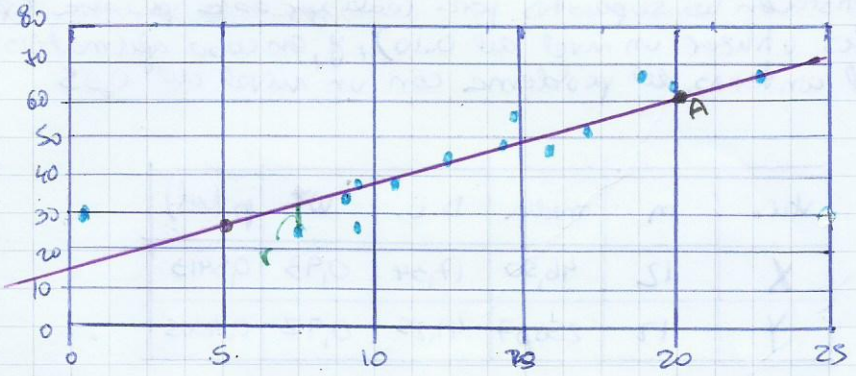
Cap 6

REGRESIÓN Y CORRELACIÓN

① Se realiza un estudio para establecer una ecuación mediante la cual se puede utilizar la concentración de cobre de una pieza (X) para predecir el porcentaje de níquel de la misma (Y). Se extrajeron al azar los sig. datos de 14 piezas fabricadas en una misma planta durante el mes pasado:

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

a) Graficar sobre el diagrama de dispersión la recta de regresión y señalar un par ordenado observado y un residuo



Coef regr.

	Est
coef const	15,85
X	2,26

del encuadrado (tabla)

$\hat{Y} = 2,26\,X + 15,85$ $\quad \hat{Y}_{(5)} = 27,15$ $\quad \hat{Y}_{(20)} = 61,05 \Rightarrow A = (20; 61,05)$

b) Verificar gráfica y analíticamente que la recta de regresión pasa por el punto $(\bar{x}; \bar{y})$ ✓

c) Indicar, a partir de la salida de computadora los parámetros estimados para el modelo y el % de variab. de Y que logra explicar X a través de la recta de regresión (lo hice en a) $\quad Y = a + bX$

Ahora lo hago analíticamente: → según calculadora: b_0: $a = 15,8525$ $\quad b_1$: $b = 2,6226$

$$\hat{Y} = 15,8525 + 2,6226\,x \qquad r = 0,914 \Rightarrow r^2 = 0,8356$$

$\hat{Y} = \beta_0 + \beta_1 x \rightarrow$

$\beta_0 = 15,8525$

$\beta_1 = 2,6226$

$R^2 = 83,56\%$

(2) En una planta química, se sospecha que la cant. de vapor utilizada por mes (Y en miles de libras) está relacionada con la temperatura ambiente promedio del mismo mes (X en °F). La sig. tabla presenta, para todo un año, el uso del vapor y la temperatura del mes correspondiente.

X	21	24	32	47	50	59	68	74	62	50	41	30
Y	240	220	320	325	280	230	200	275	270	250	200	270

Shapiro - Wilks (modificado)

a) Indicar las hipótesis a testear para responder a la sospecha de la planta química indicando el estadístico de contraste y su distribución

$n = 12$

$H_0: \rho = 0$ vs $H_1: \rho \neq 0$ $\quad T = \frac{\sqrt{10}\, r}{\sqrt{1-r^2}} \overset{a}{\sim} t_{10}$ bajo H_0

b) Indicar si se satisfacen los supuestos para realizar esta prueba (para el test de Shapiro Wilks utilizar un nivel del 0.10), y, en caso afirmativo, realizarla y concluir en el contexto del problema con un nivel del 0,05

var.	n	media	D.E.	W*	p (val.)
X	12	46,50	17,34	0,93	0,5413
Y	12	256,7	41,25	0,92	0.4005

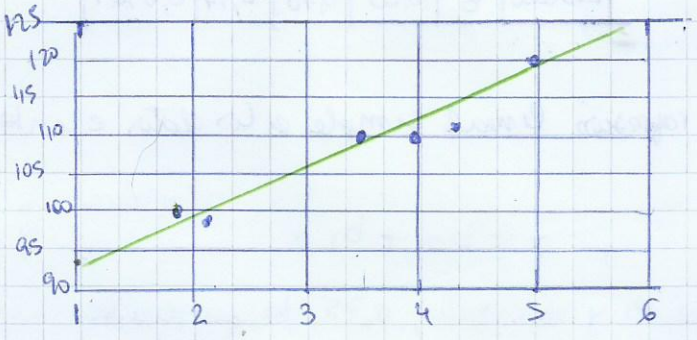
H_0: la muestra es normal vs H_1: la muestra no es normal

