Clase 2: Estadística descriptiva 2D

Matías Carrasco

1 de octubre de 2019

Índice

1. Visualizando la normal bi-variada

Supongamos que el par (X,Y) tiene distribución normal bi-variada de parámetros

$$\mu_X, \quad \sigma_X^2, \quad \mu_Y, \quad \sigma_Y^2, \quad \rho.$$

La densidad conjunta tiene una expresión matemática intimidante, que involucra una exponencial en cuyo exponente encontramos

$$\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2.$$

Por lo tanto, las curvas de nivel de la densidad normal bi-variada vienen dadas por elipses, que se obtienen igualando la expresión anterior a una constante.

Trabajando en unidades típicas

Una forma de visualizar su densidad es justamente graficando estas elipses. Para esto, hagamos primero el cambio de coordenadas

$$u = \frac{x-\mu_X}{\sigma_X}, \quad v = \frac{y-\mu_Y}{\sigma_Y}.$$

En estas coordenadas la ecuación anterior queda $u^2 - 2\rho uv + v^2$. Sea k una constante positiva, notar que la ecuación de la elipse

$$u^2 - 2\rho uv + v^2 = k^2$$

se puede escribir en forma matricial como

$$\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} \cdot \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} u-\rho v \\ -\rho u+v \end{pmatrix} \cdot \begin{pmatrix} u \\ v \end{pmatrix} = u^2 - 2\rho uv + v^2 = k^2$$

en donde el punto indica el producto escalar.

Si llamamos

$$R = \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}$$

a la matriz de correlación, queremos dibujar las soluciones a la ecuación $R(u,v)\cdot(u,v) = k^2$.

Calculemos los valores y vectores propios de R. El polinomio característico es

$$\begin{aligned} |R-\lambda I| &= \left| \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right| \\ &= \begin{vmatrix} 1-\lambda & -\rho \\ -\rho & 1-\lambda \end{vmatrix} = (1-\lambda)^2 - \rho^2 \end{aligned}$$

Igualando a cero deducimos que $\lambda = 1 \pm \rho$. Es decir, los valores propios de R son

$$\begin{aligned} \lambda_1 &= 1+\rho \\ \lambda_2 &= 1-\rho \end{aligned}$$

Por otro lado, los vectores propios los obtenemos resolviendo el sistema

$$\begin{pmatrix} 1-\lambda & -\rho \\ -\rho & 1-\lambda \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

o de forma equivalente

$$\begin{aligned} (1-\lambda)u - \rho v &= 0 \\ -\rho u + (1-\lambda)v &= 0 \end{aligned}$$

Despejando v de la primera ecuación tenemos $v = \frac{1-\lambda}{\rho}u$. Pero $1-\lambda = \pm\rho$, por lo que los subespacios propios están dados por las rectas ortogonales

$$\lambda_1 : v = -u,\ e_1 = \left(\frac{-1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right); \quad \lambda_2 : v = u,\ e_2 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right).$$

Es decir, R es diagonal en los ejes coordenados rotados 45 grados. Un poco de trigonometría, y vemos que las coordenadas de un punto (u,v) en dichos ejes son

$$(u,v) = \left(\frac{v-u}{\sqrt{2}}\right) e_1 + \left(\frac{v+u}{\sqrt{2}}\right) e_2$$

Entonces

$$R(u,v) = (1+\rho)\left(\frac{v-u}{\sqrt{2}}\right) e_1 + (1-\rho)\left(\frac{v+u}{\sqrt{2}}\right) e_2$$

y por lo tanto

$$R(u,v)\cdot(u,v) = (1+\rho)\left(\frac{v-u}{\sqrt{2}}\right)^2 + (1-\rho)\left(\frac{v+u}{\sqrt{2}}\right)^2$$

En el eje $u = v$, la ecuación de la elipse implica

$$k^2 = (1+\rho)\left(\frac{v-u}{\sqrt{2}}\right)^2 + (1-\rho)\left(\frac{v+u}{\sqrt{2}}\right)^2 = 2(1-\rho)u^2$$

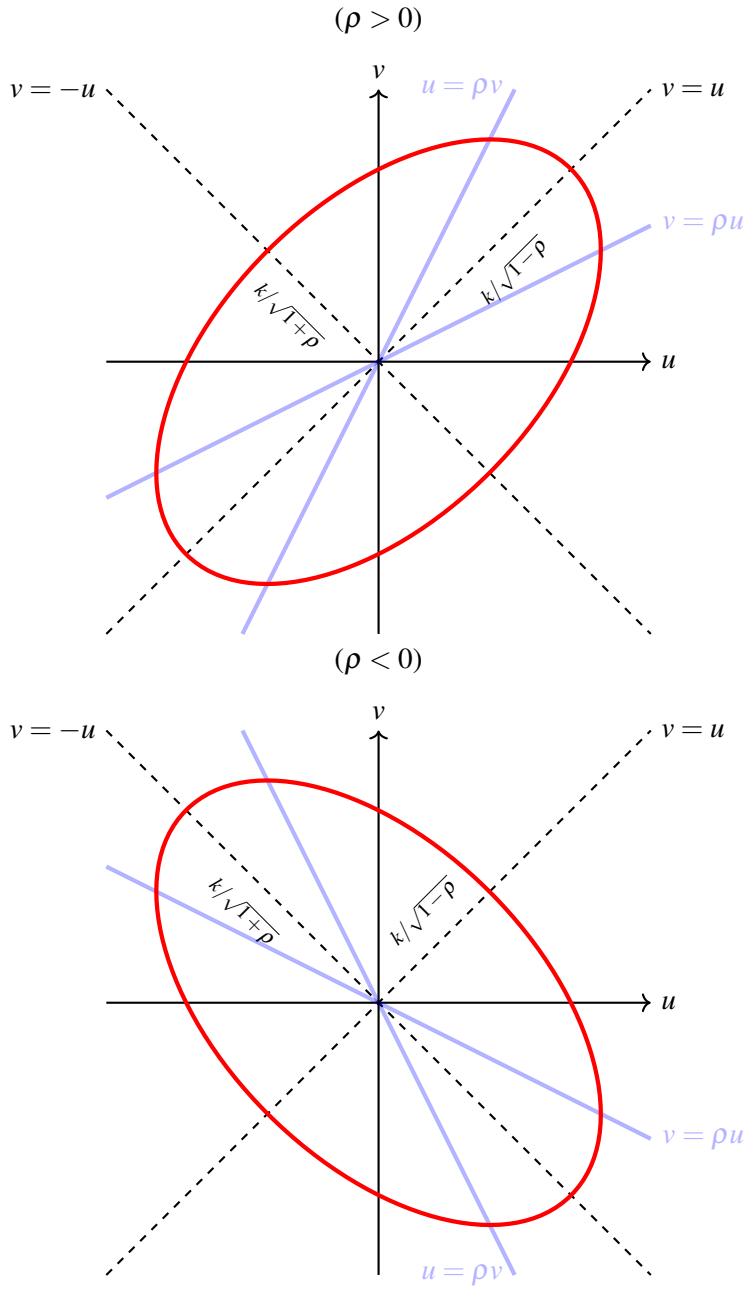


Figura 1: La elipse $u^2 - 2\rho uv + v^2 = k^2$ en rojo y las rectas de regresión $v = \rho u$, $u = \rho v$ en azul transparente.

de donde $u = \pm k/\sqrt{2(1-\rho)}$. Esto quiere decir que la mitad de la longitud del eje de la elipse en la dirección $u = v$ es $k/\sqrt{1-\rho}$. Del mismo modo se ve que la mitad de la longitud del eje de la elipse en la dirección $u = -v$ es $k/\sqrt{1+\rho}$. Cuál de estos ejes es mayor depende del signo de ρ. Ver la Figura 1.

