt{\frac{84{,}68}{12}} = 2{,}656 \quad ; \quad s_Y = \sqrt{\frac{38{,}92}{12}} = 1{,}8$$

$$r = \frac{\frac{1}{12}40{,}34}{2{,}656 \cdot 1{,}8} = 0{,}703 \qquad \textbf{[1]}$$

Recta de regresión de Y sobre X (ver **[5]** de CR IV):

$$y^* = 67{,}6 + 0{,}703 \cdot \frac{1{,}8}{2{,}656}(x - 66{,}7) = 35{,}82 + 0{,}476\, x \qquad \textbf{[2]}$$

Recta de regresión de X sobre Y (ver **[7]** de CR IV):

$$x^* = 66{,}7 + 0{,}703 \cdot \frac{2{,}656}{1{,}8}(y - 67{,}6) = -3{,}42 + 1{,}0373\, y \qquad \textbf{[3]}$$

Ver figura CR V.c.

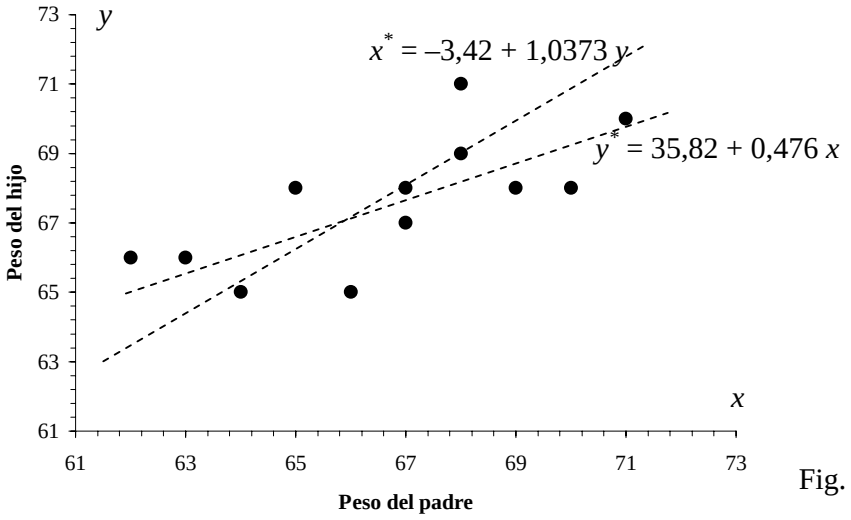

Fig. CR V.c

c. Error típico de estimación de Y sobre X (ver **[11]** de CR IV)

