Equipo 6
Álvaro Acedo 174052
Diana Santiago 175325
Jimena Reyes 173361
Paulina Mazariegos 171929
Pamela Goya 171789

ESTADÍSTICA APLICADA II
Tarea No. 2

Dr. Víctor M. Guerrero
Ago-Dic, 2021

1. Un investigador se interesó en estudiar las siguientes series de datos para una región del Reino Unido:

Año	2005	'06	'07	'08	'09	'10	'11	'12	'13	'14	'15	'16
X	60	62	61	55	53	60	63	53	52	48	49	43
Y	23	23	25	25	26	26	29	30	30	32	33	31

Donde

X = Miles de muertes de niños menores de un año y

Y = Barriles de cerveza consumida.

(a) **Calcule** el coeficiente de correlación muestral entre X y Y.

(b) Una tendencia lineal en el tiempo se ajusta a X al calcular la regresión de X sobre t. Por ejemplo, si el origen del tiempo se sitúa a la mitad de 2005 y la unidad de tiempo usada es el año, entonces el año 2012 corresponde a t = 7.

Si, en cambio, el origen se localiza al final del año 2010 (o al inicio de 2011) y la unidad de tiempo empleada es el semestre, entonces 2007 corresponde a t = -7.

Demuestre que cualquier valor *estimado por tendencia* $\hat{X}_t = b_0 + b_1 t$, no se altera por la selección del origen, ni por la unidad de medida del tiempo.

(c) Sean $\tilde{X}$ y $\tilde{Y}$ los valores de X y Y que resultan después de eliminar una tendencia lineal; o sea, $\tilde{X}_t = X_t - \hat{X}_t$ y $\tilde{Y}_t = Y_t - \hat{Y}_t$.

Calcule entonces (i) la correlación entre $\tilde{X}$ y Y, y (ii) la correlación entre $\tilde{X}$ y $\tilde{Y}$.

Compare estos valores con los de las correlaciones obtenidas en la parte (a) y **comente** acerca de las diferencias que encuentre, en particular **explique** lo que mide cada una de las correlaciones calculadas.

2. Realice la **estimación de una recta** de regresión para cada uno de los siguientes cuatro conjuntos de datos.

   Calcule también los coeficientes de correlación respectivos.

   Realice las **gráficas** que considere pertinentes.

   ¿Qué se puede **concluir** de este ejercicio?

i	X_1	Y_1	X_2	Y_2	X_3	Y_3	X_4	Y_4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.10	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.10	4	5.39	19	12.50
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89

Fuente: Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician 27*, pp. 17 – 21.

1. Un investigador se interesó en estudiar las siguientes series de datos para una región del Reino Unido:

Año	2005	'06	'07	'08	'09	'10	'11	'12	'13	'14	'15	'16
X	60	62	61	55	53	60	63	53	52	48	49	43
Y	23	23	25	25	26	26	29	30	30	32	33	31

Donde

X = Miles de muertes de niños menores de un año

Y = Barriles de cerveza consumida

a) Calcular el coeficiente de correlación muestral entre X y Y

cálculos aux:

$\sum X_t = 659$ $\qquad \sum X_t^2 = 36{,}635$

$\sum Y_t = 333$ $\qquad \sum Y_t^2 = 9{,}375$

$\sum X_t Y_t = 18{,}107$

Fórmulas:

- $r_{XY} = \dfrac{S_{XY}}{S_X S_Y} = \dfrac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$.
- $S_{XY} = \dfrac{\sum X_i Y_i - \sum X_i \sum Y_i / n}{n-1}$ = Covarianza muestral de X y Y.
- $S_X^2 = \dfrac{\sum X_i^2 - (\sum X_i)^2/n}{n-1}$ = Varianza muestral de X.