Notar que el único caso con simetría rotacional es cuando $\rho = 0$, en el cual la elipse es de hecho un círculo. Notar también que las rectas de regresión $v = \rho u$ y $u = \rho v$ pasan por los puntos medios de la elipse, ver la Figura 2

Podemos dibujar un cuadrado que sea tangente a la elipse, esto nos ayuda a dibujarla. Los

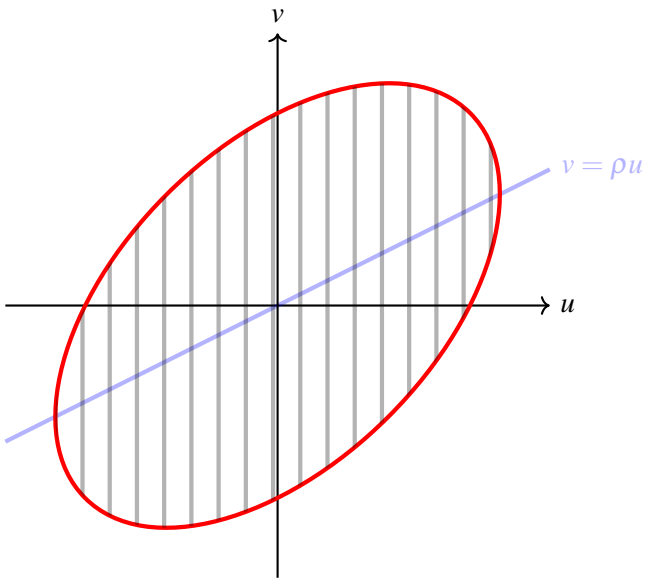


Figura 2: La recta de regresión pasa por los puntos medios de la elipse.

puntos de tangencia se dan en los cortes de la elipse con las rectas de regresión $v = \rho u$ y $u = \rho v$. Para hallar la intersección, basta por ejemplo poner $v = \rho u$ en la ecuación de la elipse:

$$\begin{aligned} k^2 &= (1+\rho)\left(\frac{v-u}{\sqrt{2}}\right)^2 + (1-\rho)\left(\frac{v+u}{\sqrt{2}}\right)^2 \\ &= \left[\frac{(1+\rho)(1-\rho)^2}{2} + \frac{(1+\rho)^2(1-\rho)}{2}\right] u^2 = (1-\rho^2)u^2. \end{aligned}$$

Es decir, los lados verticales del cuadrado pasan por $\pm k/\sqrt{1-\rho^2}$. Por simetría, los lados horizontales deben pasar por las mismas alturas. Ver la Figura 3.

Hallemos ahora el valor de k tal que la elipse tiene probabilidad 90 %. Para esto recordemos que existe Z normal estándar e independiente de U tal que

$$V = \rho U + \sqrt{1-\rho^2}Z.$$

Poniendo $v = \rho u + \sqrt{1-\rho^2}z$, un cálculo sencillo muestra que la ecuación de la elipse se transforma en

$$u^2 - 2\rho uv + v^2 = (1-\rho^2)(u^2+z^2) = k^2.$$

Basta entonces hallar k tal que

$$\mathbf{P}\left(U^2 + Z^2 \leq k^2/(1-\rho^2)\right) = 0.9$$

Pero ya hemos visto que $U^2 + Z^2$ tiene distribución exponencial de media 2 (parámetro $\lambda = 1/2$). Así que

$$0.9 = 1 - e^{-k^2/2(1-\rho^2)} \Leftrightarrow k = \sqrt{1-\rho^2}\sqrt{2\ln 10} \approx 2\sqrt{1-\rho^2}.$$

Es decir, la elipse 90 % es tangente al cuadrado que pasa por los vértices $(\pm 2, \pm 2)$ y sus ejes tienen longitud mitad aproximadamente $2\sqrt{1 \pm \rho}$. Ver la Figura 4.

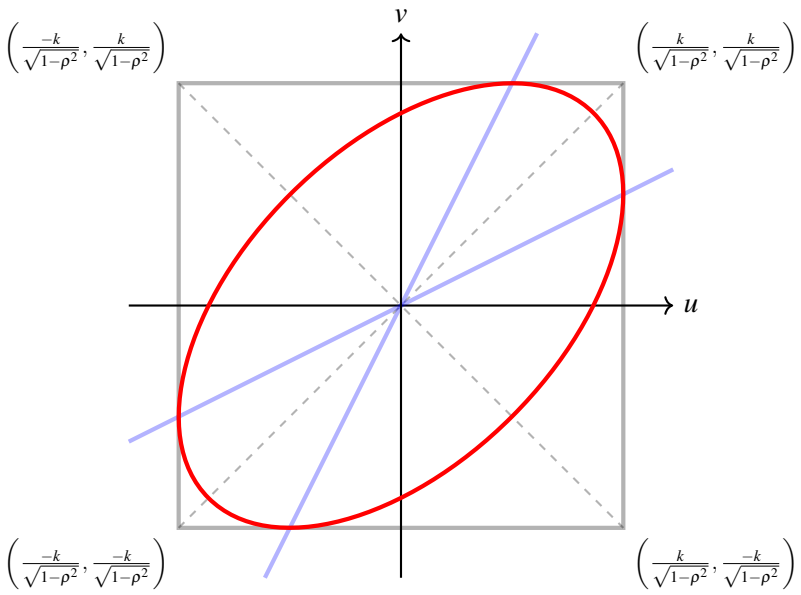


Figura 3: Cuadrado mínimo que contiene a la elipse $u^2 - 2\rho uv + v^2 = k^2$.

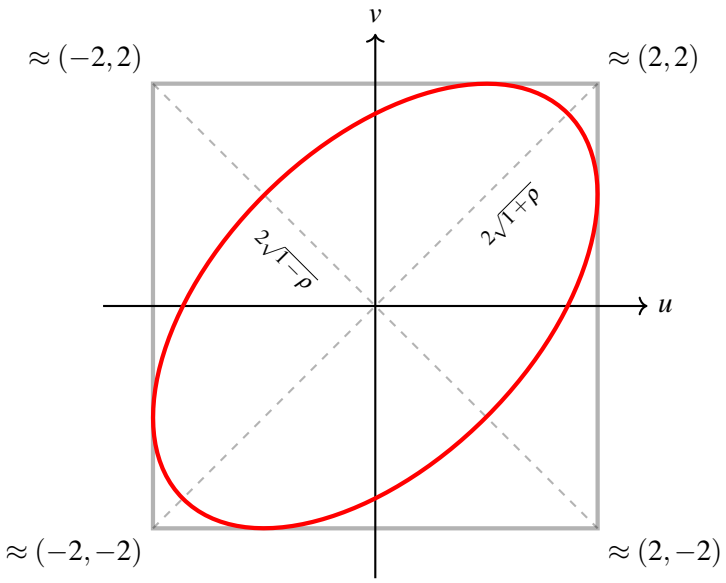


Figura 4: La elipse 90 %.