$$s_{YX} = 1{,}8\ \sqrt{1 - (0{,}703)^2} = 1{,}28$$

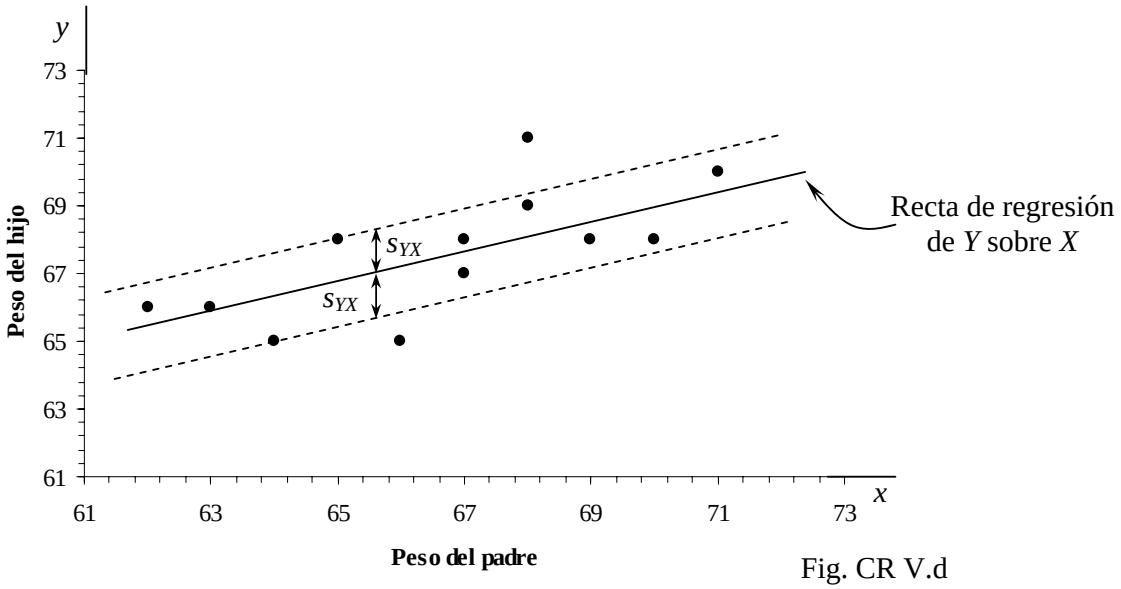

Fig. CR V.d

La recta de regresión indicada en **[2]** está dibujada en trazo grueso en la figura CR V.d. Sus paralelas a distancia vertical s_{YX} = 1,28 están dibujadas en trazo punteado. Se ve en dicha figura que 7 de los 12 puntos caen entre dichas paralelas y 2 aparecen sobre ellas. Una aritmética mas fina (mas decimales) revelaría que uno de esos dos también caería entre las paralelas. Por lo tanto $^{8}/_{12}$ = 66,66 % de los puntos caen entre las paralelas, obteniéndose así un porcentaje próximo al 68 % predicho en **h.** de CR IV. Esta diferencia es debida a la escasa cantidad de puntos del diagrama de dispersión considerado.

CRV 2

a. Supóngase que interese calcular la aceleración de la gravedad midiendo con un cronómetro de mano los períodos de oscilación de péndulos de distintas longitudes que describen oscilaciones pequeñas.
Según visto en física se tiene que:

$$T = 2\pi\sqrt{\frac{L}{g}} = \frac{2\pi}{\sqrt{g}}L^{1/2} \qquad \textbf{[1]}$$

siendo:
T = Período de oscilación (seg)
L = Longitud del péndulo (m)
g = Aceleración de la gravedad ($^{m}/_{seg^2}$)

Evidentemente, si se pudiera medir con absoluta precisión el período de oscilación de un péndulo de una longitud determinada se podría determinar a g con una única medición, pero dado lo rudimentario del método usado para medir los tiempos de oscilación en este experimento, es inevitable que los resultados obtenidos no serán coincidentes para distintas longitudes del péndulo considerado.

b. Poniendo en **[1]**:

$y = T$ $x = L$ $k = \frac{2\pi}{\sqrt{g}}$ $r = ½$ **[2]**

dicha fórmula toma el aspecto:

$$y = k\,x^r \qquad \textbf{[3]}$$

Efectuando el experimento para diversas longitudes del péndulo se obtiene un cierto diagrama de dispersión, y por conocimiento previo del fenómeno físico involucrado, se sabe que la curva de aproximación adecuada al caso será de la forma:

$$y^* = k\,x^r \qquad \textbf{[4]}$$

Esta curva de aproximación es intrínsecamente lineal (ver CR II.c) ya que haciendo el cambio de variables:

$$x' = lg_e x \qquad y' = lg_e y \qquad \textbf{[5]}$$

se obtiene la curva de aproximación rectilínea:

$$y'^* = lg_e k + r\,x' \qquad \textbf{[6]}$$

c. Sean los datos experimentales "crudos" indicados en la figura CR VI.a a los cuales corresponden los datos modificados según **[5]** indicados en la figura CR VI.b.

$x' = L$ (m)	$y = T$ (seg)
1,025	2,025
0,805	1,813
0,745	1,739
0,675	1,650
0,615	1,573
0,515	1,441
0,435	1,338
0,370	1,232
0,325	1,149
0,270	1,051
0,205	0,912

Fig. CR VI.a

$x' = lg_e x = lg_e L$	$y' = lg_e y = lg_e T$
0,0246926	0,7055697
-0,216913	0,5949829
-0,294371	0,5533102
-0,3930425	0,5007753
-0,486133	0,4529846
-0,6635883	0,3653373
-0,8324092	0,2911760
-0,9942522	0,2086389
-1,1239301	0,1388920
-1,3093333	0,0497421
-1,5847453	-0,0921152