Rechazo H_0 si p valor $\leq \alpha$ $\quad \alpha = 0,10$

(3) Se quiere estudiar la asociación lineal entre consumo diario de sal y tensión arterial. A una serie de voluntarios se les interroga sobre dosis de sal en su dieta y se mide simultáneamente su tensión arterial. Los datos están en la tabla:

consumo sal (gr/dia)	1,8	2,2	3,5	4,0	4,3	5
tensión art. mmHg	100	98	110	110	112	120

$n = 6$



$\hat{Y} = \beta_0 + \beta_1 X$

$Y = b_0 + b_1 X = 86,37 + 6,33 X$

Si se sabe que ambas variables tienen distr. Normal conjunta,

a) de la observación del diagrama de dispersión ¿le parece razonable testear la hipótesis de asociación lineal? Justificar

Se observa que una recta podría definir la trayectoria no teniendo muchos residuos. Para mí, se ve que es lineal

b) ¿qué test correspondería aplicar en caso afirmativo?

Si fuese alternativo, sería: $H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$

$$T = \frac{2r}{\sqrt{1-r^2}} \sim t_2 \text{ bajo } H_0$$

c) Establecer las hipótesis, indicar estadístico de contraste, la región crítica y la decisión al 5%

$\alpha = 0,05 \rightarrow \frac{\alpha}{2} = 0,025 \rightarrow t_{4,0.025} = 2,776$

Rechazo H_0 si $|T_{obs}| > t_{4,0.025}$

según calculadora: $r = 0,966 \rightarrow t_{obs} = \frac{2 \times 0.9666}{\sqrt{1-0,9666^2}} = 7,5462 > 2,776$ ($t_{4,0.025}$)

$\rightarrow$ Rechazo H_0

la regresión es estadísticamente significativa

④ Los sig. datos proporcionan información acerca del contenido de agua de nieve (X) en cm y la afluencia de abril a junio en pulgadas (Y) en la cuenca del río Salado. Se sabe que ambas variables tienen distribución Normal.

Shapiro Wilks

X	23,1	32,8	31,8	32	30,4	24	39,5	24,2
Y	10,5	16,7	18,2	17	16,3	10,5	23,1	12,4

Variable	n	media	d.e.	W*	p-valor
Residuo	8	0,00	0,78	0,94	0,6941

a) Ajustar el modelo de regresión lineal simple a los datos e interpretar los coeficientes obtenidos

$\hat{Y} = \beta_0 + \beta_1 X$, $y = b_0 + b_1 x \xrightarrow{\text{calculadora}} \hat{y} = -6,68 + 0,75\,x$

b_0 es el valor de y con $x=0$, 0,75 la pendiente

b) ¿Se satisface el supuesto de normalidad de los errores para este modelo? utilizar un 10% de significación

H_0: es normal vs H_1: no es normal

Rechazo H_0 si p valor < nivel de significación | 0,6941 0,10 | p valor > α | No rechazo → satisface supuesto de normalidad

d) testear la significación del modelo con nivel 0,05

$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

$\alpha = 0.05 \rightarrow \frac{\alpha}{2} = 0,025$

$n = 8$

$T = \frac{b_1}{S/\sqrt{S_{xx}}} \sim t_{n-2}$ bajo H_0 | Rechazo H_0 si $|T_{obs}| > t_{n-2,\frac{\alpha}{2}}$

$b_1 = 0,75$

$S^2 = \left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right)\frac{1}{n-2} = \left(128,93 - \frac{166,41^2}{222,14}\right)\cdot\frac{1}{6} = 0,7109 \rightarrow S = 0,843$

$|T_{obs}| = \frac{0,75}{0,843/\sqrt{222,14}} = 13,26$

$t_{6,0.025} = 2,447$

$|T_{obs}| > t_{6,0,025} \rightarrow$ Rechazo H_0

→ la regresión es estadísticamente significativa

⑤ Una compañía productora de energía eléctrica está interesada en desarrollar un modelo que relacione la demanda máxima por hora (D, en Kw) con el uso de la energía total al mes (U, en Kwh) las variables se distribuyen en forma Normal bivariada