Trabajando en las unidades originales

Para obtener la elipse 90 % en las unidades originales, basta deshacer el cambio de coordenadas, recordando que

$$x = \mu_X + \sigma_X u, \quad y = \mu_Y + \sigma_Y v.$$

En las coordenadas originales, los ejes $v = \pm u$ de la elipse se transforman en las rectas de ecuaciones

$$y = \left(\frac{\sigma_Y}{\sigma_X}\right) x + \left(\mu_Y - \frac{\sigma_Y}{\sigma_X}\mu_X\right), \quad y = \left(-\frac{\sigma_Y}{\sigma_X}\right) x + \left(\mu_Y + \frac{\sigma_Y}{\sigma_X}\mu_X\right).$$

Estas rectas pasan por el punto (μ_X, μ_Y) y tienen pendiente $\pm\sigma_Y/\sigma_X$. Se las conoce con el nombre de rectas de desvíos, pues su pendiente indica le relación entre los desvíos de Y y X. No se debe confundir la recta de desvíos con la recta de regresión, cuya ecuación es muy similar pero en la cual aparece el coeficiente de correlación ρ:

$$\text{recta de regresión} : y = \left(\rho\frac{\sigma_Y}{\sigma_X}\right)x + \left(\mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X\right).$$

Tampoco se debe confundir las rectas de los desvíos con los ejes de la elipse, pues éstos deben ser perpendiculares entre sí.

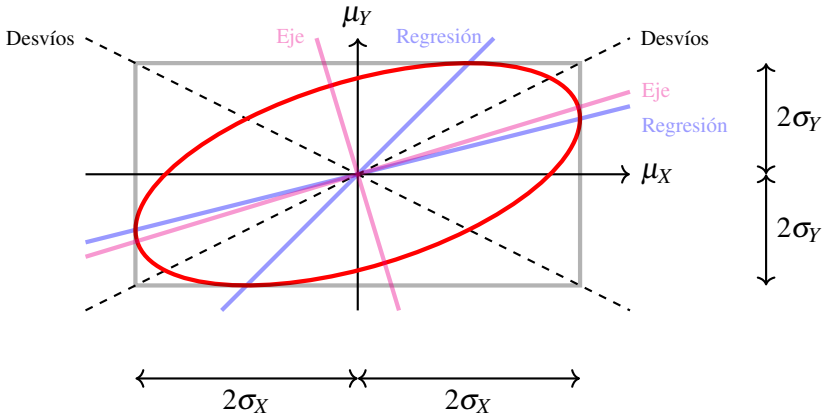


Figura 5: La elipse 90 % en las coordenadas originales (x, y).

La recta de regresión sigue teniendo la propiedad de pasar por los puntos medios de la elipse, pues esta propiedad es invariante por transformaciones lineales. También se sigue verificando que la recta de regresión pasa por los puntos de tangencia de la elipse con la caja mínima que la contiene. En estas coordenadas, la caja mínima tiene lado horizontal de longitud $4\sigma_X$ y vertical de longitud $4\sigma_Y$. Claramente la elipse y la caja están centradas en los promedios (μ_X, μ_Y).

Hallar la dirección del eje es un poco más difícil en este caso. Una forma de hacer es escribir a los puntos de la elipse en el plano (x, y) en coordenadas polares con centro en el punto (μ_X, μ_Y):

$$x - \mu_X = r\cos\theta, \quad y - \mu_Y = r\sin\theta.$$

No olvidarse que $r = r(\theta)$ depende del ángulo θ. Los ejes están en las direcciones para las cuales $r'(\theta) = 0$. Por la ecuación de la elipse, tenemos

$$\begin{aligned} k^2 &= r^2\left[\frac{1}{\sigma_X^2}\cos^2\theta - 2\frac{\rho}{\sigma_X\sigma_Y}\cos\theta\sin\theta + \frac{1}{\sigma_Y^2}\sin^2\theta\right] \\ &= \frac{r^2}{\sigma_X^2\sigma_Y^2}\left[\frac{\sigma_Y^2 - \sigma_X^2}{2}\cos 2\theta - \sigma_{XY}\sin 2\theta + \frac{\sigma_X^2 + \sigma_Y^2}{2}\right] \end{aligned}$$

en donde en la última igualdad hemos usado la identidad trigonométrica $\cos^2\theta = \frac{1+\cos 2\theta}{2}$.

Esta ecuación tiene la forma $r^2 C(\theta) = \text{cte}$. Tomando logaritmo y derivando respecto de θ, se obtiene la ecuación

$$\frac{2r'}{r} + \frac{C'}{C} = 0.$$

La ecuación $r'(\theta) = 0$ es equivalente entonces a la ecuación $C'(\theta) = 0$. Es decir, queremos resolver

$$C'(\theta) = -(\sigma_Y^2 - \sigma_X^2)\sin 2\theta - 2\sigma_{XY}\cos 2\theta = 0$$

cuya solución es

$$\tan 2\theta = \frac{2\sigma_{XY}}{\sigma_X^2 - \sigma_Y^2}$$

o de forma equivalente

$$\theta = \frac{1}{2}\arctan\left(\frac{2\sigma_{XY}}{\sigma_X^2 - \sigma_Y^2}\right).$$

Esta es la dirección de uno de los ejes de la elipse, el otro eje está en la dirección perpendicular.

A modo de ejemplo, supongamos que $\sigma_X = 1$, $\sigma_Y = 1/2$ y $\rho = 1/2$. Entonces, las direcciones de los ejes son

$$\begin{cases} \theta_1 = \frac{1}{2}\arctan\left(\frac{2\sigma_{XY}}{\sigma_X^2 - \sigma_Y^2}\right) = \frac{1}{2}\arctan\left(\frac{2(1/4)}{1-(1/4)}\right) = 0.294 \text{ rad} = 16.8^\circ \\ \theta_2 = 1.865 \text{ rad} = 106.8^\circ \end{cases}$$

Ver la Figura 5.

2. Bajando a tierra: el diagrama de dispersión

La siguiente tabla muestra los datos recolectados en un estudio que realizó Karl Pearson (1857-1936, estudiante de Francis Galton). El propósito del estudio era entender las propiedades hereditarias de la altura.

TABLE XXII.

Father's Stature and Son's Stature.

Father's Stature.