Fig. CR VI.b

Según indicado en **[3]** de CR IV se tiene que:

$$y'^{*} = \bar{y}' + \frac{\sum_{i=1}^{11}(x'_i - \bar{x}')y'_i}{\sum_{i=1}^{11}(x'_i - \bar{x}')^2}(x' - \bar{x}') = \underbrace{\bar{y}' - \frac{\sum_{i=1}^{11}(x'_i - \bar{x}')y'_i}{\sum_{i=1}^{11}(x'_i - \bar{x}')^2}\bar{x}'}_{a_0} + \underbrace{\frac{\sum_{i=1}^{11}(x'_i - \bar{x}')y'_i}{\sum_{i=1}^{11}(x'_i - \bar{x}')^2}}_{a_1}x' \qquad \textbf{[7]}$$

Resumiendo:

$$y'^{*} = a_0 + a_1 x'$$

y comparando con **[6]** resulta que:

$$a_0 = lg_e k = lg_e \frac{2\pi}{\sqrt{g}}$$

Ver **[2]**

y por lo tanto:

$$e^{a_0} = \frac{2\pi}{\sqrt{g}} \Rightarrow g = \left(\frac{2\pi}{e^{a_0}}\right)^2$$

Efectuando los cálculos indicados en **[7]** (evidentemente con la ayuda de una computadora), en base a los datos de la figura CR VI.b se llega a que:

$$a_0 = 0{,}7172 \Rightarrow g = 9{,}406$$

lo cual constituye una aproximación razonable de la magnitud de la aceleración de la gravedad, visto y considerando la precariedad del método empleado.

CR VII

Regresión curvilínea

a. Si el diagrama de dispersión indica que una recta no se adaptará satisfactoriamente a los datos obtenidos debido a la no linealidad de la tendencia observada, debe adoptarse otro tipo de curva de aproximación.
Si no hay razones básicas que sugieran un cierto tipo de curva, generalmente se usarán curvas polinómicas por su simplicidad y flexibilidad.
Por inspección puede a menudo determinarse el grado de la curva polinómica mas sencilla que se ajuste a los valores obtenidos.

b. A título de ejemplo se supondrá que la curva del caso es una parábola cuya ecuación es:

$$y^{*} = a_0 + a_1 x + a_2 x^2$$

Entonces se tiene que:

$$D = d_1^2 + ... + d_n^2 = \sum_{i=1}^{n} (y_i - y_i^*)^2 = \sum_{i=1}^{n} \left[y_i - a_0 - a_1 x_i - a_2 x_i^2 \right]^2$$

Los valores de a_0, a_1 y a_2 que hacen mínima a D han de ser tales que:

$$\begin{cases} \dfrac{\partial D}{\partial a_0} = \displaystyle\sum_{i=1}^{n} 2\left[y_i - a_0 - a_1 x_i - a_2 x_i^2 \right](-1) = 0 \\ \dfrac{\partial D}{\partial a_1} = \displaystyle\sum_{i=1}^{n} 2\left[y_i - a_0 - a_1 x_i - a_2 x_i^2 \right](-x_i) = 0 \\ \dfrac{\partial D}{\partial a_2} = \displaystyle\sum_{i=1}^{n} 2\left[y_i - a_0 - a_1 x_i - a_2 x_i^2 \right](-x_i^2) = 0 \end{cases}$$

Simplificando estas expresiones se obtiene:

$$\begin{cases} n a_0 + a_1 \displaystyle\sum_{i=1}^{n} x_i + a_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i \\ a_0 \displaystyle\sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 + a_2 \sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} x_i y_i \\ a_0 \displaystyle\sum_{i=1}^{n} x_i^2 + a_1 \sum_{i=1}^{n} x_i^3 + a_2 \sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} x_i^2 y_i \end{cases}$$

Resolviendo este sistema se hallan los coeficientes a_0, a_1 y a_2 que proporcionan el mejor ajuste posible de la parábola a los datos del problema.