La sig. tabla muestra los datos obtenidos de una muestra de 11 clientes elegidos al azar entre los de esta compañía:

X	D	679	292	1012	493	582	1156	997	2189	1087	2077	2078
Y	U	0,79	0,44	0,56	0,79	2,7	3,64	4,73	9,5	5,34	6,88	6,85

a) Ajustar al modelo de regresión lineal simple a los datos e indicar si el ajuste es adecuado con un nivel de significación del 1%

$Y = \beta_0 + \beta_1 X + \varepsilon$ Recta de regresión $\hat{y} = b_0 + b_1 x$

$\hat{Y} = -0,88 + 0,0041 X$ $b_0 = -0,88$ $b_1 = 0,0041$

$n = 11$

$\alpha = 0,01$ $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

$T_{obs} = \frac{b_1}{S/\sqrt{S_{xx}}} \sim t_9$ bajo H_0 . Rechazo H_0 si $|T_{obs}| > t_{9,0.005}$

$S^2 = \left(S_{YY} - \frac{S_{XY}^2}{S_{XX}}\right) \cdot \frac{1}{n-2} = \left(95,64 - \frac{18717,98^2}{4558764,18}\right) \frac{1}{9} = 2,01 \rightarrow \boxed{S = 1,42}$

$T_{obs} = \frac{0,0041}{1,42} \cdot \sqrt{4558764,18} = 6,1652$

$t_{9,0.005} = 3,250 \longrightarrow T_{obs} > t_{9,0.005} \rightarrow$ Rechazo H_0

$\rightarrow$ el ajuste es significativo

b) Encontrar un intervalo de predicción para $x_0 = 900$ Kw con un nivel del 99%. Interpretarlo e indicar la diferencia con un intervalo de confianza para la media con el mismo valor dado

$\hat{Y}_0 = -0,88 + 0,0041 \cdot 900 = 2,81 = \hat{Y}_0$, $\bar{X} = 1149,27$, $\alpha = 0,99$

$\boxed{IP_r = [1.31, 4.30]}$ $\boxed{IC = [-2.04, 7.66]}$

c) ¿Puede estimarse a partir de estos datos el valor esperado de Y para $X = 156$?

No, porque $x = 156$ está lejos del recorrido de X

⑥ Se realizó un estudio para estimar los efectos producidos en las personas debido a la exposición al ruido. Ocho personas participaron en este estudio. Se registraron los aumentos en la presión sanguínea del individuo y el nivel de presión sonora, distribuido normalmente, a que se sometieron. Los datos se expresan en la tabla,

X: Nivel de Presión Sonora	60	63	70	80	85	90	94	100
Y: Aumento pr. Sang.	1	0	4	3	5	8	7	6

$n = 8$

$\sum_{i=1}^{n} x_i = 642$ $\quad \sum_{i=1}^{n} y_i = 34$ $\quad \sum_{i=1}^{n} x_i^2 = 53030$ $\quad \sum_{i=1}^{n} y_i^2 = 200$ $\quad \sum_{i=1}^{n} x_i y_i = 2983$

a) Hallar la recta de regresión estimada

$Y = \beta_0 + \beta_1 X + \varepsilon$ $\qquad \hat{Y} = b_0 + b_1 x \xrightarrow{\text{calculadora}}$ $b_0 = -9{,}28$, $b_1 = 0{,}169$

$$\boxed{\hat{Y} = -9{,}28 + 0{,}17x}$$

b) Hallar el coeficiente de determinación e interpretarlo

$$r^2 = \frac{S^2_{xy}}{S_{xx} S_{yy}} = \frac{254{,}5^2}{1509{,}5 \times 55{,}5}$$

$$\boxed{r^2 = 0{,}77}$$

El 77 % de la variabilidad de Y queda explicado por la variabilidad de X a través de la recta de regresión

(7) Una empresa de servicios informáticos está preocupado por la proliferación de cierto tipo de virus. La sig. tabla registra el número de días que han transcurrido desde que se ha detectado un nuevo virus (X) y el número de ordenadores (Y) que han requerido los servicios de esta empresa.