sea $n = 12$, entonces

$S_{xy} = (12-1)^{-1}\left[(18{,}107) - (659)(12^{-1})(333)\right] = -16.38636$

$S_x^2 = (12-1)^{-1}\left[(36{,}635) - (12^{-1})(659)^2\right] = 40.44697$

$S_y^2 = (12-1)^{-1}\left[(9{,}375) - (12^{-1})(333)^2\right] = 12.204545$

$S_x = (S_x^2)^{1/2} = 6.3597932$

$S_y = (S_y^2)^{1/2} = 3.4935005$

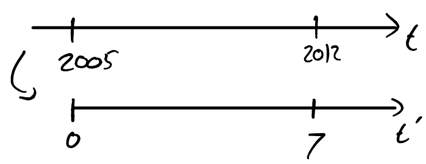
$\therefore \ r_{xy} = \dfrac{S_{xy}}{S_x S_y} = \dfrac{-16.38636}{(6.3597932)(3.4935005)}$

$= -0.737528$ Bien

(b) Una tendencia lineal en el tiempo se ajusta a X al calcular la regresión de X sobre t. Por ejemplo, si el origen del tiempo se sitúa a la mitad de 2005 y la unidad de tiempo usada es el año, entonces el año 2012 corresponde a t = 7.

Si, en cambio, el origen se localiza al final del año 2010 (o al inicio de 2011) y la unidad de tiempo empleada es el semestre, entonces 2007 corresponde a t = -7.

Demuestre que cualquier valor *estimado por tendencia* $\hat{X}_t = b_0 + b_1 t$, no se altera por la selección del origen, ni por la unidad de medida del tiempo.



Si definimos

t_0: el nuevo origen (centro)

K: una constante que nos da el número de las nuevas unidades de tiempo en un año

Podemos ver el nuevo tiempo t_i como:

$t' := K(t - t_0)$ Bien $\qquad t_0, K \in \mathbb{R}$,

Sean $\hat{X}_i = b_0 + b_1 t_i$ la recta de regresión con t original

$\hat{X}_i' = b_0' + b_1' t_i'$ la recta de regresión con t'

$\underline{Pd}$ $\hat{X}_i = \hat{X}_i'$

Tenemos que $b_1 = \dfrac{\sum (t_i - \bar{t})(X_i - \bar{X})}{\sum (t_i - \bar{t})^2}$

Luego

$$b_1' = \frac{\sum (t_i' - \bar{t})(X_i - \bar{X})}{\sum (t_i' - \bar{t})^2} = \frac{\sum (K(t_i - t_0) - K(\bar{t} - t_0))(X_i - \bar{X})}{\sum (K(t_i - t_0) - K(\bar{t} - t_0))^2}$$

$$= \frac{\sum (Kt_i - \cancel{Kt_0} - K\bar{t} + \cancel{Kt_0})(X_i - \bar{X})}{\sum (Kt_i - \cancel{Kt_0} - K\bar{t} + \cancel{Kt_0})^2} = \frac{\sum K(t_i - \bar{t})(X_i - \bar{X})}{\sum (K(t_i - \bar{t}))^2}$$

$$= \frac{\cancel{K} \sum (t_i - \bar{t})(X_i - \bar{X})}{K^2 \sum (t_i - \bar{t})} = \frac{1}{K} b_1 \qquad (K \neq 0)$$

$$\Rightarrow b_1' = \frac{1}{K} b_1 \qquad (*)$$

Por otro lado

$$b_0 = \hat{X}_i - b_1 t_i \qquad (**)$$

$$b_0' = \hat{X}_i - b_1' t_i' \underset{(*)}{=} \hat{X}_i - \frac{1}{K} b_1 t_i' = \hat{X}_i - \frac{1}{\cancel{K}} b_1 (\cancel{K}(t_i - t_0))$$