c. La generalización de lo visto en **b.** al caso en que la curva de aproximación sea un polinomio de grado k es obvia.
Se obtendrá un sistema:

$$\begin{cases} n a_0 + a_1 \displaystyle\sum_{i=1}^{n} x_i + \ldots + a_k \sum_{i=1}^{n} x_i^k = \sum_{i=1}^{n} y_i \\ a_0 \displaystyle\sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 + ... + a_k \sum_{i=1}^{n} x_i^{k+1} = \sum_{i=1}^{n} x_i y_i \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_0 \displaystyle\sum_{i=1}^{n} x_i^k + a_1 \sum_{i=1}^{n} x_i^{k+1} + ... + a_k \sum_{i=1}^{n} x_i^{2k} = \sum_{i=1}^{n} x_i^k y_i \end{cases}$$

y resolviendo este sistema se hallarán los valores a_0, a_1, …, a_k que implican el mejor ajuste posible de la curva polinómica:

$$y^* = a_0 + a_1 x + \ldots + a_k x^k$$

a los datos del problema.

c. En este caso de regresión polinómica, el error típico de estimación de *Y* sobre *X* viene dado por la expresión:

$$s_{YX} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - \ldots - a_k x_i^k)^2}{n}}$$

d. En **a.**, **b.** y **c.** se ha tratado el caso de una regresión polinómica de *Y* sobre *X*.
Una regresión polinómica de *X* sobre *Y* sería tratada de una manera análoga.

CR VIII

Observaciones

a. Lo visto en este capítulo constituye apenas un "viaje exploratorio" al mundo de la correlación y regresión, habiéndose tratado únicamente de introducir los conceptos fundamentales de dichos temas.
Un tratamiento completo insumiría unas 100 o mas páginas.

b. Los problemas de correlación y regresión implican por lo general una dosis inaceptable de aritmética si ésta se hace a mano.
Si se trabaja profesionalmente en este tema es prácticamente indispensable el uso de una computadora dotada de un software adecuado.
Hoy en día muchas calculadoras científicas tienen el modo "reg" que con solo ingresar los pares de datos se obtienen los parámetros de la recta de ajuste (a_0 y a_1).

c. Vale la pena citar la reflexión hecha por M. I. Moroney en su pequeño gran libro "Hechos y Estadísticas".
"La rama de la estadística que mayor similitud tiene con una máquina de hacer salchichas es el análisis de la correlación y regresión. El problema de la interpretación de los resultados es siempre mucho mas difícil que las manipulaciones estadísticas".
"Para hacer una interpretación correcta de los resultados no hay substituto para el conocimiento profundo y detallado del problema que se tiene entre manos y de sus condiciones de contorno".
"El estadístico puede ayudar al especialista en su campo, pero no puede nunca sustituirlo".
"Quien usa sin precauciones herramientas de alto filo tiene un gran riesgo de cortarse".

APÉNDICE

Curvas de aproximación intrínsecamente lineales

a. En el párrafo CR II **c.** se definió lo que son las curvas de aproximación intrínsecamente lineales.
En la tabla de la figura A.CR.a se indican algunas de las curvas de aproximación intrínsecamente lineales mas comunes junto con los correspondientes cambios de variables que determinan que a los resultados experimentales modificados por dichos cambios corresponda una curva de aproximación lineal.

Caso	Curva de aproximación intrínsecamente lineal	Cambios de variables a aplicar a los resultados experimentales "crudos"	Curva de aproximación después de efectuado el cambio de variables
(a)	$y^* = a\, e^{bx}$	$x' = x\ ;\ y' = lg_e\, y$	$y'^* = lg_e\, a + bx'$
(b)	$y^* = a\, x^b$	$x' = lg_e\, x\ ;\ y' = lg_e\, y$	$y'^* = lg_e\, a + bx'$
(c)	$y^* = a + b\, lg_e\, x$	$x' = lg_e\, x\ ;\ y' = y$	$y'^* = a + bx'$
(d)	$y^* = a + {}^{b}\!/_{x}$	$x' = {}^{1}\!/_{x}\ ;\ y' = y$	$y'^* = a + bx'$

Fig. A.CR.a

En la figura A.CR.b se han graficado las curvas de aproximación intrínsecamente lineales indicadas en la segunda columna de la tabla A.CR.a.