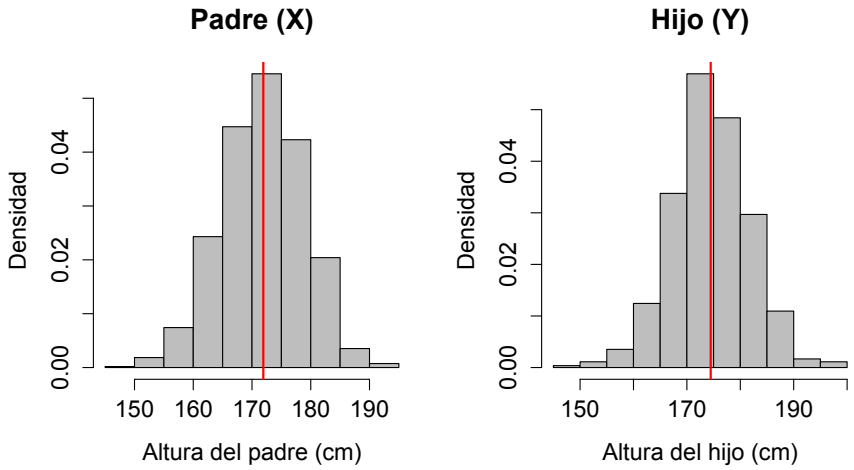
Son's Stature.	58·5—59·5	59·5—60·5	60·5—61·5	61·5—62·5	62·5—63·5	63·5—64·5	64·5—65·5	65·5—66·5	66·5—67·5	67·5—68·5	68·5—69·5	69·5—70·5	70·5—71·5	71·5—72·5	72·5—73·5	73·5—74·5	74·5—75·5	Totals
59·5—60·5	—	—	—	—	·5	·5	1	—	—	—	—	—	—	—	—	—	—	2
60·5—61·5	—	—	—	—	·5	—	—	—	1	—	—	—	—	—	—	—	—	1·5
61·5—62·5	—	·25	·25	—	·5	1	·25	·25	·5	·5	—	—	—	—	—	—	—	3·5
62·5—63·5	—	·25	·25	2·25	2·25	2	4	5	2·75	1·25	—	·25	.25	—	—	—	—	20·5
63·5—64·5	1	—	1·5	3·75	3	4·25	8	9·25	3	1·25	1·5	·75	1·25	—	—	—	—	38·5
64·5—65·5	2	·1	·5	2	3·25	9·5	13·5	10·75	7·5	5·5	3·5	2·5	—	—	—	—	—	61·5
65·5—66·5	—	·5	1	2·25	5·25	9·5	10	16·75	17·5	16	5·25	2	2·5	1	—	—	—	89·5
66·5—67·5	—	1·5	2	4.75	3·5	13·75	19·75	26·5	25·75	19·5	12·5	13·75	3·25	·5	1	—	—	148·0
67·5—68·5	—	—	1·5	2	7·5	10	10·25	24·25	31·5	23·5	29·5	13·25	8·5	9·5	2·25	—	—	173·5
68·5—69·5	—	—	1	—	5·25	5	12·75	18·25	16	24	29	21·5	10	3·5	2·25	—	1	149·5
69·5—70·5	—	—	—	—	1	2·5	5·75	18·75	11·75	19·5	22·5	19·5	14·5	6·25	3·5	1·5	1	128·0
70·5—71·5	—	—	—	—	—	3·25	5	8·75	10·75	19	14·75	20·75	10·75	8	5	1	1	108·0
71·5—72·5	—	—	—	—	—	·25	3	1.25	7	7·75	10·75	11·25	10	8·5	2·75	·5	—	63·0
72·5—73·5	—	—	—	—	—	—	·75	·75	2·5	7·5	6·5	6	7·5	6·25	3·25	·5	·5	42·0
73·5—74·5	—	—	—	—	1	—	1·5	1·5	—	5·25	2·25	2·5	6·5	3·25	3·25	—	2	29·0
74·5—75·5	—	—	—	—	—	—	—	—	—	1	2	—	2·5	·75	1·75	·5	—	8·5
75·5—76·5	—	—	—	—	—	—	—	—	—	1·25	·25	—	·5	1	1	—	—	4·0
76·5—77·5	—	—	—	—	—	—	—	—	—	1·25	·25	1	—	—	1·5	—	—	4·0
77·5—78·5	—	—	—	—	—	—	—	—	—	—	1	1	—	·25	·75	—	—	3·0
78·5—79·5	—	—	—	—	—	—	—	—	—	—	—	—	—	·25	·25	—	—	·5
Totals	3	3·5	8	17	33·5	61·5	95·5	142	137·5	154	141·5	116	78	49	28·5	4	5·5	1078

Pearson observó la estatura de 1078 pares de padres e hijos. ¿De tal palo, tal astilla?

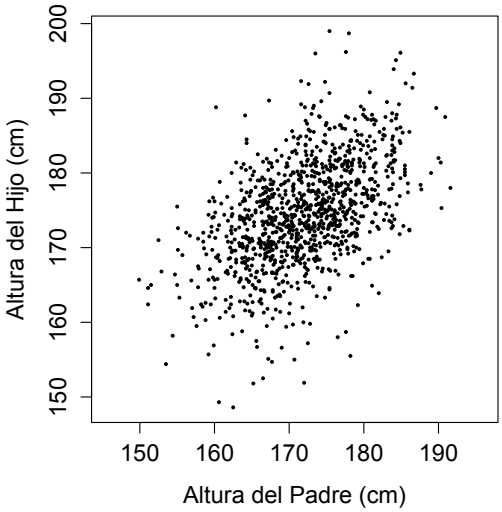
Las alturas están en pulgadas, como es usual medir la altura en Inglaterra. No es importante leer la tabla en detalle, pero uno puede reconocer la forma ovalada en la que se agrupan los

datos. Más aún, uno puede entrever una asociación entre la altura del padre y la altura del hijo: cuanto más alto es el padre más alto es el hijo y viceversa. Al menos en promedio.

La fila inferior muestra las frecuencias de la distribución de alturas de los padres. Llamemos X a esta variable. La última columna a la derecha muestra la de los hijos, variable que llamaremos Y. La tabla muestra entonces una aproximación de la distribución conjunta de (X,Y). Es mucho más fácil entender un gráfico que una tabla. Por ejemplo, las marginales de X e Y se muestran abajo.



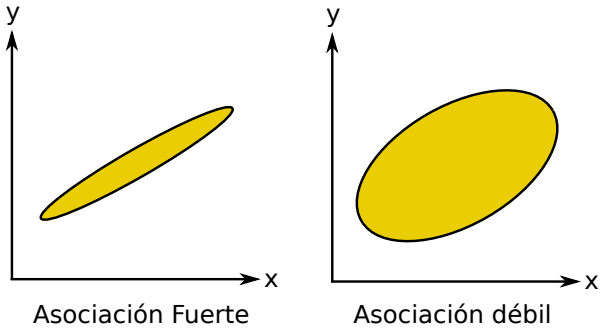
Pero en este tipo de gráficos no podemos ver la relación entre X e Y. El gráfico más común para describir la relación entre dos variables es el *diagrama de dispersión*.



Este tipo de diagrama se utiliza cuando las observaciones están apareadas. Ponemos un punto por cada par $p_i = (x_i, y_i)$, $i = 1, \ldots, n$, en donde $x_1, \ldots, x_n$ son los valores muestreados de X e $y_1, \ldots, y_n$ los de Y.

Nuestro objetivo es estimar el valor promedio de Y entre aquellos hijos que tienen padres con una determinada altura. No esperamos poder predecir mucho más, ya que por ejemplo la franja de puntos con $x_i \in [179, 181]$ tiene mucha variabilidad.