x_i	y_i	$\ln(y_i)$
1	254	5,54
2	1,5	0,41
4	2,105	0,74
5	5,05	1,62
8	16,2	2,78
10	45,32	3,81
11	58,57	4,07
14	375,8	5,93
16	1525d40	14,24
20	2577000	14,76

$m = 10$

a) A partir de los diagramas de dispersión presentados (ver guía) decidir cuál es la regresión lineal más adecuada y estimar los coeficientes correspondientes

$\hat{Y} = -579816 + 108808X$

$\widehat{\ln(Y)} = -0,714 + 0,67X$

$\hat{Y} = e^{(-0,714 + 0,67X)}$

la más conveniente

b) Estimar puntualmente el número de consultas que tendrá la empresa a los 15 días de la detección del virus

$\hat{Y}_{(15)} = 11.486$

con $\ln(Y)$

(9) [T] Demostrar que la suma de cuadrados totales es igual a la adición de la suma de sus cuadrados residuales y la suma de cuadrados de regresión; es decir:

$$\sum_{i=1}^{m}(Y_i-\bar{Y})^2=\sum_{i=1}^{m}(Y_i-\hat{Y}_i)^2+\sum_{i=1}^{m}(\hat{Y}_i-\bar{Y})^2$$

$$\sum_{i=1}^{m}(Y_i-\bar{Y})^2=\sum_{i=1}^{m}\left[(Y_i-\hat{Y}_i+\hat{Y}_i-\bar{Y})\right]^2=\sum_{i=1}^{m}\left[(Y_i-\hat{Y}_i)+(\hat{Y}_i-\bar{Y})\right]^2=$$

($\hat{Y}_i-\hat{Y}_i=0$)

$$=\sum_{i=1}^{m}\left[(Y_i-\hat{Y}_i)^2+2(Y_i-\hat{Y}_i)(\hat{Y}_i-\bar{Y})+(\hat{Y}_i-\bar{Y})^2\right]=$$

$$=\sum_{i=1}^{m}(Y_i-\hat{Y}_i)^2+2\underbrace{\sum_{i=1}^{m}(Y_i-\hat{Y}_i)(\hat{Y}_i-\bar{Y})}_{0}+\sum_{i=1}^{m}(\hat{Y}_i-\bar{Y})^2=$$

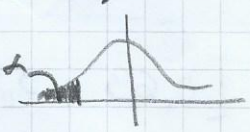
$$=\sum_{i=1}^{m}(Y_i-\hat{Y}_i)^2+\sum_{i=1}^{m}(\hat{Y}_i-\bar{Y})^2$$

- $\sum_{i=1}^{m}\overbrace{(Y_i-\hat{Y}_i)}^{\text{el error}}(\hat{Y}_i-\bar{Y})=\sum_{i=1}^{m}e_i(\hat{Y}_i-\bar{Y})$ → considerando que $\bar{Y}=b_0+b_1X_i$ y la 2ª ecuación normal queda demostrado que da cero

---

(10) Con respecto a los resultados de un análisis de regresión lineal simple ¿cuál de las afirmaciones puede ser cierta? Justificar.

- $SC_{\text{TOTAL}}\ (1) < SC_{\text{ERROR}}\ (2) + SC_{\text{REGRES}}\ (3)$ F $\quad S_{YY}\ (1) < \overbrace{S_{YY}-\frac{S_{XY}^2}{S_{XX}}}^{(2)} + \overbrace{\frac{S_{XY}^2}{S_{XX}}}^{3}$ ← absurdo
- $R^2=-0,65$ F, $R^2>0\ \forall R\in\mathbb{R}$
- $S_b=-1,25$ F $\quad S_b>0$
- $t_{obs}=-2,45$ V



$t_{obs}=-2,45 \rightarrow \alpha=0,01$

Supongo $n=30$

(11) En varios estudios se ha demostrado que los líquenes, organismos compuestos por un alga y un hongo, son buenos indicadores biológicos de la contaminación del aire. Los sig. datos corresponden a las variables X (depósito de nitratos en g/cm²) e Y (nitrógeno en el líquen, en % de peso seco