$$= \hat{X}_i - b_1 (t_i - t_0)$$

$$= \underbrace{\hat{X}_i - b_1 t_i}_{b_0 \ (**)} + b_1 t_0$$

$$= b_0 + b_1 t_0$$

$$\Rightarrow b_0' = b_0 + b_1 t_0$$

Finalmente,

$$\hat{X}_i' = b_0' + b_1' t_i' = b_0 + b_1 t_0 + \frac{1}{K} b_1 t_i'$$

$$= b_0 + b_1 t_0 + \frac{1}{\cancel{K}} b_1 (\cancel{K}(t_i - t_0))$$

$$= b_0 + \cancel{b_1 t_0} + b_1 t_i - \cancel{b_1 t_0}$$

$$= b_0 + b_1 t_i = \hat{X}_i$$

Bien

$\therefore$ El valor estimado por tendencia no se altera por selección del origen //

c) Sean $\tilde{X}$ y $\tilde{Y}$ los valores de X y Y que resultan después de eliminar una tendencia lineal; o sea, $\tilde{X}_t = X_t - \hat{X}_t$ y $\tilde{Y}_t = Y_t - \hat{Y}_t$.

Calcule entonces (i) la correlación entre $\tilde{X}$ y Y, y (ii) la correlación entre $\tilde{X}$ y $\tilde{Y}$.

Compare estos valores con los de las correlaciones obtenidas en la parte (a) y **comente** acerca de las diferencias que encuentre, en particular **explique** lo que mide cada una de las correlaciones calculadas.

i. Corr entre $\tilde{X}$ y Y

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})^2}}.$$

No entiendo para qué hicieron esto y veo que los cálculos que siguen no tienen much

Sea $\mathbb{E}(\tilde{X}) = 0$, la fórmula se simplifica

$$r_{\tilde{X}Y} = \frac{\sum (\tilde{X}_i)(Y_i - \overline{Y})}{\sqrt{(\sum \tilde{X}_i^{\,2})(\sum (Y_i - \overline{Y})^2)}}$$

Sea $\sum (\tilde{X}_i)(Y_i - \overline{Y}) = 0 \Rightarrow r_{\tilde{X}Y} = 0$ (calculado en excel)

En este caso, la covarianza muestral es 0 por lo que la interpretación es que no existe una relación lineal entre las variables. Consecuentemente, la correlación es 0.
Esta correlación mide la dirección y el grado de asociación lineal entre las observaciones de Y, y las observaciones de X disminuidas por su estimación correspondiente usando el método de regresión lineal simple. Es decir, mide la asociación entre las observaciones de Y y el error en la estimaciones de X. Al ser variables que no tienen una relación directa, suena lógico que la correlación sea 0

Esta conclusión es errónea.

ii. Corr. entre $\tilde{X}$ y $\tilde{Y}$

Sea también $\mathbb{E}(\tilde{Y})=0$, la fórmula se simplifica

$$r_{\tilde{X}\tilde{Y}} = \frac{\sum(\tilde{X}_i)(\tilde{Y}_i)}{\sqrt{(\sum \tilde{X}_i^2)(\sum \tilde{Y}_i^2)}} = 0.7375 = -r_{XY}$$ (calculado en excel)

De nuevo, este resultado es incorrecto. -8

Esta correlación mide la dirección y el grado de asociación lineal entre las observaciones de Y disminuidas por la estimación correspondiente, y las observaciones de X disminuidas por la estimación correspondiente usando el método de regresión lineal simple. Es decir, mide la asociación entre el error en las estimaciones de X y Y.

Notar que la correlación entre las observaciones de X y Y es negativa, por lo que existe una relación lineal inversamente proporcional entre dichas variables.

La correlación entre los errores de estimación de X y Y tiene exactamente la misma magnitud, pero es positiva, lo que indica que el grado de asociación es el mismo pero con dirección opuesta: en este caso las relación es directamente proporcional.

2. Realice la **estimación de una recta** de regresión para cada uno de los siguientes cuatro conjuntos de datos.

   Calcule también los coeficientes de correlación respectivos.

   Realice las **gráficas** que considere pertinentes.