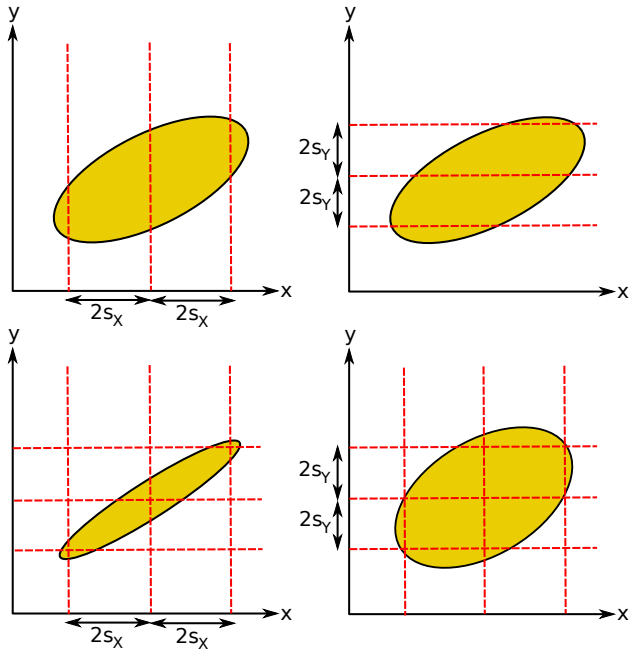
En realidad, cuán difícil sea estimar Y para valores dados de X depende de cuán fuerte sea la asociación entre ambas. Siempre es útil hacer un grueso bosquejo de un diagrama de dispersión dibujando una elipse que represente la nube de puntos. El centro de la elipse es $\overline{p} = (\overline{x}, \overline{y})$, el punto cuyas coordenadas son los promedios, pero el ancho depende de cuán fuerte sea la asociación entre X e Y.



Se puede pensar en analogía a un juego de adivinanzas. Si hay asociación fuerte entre X e Y, saber una ayuda a predecir la otra, pero si la asociación es débil, información sobre una variable no ayudará mucho a adivinar la otra.

Esquema general

Supondremos de ahora en más que el par (X,Y) se ajusta a una distribución normal bivariada. Esto nos permitirá usar algunas reglas prácticas, que estrictamente hablando son válidas solamente para la distribución normal.



¿Cómo resumir un diagrama de dispersión? La primera observación es sobre la dispersión en cada variable:

- el 95 % de las $x_i's$ caen en el intervalo $[\overline{x} \pm 2s_X]$;
- el 95 % de las $y_i's$ caen en el intervalo $[\overline{y} \pm 2s_Y]$.

Un cálculo rápido nos indica que aproximadamente $0.95^2 \approx 0.9$, es decir 90 % de los pares caen en la caja. Pero solamente las medias y los desvíos no alcanzan, falta un número que nos diga cuán fuerte es la asociación entre X e Y.

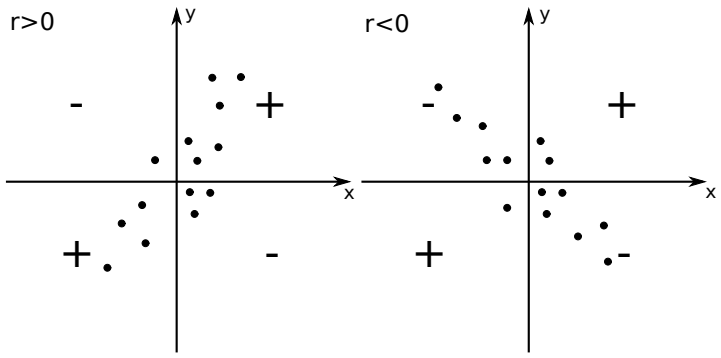
Nos está faltando el coeficiente de correlación ρ. En realidad ρ mide la fuerza de la asociación lineal entre dos variables, pero como estamos asumiendo datos normales, es equivalente a conocer la dependencia entre ellas.

Usaremos la letra r para indicar un estimador de la correlación ρ. Más precisamente definimos

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_X} \right) \left(\frac{y_i - \overline{y}}{s_Y} \right),$$

lo cual no es sorprendente por su analogía con la definición de ρ. Sin embargo daremos una justificación más satisfactoria más adelante.

En palabras, r mide cuán agrupados en una linea recta están los puntos en el diagrama de dispersión, siendo su signo positivo cuando la asociación es positiva.



El diagrama de dispersión, que representa la aproximación a la distribución conjunta de (X,Y) queda resumido entonces por los cinco números:

$$\overline{X},\ s_X,\ \overline{Y},\ s_Y,\ r.$$

Estos son los estimadores de los cinco parámetros que definen la distribución normal bivariada.

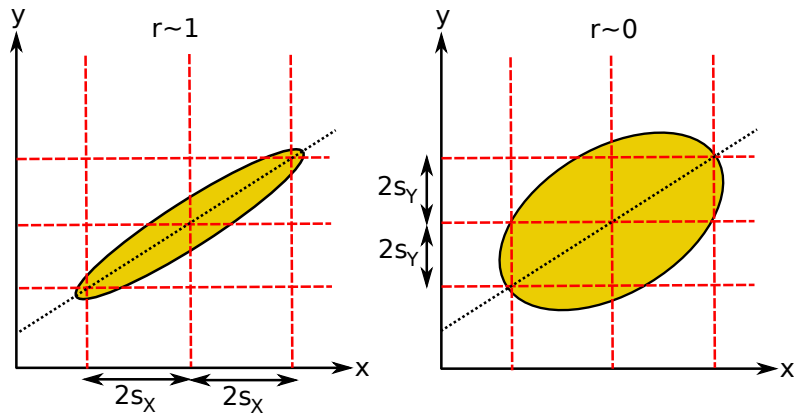
La recta de desvíos

La nube de puntos en un diagrama de dispersión tiende a inclinarse en la dirección de la *recta de desvíos*. La recta de desvíos es aquella que pasa por los promedios y tiene pendiente $\pm s_Y/s_X$. Esto se ve claramente en el diagrama de puntos de los datos de Pearson. Los desvíos de las dos variables son casi iguales en este caso, por lo que las variaciones verticales pueden ser comparadas con las horizontales, pero esto no será siempre así.