X	0,05	0,1	0,11	0,12	0,31	0,37	0,42	0,58	0,68	0,68	0,73	0,85	0,82
Y	0,48	0,55	0,48	0,5	0,58	0,52	1,02	0,86	0,86	1	0,88	1,04	1,7

a) Indicar el modelo estimado para estos datos por mínimos cuadrados

x enunciado const: 0,37 , X = 0,97

$$\hat{Y} = 0,37 + 0,97X$$

b) Completar los p valores de los solidos e indicar la significación del modelo

Datos: $T \to 3,69$ / $5,29$ $\to t_{obs_1} = 3,69$

$m = 13$

$\to t_{11,\frac{\alpha}{2}} = 3,69 \to \frac{\alpha}{2} = 0,0018$

$\alpha_1 = 0,0036$

$t_{obs_2} = 5,29 \to t_{11,\frac{\alpha}{2}} = 5,29 \to \frac{\alpha}{2} = 0,00015 \to \alpha_2 = 0,0003$

p valor $_1 = 0,0036$ p valor $_2 = 0,0003$

es significativa

c) Construir un IC. de nivel 95% para el valor esperado de nitrógeno en el líquen para un depósito de nitrato de 0,4 g/cm²

$X_0 = 0,4$

$\bar{X} = 0,455$

$\hat{Y}_{0(0,4)} = 0,758$

$t_{11,0,025} = 2,201$

$\alpha = 0,05$

$$IC = \left[\hat{y}_0 - t_{11,0025}\, S\sqrt{\frac{1}{m} + \frac{(x_0 - \bar{x})^2}{Sxx}} \; ; \; \hat{y}_0 + t_{11,0025}\, S\sqrt{\frac{1}{m} + \frac{(x_0 - \bar{x})^2}{Sxx}}\right]$$

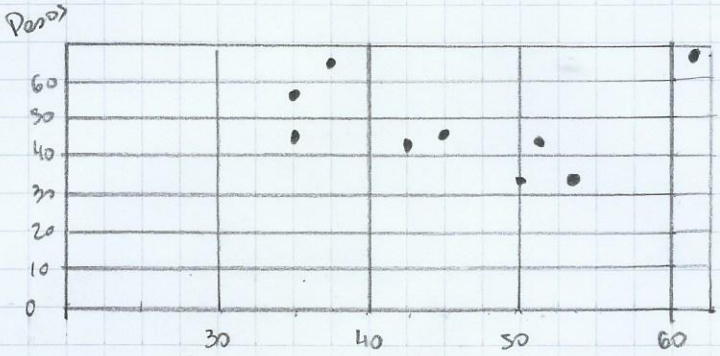
$$\Rightarrow \quad IC = [0,63 \; ; \; 0,88]$$

$$S^2 = \left(Syy - \frac{S_{xy}^2}{Sxx}\right)\frac{1}{n-2} = \left(1,4533 - \frac{1,0785^2}{1,1155}\right)\frac{1}{11} = 0,037 \to S = 0,193$$

(12) En la sig. tabla se han registrado las sup. y los pesos de ciertas piezas metálicos producidos por una máquina

X	Superficie	44,09	36,67	51,72	36,04	38,97	61,4	42,06	53,33	50,01
Y	Peso	47,6	45	43,04	56,88	63,11	68,7	41,8	34,3	31,2
		1	2	3	4	5	6°	7	8	9

a) Graficar el diagrama de dispersión aproximado de Peso vs Superficie ¿Hay alguna observación que le llame la atención?



Parecen puntos muy dispersos, comparando a la mayoría con el de 61, que queda muy alejado → la 6° observación

b) Estimar los coeficientes del modelo de regresión lineal utilizando toda la información disponible

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$\hat{Y} = b_0 + b_1 x \rightarrow \boxed{\hat{Y} = 49,23 - 0,027 X}$$ ✓

c) Calcular los residuos ¿hay alguno que le parezca especial? $r_i = Y_i - (b_0 + b_1 X_i)$

$r_1 = -0,44$ $r_3 = -4,79$ $r_5 = 14,93$ $r_7 = -6,29$ $r_9 = -16,08$

$r_2 = -3,24$ $r_4 = 8,54$ $\boxed{r_6 = 21,13}$ $r_8 = -13,49$

d) Eliminar la 6° observación y repetir los dos ítems anteriores.

$$\hat{Y} = 95,58 - 1,138 X$$

$r_1 = 2,19$ $r_3 = 6,32$ $r_5 = 11,86$ $r_8 = -0,56$

$r_2 = -8,85$ $r_4 = 2,23$ $r_7 = -5,95$ $r_9 = -7,37$

tiene menor rango de diferencias