   ¿Qué se puede **concluir** de este ejercicio?

i	X_1	Y_1	X_2	Y_2	X_3	Y_3	X_4	Y_4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.10	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.10	4	5.39	19	12.50
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89

Fuente: Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician 27*, pp. 17 – 21.

Cálculos auxiliares:

$\sum X_1 = 99$	$\sum X_2 = 99$	$\sum X_3 = 99$	$\sum X_4 = 99$
$\sum Y_1 = 82.51$	$\sum Y_2 = 82.51$	$\sum Y_3 = 82.5$	$\sum Y_4 = 82.51$
$\sum X_1 Y_1 = 797.6$	$\sum X_2 Y_2 = 797.59$	$\sum X_3 Y_3 = 797.47$	$\sum X_4 Y_4 = 797.58$
$\sum X_1^2 = 1001$	$\sum X_2^2 = 1001$	$\sum X_3^2 = 1001$	$\sum X_4^2 = 1001$
$\sum Y_1^2 = 660.1727$	$\sum Y_2^2 = 660.1763$	$\sum Y_3^2 = 659.9762$	$\sum Y_4^2 = 660.1325$
$\overline{X}_1 = 9$	$\overline{X}_2 = 9$	$\overline{X}_3 = 9$	$\overline{X}_4 = 9$
$\overline{Y}_1 = 7.5009$	$\overline{Y}_2 = 7.5009$	$\overline{Y}_3 = 7.5$	$\overline{Y}_4 = 7.5009$
$S_{X_1^2} = 11$	$S_{X_2^2} = 11$	$S_{X_3^2} = 11$	$S_{X_4^2} = 11$
$S_{Y_1^2} = 4.1272$	$S_{Y_2^2} = 4.1276$	$S_{Y_3^2} = 4.1262$	$S_{Y_4^2} = 4.1234$
$S_{X_1Y_1} = 5.501$	$S_{X_2Y_2} = 5.5$	$S_{X_3Y_3} = 5.497$	$S_{X_4Y_4} = 5.499$

Recordemos las fórmulas:

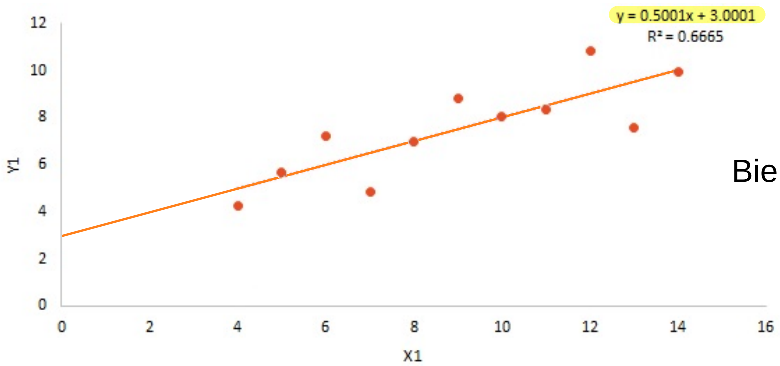
- $$b_1 = \frac{S_{XY}}{S_X^2}$$
- $$b_0 = \overline{Y} - b_1\overline{X}$$
- $$S_{XY} = \frac{\sum X_i Y_i - \sum X_i \sum Y_i / n}{n - 1}$$
- $$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})^2}}$$

$\therefore$ Tenemos los siguientes resultados:

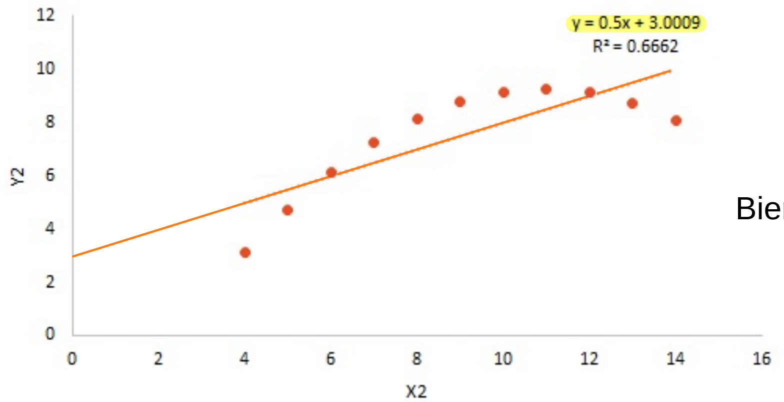
	$X_1 Y_1$	$X_2 Y_2$	$X_3 Y_3$	$X_4 Y_4$
b_1	0.5001	0.5	0.4497	0.4999
b_0	3.0001	3.0009	3.0025	3.0017
r_{XY}	0.8164	0.8162	0.8163	0.8165
R^2	0.6665	0.6662	0.6663	0.6667

} coeficientes de correlación

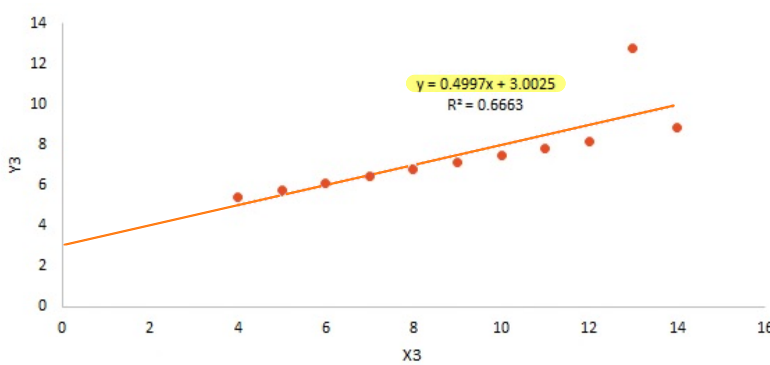
Gráficas y estimación de las rectas:



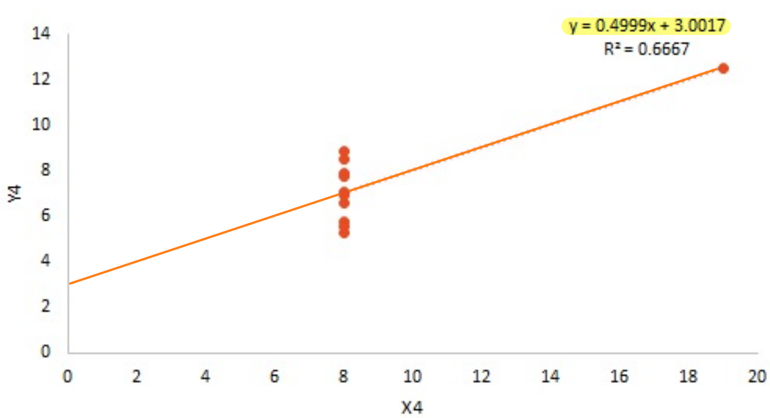
Bien



Bien



Bien



Bien

Conclusion Esto es lo importante

- A pes nida para los cuatro conjuntos de datos es prácticamente la misma, el comportamiento de las muestras es muy diferente.
- La estimación de la recta para el conjunto 1 ajusta apropiadamente a sus datos muestrales; sin embargo, no ocurre lo mismo con el conjunto 2, pues la muestra se comporta de manera cuadrática
- La estimación de la recta para el conjunto 3 tampoco describe correctamente a sus datos muestrales, pues existe un outlier en la muestra.
- Con respecto al conjunto 4, el valor X=8 contribuye más que el de X=19 al cálculo del coeficiente de correlación, el cual mide la asosiación lineal.
- El comportamiento de cada conjunto de muestras es muy diferente y, a pesar de ello, su asosiación lineal solo difiere en milésimas. Por ello es que se produce casi la misma recta, aunque no sea la más adecuada para todos estos conjuntos.