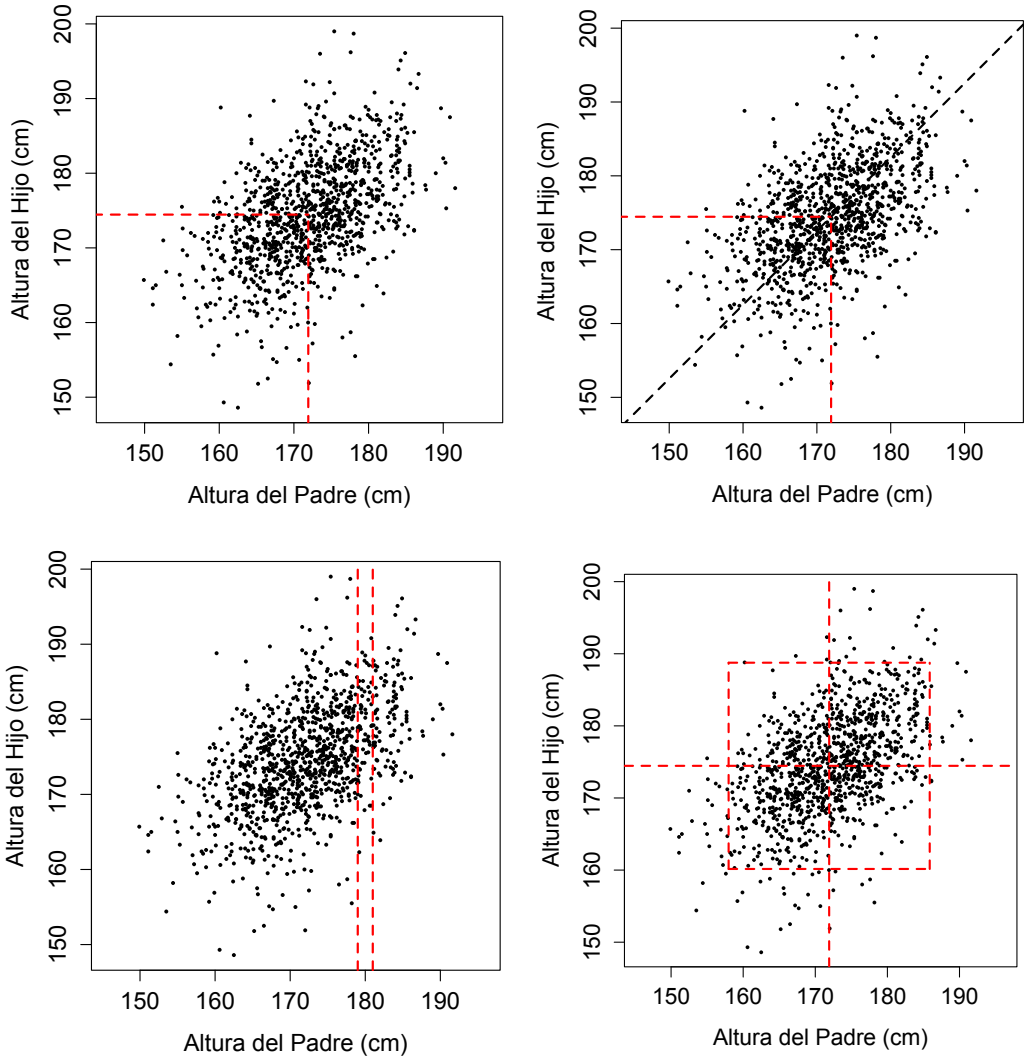
En general, la recta de desvíos mide la relación de variabilidad entre X e Y, y el signo de la pendiente depende del signo de la correlación r. Aunque no la usaremos mucho, la ecuación de la recta es

$$y - \overline{y} = \left(\text{signo}(r) \frac{s_Y}{s_X} \right) (x - \overline{x}).$$

Observar que fijados los desvíos, la correlación r mide si la nube es fina o gruesa. Se puede tener dos nubes de puntos con idénticas rectas de desvíos, pero con correlaciones muy diferentes.



La recta de desvíos mide simplemente la distorsión de escalas producida por la diferencia entre s_X y s_Y. Cuanto mayor sea s_Y, por ejemplo, más estirada hacia arriba será la nube de puntos.



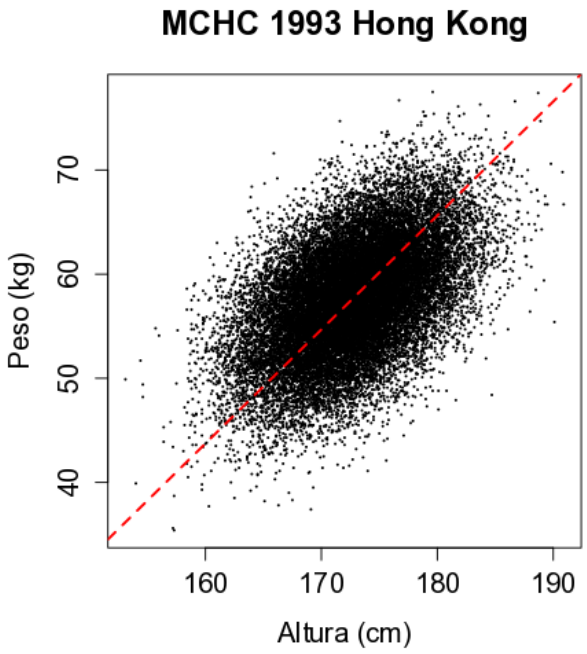
3. La recta de regresión empírica

Como vimos la clase pasada, el método de regresión consiste en estimar el valor promedio de Y correspondiente a un determinado valor de X. Esto da lugar a la definición de la función de regresión

$$R(x) = \mathbf{E}(Y|X = x) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X).$$

Lo que haremos ahora es estimar esta función a partir de la nube de puntos, e interpretar su significado empírico.

Tomemos como ejemplo el siguiente estudio realizado en 1993 en Hong Kong por el Maternal and Child Health Centres. Es un estudio de 25 000 adolescentes de 18 años, en el cual se midieron X la altura en cm, e Y el peso en kg. La figura siguiente muestra el diagrama de dispersión:

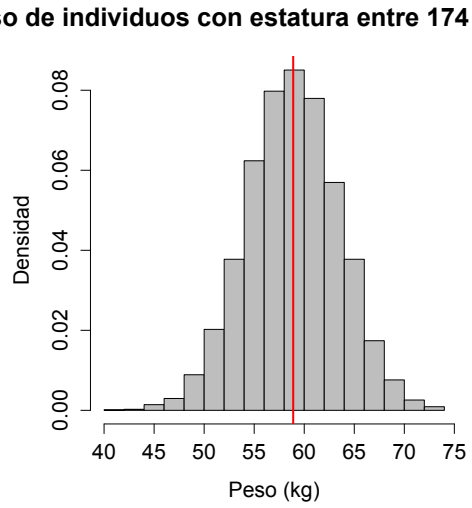


El resumen numérico del diagrama es

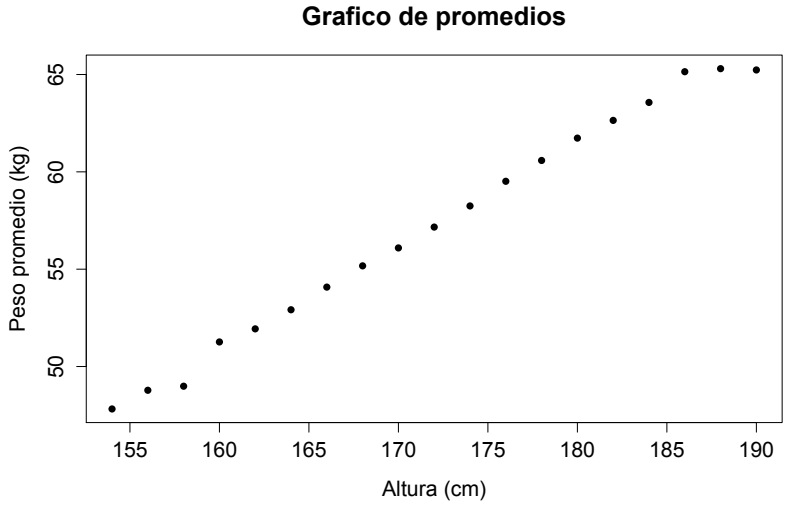
$\bar{x}$ (cm)	s_X (cm)	$\bar{y}$ (kg)	s_Y (kg)	r
172.7	4.8	57.6	5.3	0.5

Es claro que los altos pesan en general más que los bajos, pero queremos cuantificar el incremento en peso debido a un incremento en altura.

La dificultad es, al igual que en los datos de Pearson, la gran variabilidad de peso entre individuos que miden aproximadamente lo mismo. El siguiente gráfico muestra cómo se distribuye el peso entre los adolescentes que tienen una estatura entre 174 cm y 176 cm.