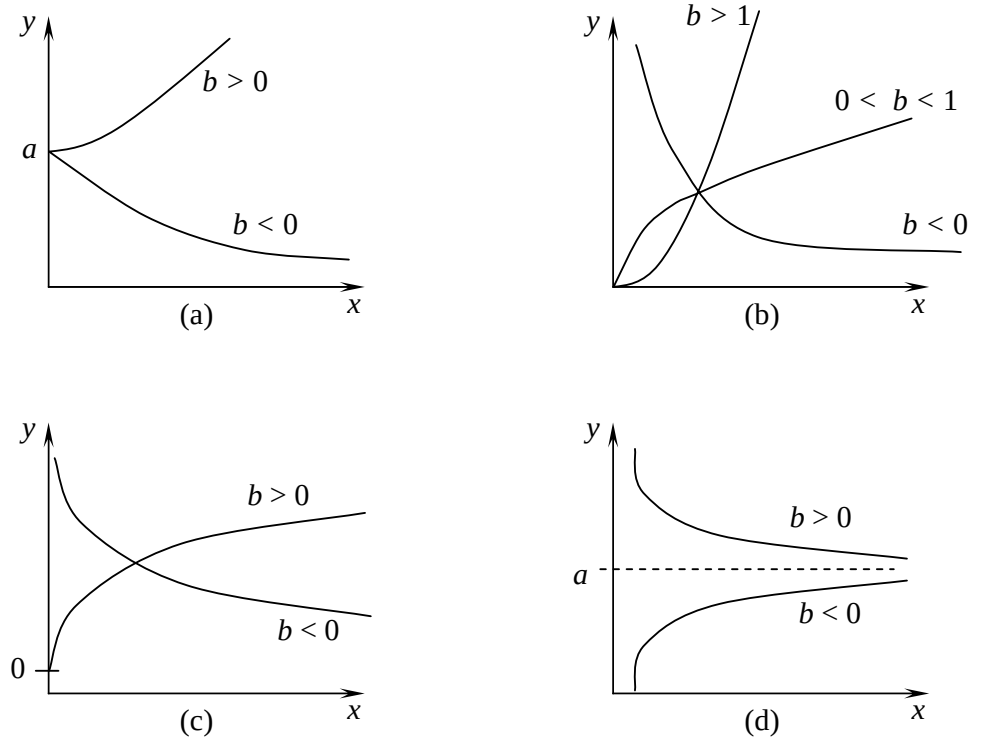

Fig. A.CR.b

Ejercicios sobre Correlación y Regresión

CR 1 Demostrar que la fórmula **[1]** de CR I (definición del coeficiente de correlación) también podría ser expresada bajo la forma (mucho mas económica en aritmética):

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

CR 2 Explicar porque no sería sorprendente encontrar una alta correlación entre el tráfico en la Panamericana y la altura de la marea en Río Gallegos. Supóngase que se hacen mediciones cada hora entre las 6 y 10 de la mañana y que la marea máxima en Río Gallegos ocurra a las 8 de la mañana.

CR 3 Cual sería el efecto en el valor *r* del coeficiente de correlación entre el peso y la estatura de los varones de todas las edades si solo fueran muestreados varones entre 20 y 25 años. Haga un diagrama de dispersión para ayudar su contestación.

CR 4 Si la recta de regresión de *Y* sobre *X* es $y^* = a_0 + a_1 x$, y la recta de regresión de *X* sobre *Y* es $x^* = b_0 + b_1 y$, probar que $a_1 b_1 = r^2$.

CR 5 La tabla adjunta muestra los índices de precios de la alimentación *X* y de los gastos médicos *Y* a los largo de 9 años. Se pide hallar:

- El coeficiente de correlación lineal de *Y* y *X*.
- La recta de regresión de *Y* sobre *X*.
- La recta de regresión de *X* sobre *Y*.
- El error típico de estimación de *Y* sobre *X*.

X	175	181	192	211	235	255	275	286	292
Y	169	185	202	219	240	266	295	329	357

CR 6 Ajustar una parábola al conjunto de datos de la tabla adjunta:

X	1,2	1,8	3,1	4,9	5,7	7,1	8,6	9,8
Y	4,5	5,9	7,0	7,8	7,2	6,8	4,5	2,7

(La cantidad de aritmética que requiere este problema implica el uso de una computadora).