Pero al tomar promedios deberíamos ver puntos alineados, ya que para datos normales la función de regresión es lineal. Esto es exactamente lo que ocurre:



¿Cómo son los incrementos? El incremento en ΔX es siempre de 2 cm, y el de ΔY es

ΔY	0.96	0.21	2.27	0.67	0.98	1.16	1.09	0.92	1.07
	1.09	1.27	1.07	1.15	0.91	0.92	1.58	0.16	−0.07

El incremento promedio (eliminando los extremos que son un poco anómalos) es 1.06 kg. ¿Y si los medimos en unidades típicas?

$$\begin{cases} 1 \text{ unidad típica de } X = 1 \text{ desvío } s_X = 4.8 \text{ cm.} \\ 1 \text{ unidad típica de } Y = 1 \text{ desvío } s_Y = 5.3 \text{ kg.} \end{cases}$$

Los incrementos quedan:

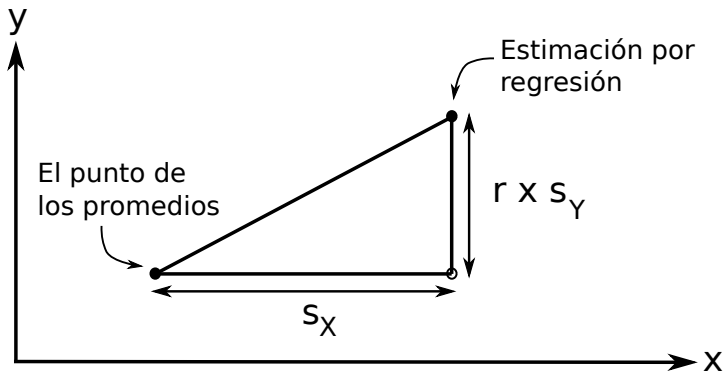
$$\Delta X = 0.42 \text{ u.t.} \quad \text{Promedio}\,(\Delta Y) = 0.2 \text{ u.t.} \quad \text{Promedio}\left(\frac{\Delta Y \text{ en u.t.}}{\Delta X \text{ en u.t.}}\right) = 0.48 \approx r = 0.5$$

Los promedio representados por los puntos en el gráfico anterior son estimaciones de la función $R(x) = \mathbf{E}\,(Y|X = x)$. Es decir, la recta de regresión teórica debe estar cerca de la

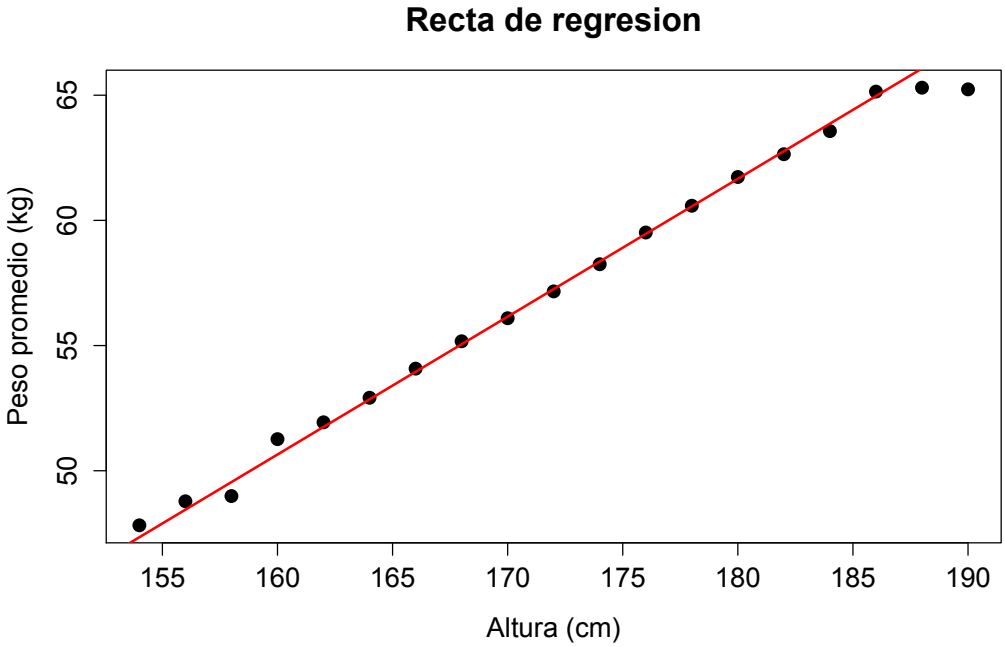
"recta" del gráfico de promedios. Es entonces natural estimar esta recta con la *recta de regresión empírica* que cumple:

- pasa por el punto de los promedios $\overline{p} = (\overline{x}, \overline{y})$;
- tiene Pendiente $= r\frac{s_Y}{s_X}$.

Por lo tanto, la ecuación de la recta es $y - \overline{y} = r\frac{s_Y}{s_X}(x - \overline{x})$, que es la ecuación análoga de la función de regresión teórica que vimos la clase pasad.

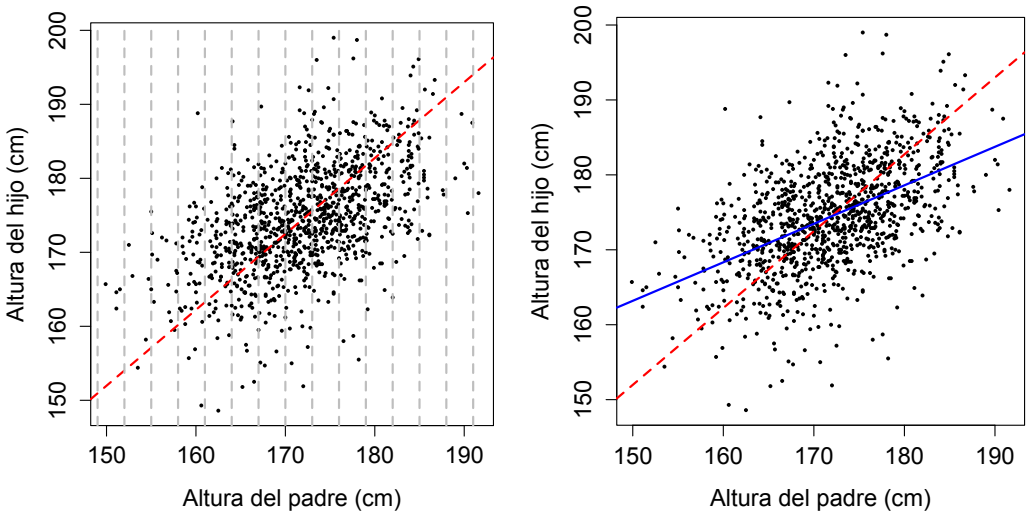


Notar qué bien se ajusta la recta de regresión empírica a la recta de los promedios en el estudio de los adolescentes de Hong Kong:



Regresión a la media

En la clase pasada vimos el fenómeno de regresión a la media. Usando la nube de puntos podemos entenderlo de forma visual. Volvamos a los datos de las alturas de Pearson. Notar lo que ocurre al dividir por franjas:



Vemos que el promedio (en cada franja) en la parte derecha de la nube está por debajo de la recta de desvíos, mientras que en la parte izquierda ocurre lo contrario. Lo mismo pasa con la recta de regresión porque el coeficiente r es menor o igual a 1.

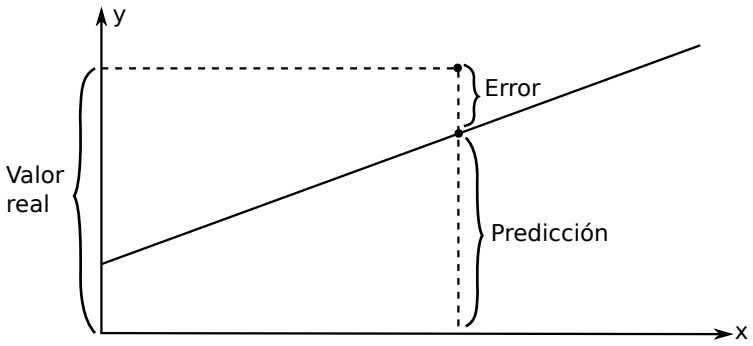
El Error Cuadrático Medio en regresión

El error de predicción para el i-ésimo individuo es $\varepsilon_i = y_i - \text{reg}(x_i)$ en donde $\text{reg}(x_i)$ es el valor predicho por la recta de regresión:

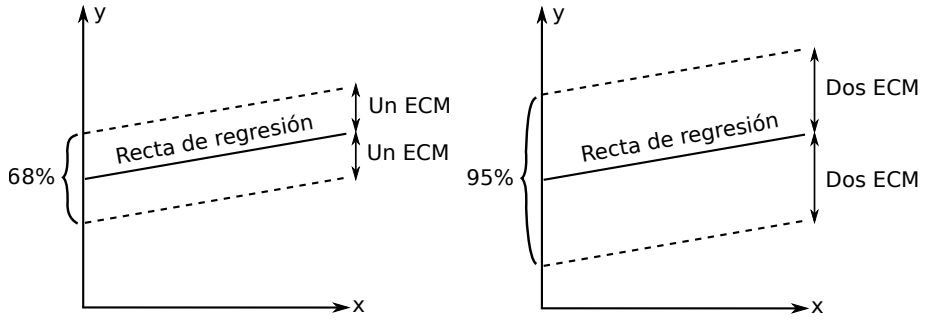
$$\text{reg}(x) = r\frac{s_Y}{s_X}(x-\bar{x})+\bar{y}.$$

El *error cuadrático medio* de predicción es por definición

$$\text{ECM} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\varepsilon_i^2}.$$



En palabras, el ECM dice cuán lejos en promedio están los pares $p_i = (x_i, y_i)$ de la recta de regresión:



Recordar que la distribución de Y dado que $X = x$ es normal de esperanza $R(x)$ y desvío

$$\sigma_Y \sqrt{1-\rho^2}.$$

Entonces, intuitivamente esperamos que el ECM sea un estimador de la varianza condicional de Y.

De hecho, el ECM se puede calcular mediante

$$\text{ECM} = \sqrt{1-r^2}\, s_Y.$$

La prueba es una cuenta relativamente sencilla:

$$\begin{aligned}\text{ECM}^2 &= \frac{1}{n-1}\sum_{i=1}^{n} \varepsilon_i^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y} - r\frac{s_Y}{s_X}(x_i - \overline{x})\right)^2 \\ &= s_Y^2 - 2r^2 s_Y + r^2 s_Y^2 \\ &= (1-r^2)s_Y^2.\end{aligned}$$

4. Variabilidad explicada por la regresión

Recordar que estamos siempre suponiendo que el par (X,Y) tiene distribución normal bivariada. Cuando existe una relación entre X e Y, parte de la variación de Y se explica por el hecho de que cuando X cambia, arrastra consigo a Y. ¿Cuánto influye esto en la variación total de Y?

Observar que la media de las predicciones de regresión $\text{reg}(x_i)$ coincide con la de Y:

$$\frac{1}{n}\sum_{i=1}^{n} \text{reg}(x_i) = \frac{1}{n}\sum_{i=1}^{n} r\frac{s_Y}{s_X}(x_i - \overline{x}) + \overline{y} = \overline{y}.$$

La *variación explicada* por la regresión es por definición

$$S_{\text{reg}}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (\text{reg}(x_i) - \overline{y})^2.$$

Podemos descomponer la variación total de Y como $s_Y^2 = \text{ECM}^2 + S_{\text{reg}}^2$. La prueba es una cuenta: primero observar que

$$\sum_{i=1}^{n} (y_i - \text{reg}(x_i))\,(\text{reg}(x_i) - \overline{y}) = 0$$

De aquí resulta

$$\begin{aligned} s_Y^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \mathrm{reg}(x_i) + \mathrm{reg}(x_i) - \overline{y}\right)^2 \\ &= \mathrm{ECM}^2 + \frac{2}{n-1}\sum_{i=1}^{n}\left(y_i - \mathrm{reg}(x_i)\right)\left(\mathrm{reg}(x_i) - \overline{y}\right) + s_{\mathrm{reg}}^2 = \mathrm{ECM}^2 + S_{\mathrm{reg}}^2. \end{aligned}$$

La fórmula anterior se interpreta así:

$$\overbrace{s_Y^2}^{\substack{\text{Variación}\\ \text{total de } Y}} = \overbrace{ECM^2}^{\substack{\text{Variación}\\ \text{residual}}} + \overbrace{S_{\mathrm{reg}}^2}^{\substack{\text{Variación}\\ \text{de regresión}}}$$

Pero recordando que $ECM^2 = s_Y^2(1-r^2)$, podemos re-escribir la relación anterior como

$$\overbrace{s_Y^2}^{\substack{\text{Variación}\\ \text{total de } Y}} = \overbrace{s_Y^2(1-r^2)}^{\substack{\text{Variación}\\ \text{residual}}} + \overbrace{s_Y^2 r^2}^{\substack{\text{Variación}\\ \text{de regresión}}}$$

Es por esto que podemos interpretar los cocientes

$$\frac{r^2}{1-r^2} = \frac{\text{Variación de regresión}}{\text{Variación residual}} \qquad r^2 = \frac{\text{Variación de regresión}}{\text{Variación total de } Y}$$

como una comparación de dos varianzas.

5. Regresión y mínimos cuadrados

Busquemos la recta que mejor aproxima los puntos $p_i = (x_i, y_i)$ en el diagrama de dispersión usando el método de mínimos cuadrados.

Entre todas las rectas $y = \beta x + \alpha$ queremos aquella que minimiza el

$$\mathrm{ECM}(\beta,\alpha)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \beta x_i - \alpha\right)^2.$$

Derivamos e igualamos a cero

$$\begin{aligned} \frac{\partial \mathrm{ECM}^2}{\partial \beta} &= \frac{2}{n-1}\sum_{i=1}^{n}\left(y_i - \beta x_i - \alpha\right)x_i = 0 \\ \frac{\partial \mathrm{ECM}^2}{\partial \alpha} &= \frac{2}{n-1}\sum_{i=1}^{n}\left(y_i - \beta x_i - \alpha\right) = 0 \end{aligned}$$

Haciendo cuentas, llegamos a

$$\begin{aligned} \beta &= \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2} \\ \alpha &= \overline{y} - \beta\overline{x}. \end{aligned}$$

Y haciendo un poco más de cuentas

$$\beta = r\frac{s_Y}{s_X}.$$

Es decir, que la recta de regresión es la recta que mejor aproxima los puntos del diagrama de dispersión por el método de mínimos cuadrados.