

Tarea #2

EQUIPO #3

Sofia Alejandra Diaz Miranda 172360

David Isaac Lopez Romero 173993

Sofia Oliva Ruiz 164595

Adriana Alvarez Lujano 163480

Diego Carlos Krafft de Silva 173246

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.

$$\bar{x} = \frac{60 + 62 + 61 + 55 + 53 + 60 + 63 + 53 + 52 + 48 + 49 + 43}{12}$$
$$= 54.9166$$
$$\approx 54.92$$

$$\bar{y} = \frac{23 + 23 + 25 + 25 + 26 + 26 + 29 + 30 + 30 + 32 + 33 + 31}{12}$$
$$= 27.75$$

$$\Rightarrow r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$
$$= -0.7375$$

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.

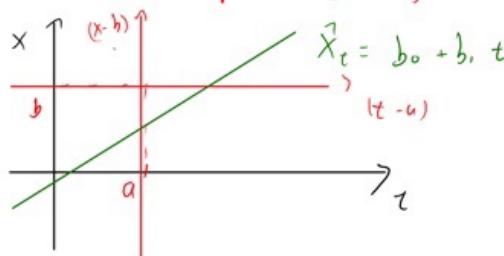
Podemos ver la tendencia como modelo de regresión lineal

$$\hat{X}_t = b_0 + b_1 t = \frac{\sum (t_i - \bar{t})(x_i - \bar{x})}{\sum (t_i - \bar{t})^2}$$

$$b_0 = \bar{X} - b_1 \bar{t}, \quad b_1 = \frac{n \sum t_i x_i - \sum t_i \sum x_i}{n \sum t_i^2 + (\sum t_i)^2} = \frac{\sum t_i x_i - \bar{t} \sum x_i - \sum \frac{x_i}{n}}{\sum t_i^2 - \frac{1}{n} (\sum t_i)^2}$$

① Venemos que no se altera con selección del origen

Consideremos el plano $(t-a, x-b)$



Entonces, la tendencia lineal en el plano $(t-a, x-b)$

tiene coeficientes b_0' y b_1' como $t_i' = t_i - a$
 $x_i' = x_i - b$

$$\begin{aligned}
 b_1' &= \frac{n \sum t_i' x_i' - \sum t_i' \sum x_i'}{n \sum t_i'^2 - (\sum t_i')^2} \\
 &= \frac{n \sum (t_i - a)(x_i - b) - \sum (t_i - a) \sum (x_i - b)}{n \sum (t_i - a)^2 - \left\{ \sum (t_i - a) \right\}^2} \\
 &= \frac{n \sum \{t_i x_i - b t_i - a x_i + a b\} - \{(n \bar{t} - n a)(n \bar{x} - n b)\}}{n \sum (t_i^2 - 2 t_i a + a^2) - \{n \bar{t} - n a\}^2} \\
 &= \frac{n \left\{ \sum (t_i x_i) - b n \bar{t} - a n \bar{x} + n a b \right\} - \{n^2 \{ \bar{t} \bar{x} - b \bar{t} - a \bar{x} + a b \}\}}{n \sum (t_i^2 - 2 n \bar{t} + n a^2) - n^2 \{ \bar{t} - a \}^2} \\
 &= \frac{n \left\{ \sum t_i x_i - b n \bar{t} - a n \bar{x} + n a b - n (\bar{t} \bar{x} - b \bar{t} - a \bar{x} + a b) \right\}}{n \left\{ \sum t_i^2 - 2 n \bar{t} + n a^2 - n \{ \bar{t} - a \}^2 \right\}} \\
 &= \frac{\sum t_i x_i - n b \bar{t} - a n \bar{x} + n a b - n \bar{t} \bar{x} + n b \bar{t} + n a \bar{x} - n a b}{\sum t_i^2 - 2 n \bar{t} + n a^2 - n (\bar{t}^2 - 2 \bar{t} a + a^2)} \\
 &= \frac{\sum t_i x_i - n \bar{t} \bar{x}}{\sum t_i^2 - 2 n \bar{t} + n a^2 - n \bar{t}^2 + 2 n \bar{t} a - n a^2} \\
 &= \frac{\sum t_i x_i - n \bar{t} \bar{x}}{\sum t_i^2 - n \bar{t}^2} = b_1 \quad \text{Original de la tendencia lineal} \\
 &\qquad\qquad\qquad \hat{x}_t = b_0 + b_1 t
 \end{aligned}$$

Ahora,

$$\begin{aligned} b_0' &= \bar{x}' - b_1' \bar{t}' = \frac{1}{n} \sum_{i=1}^n x_i' - b_1 \frac{1}{n} \sum_{i=1}^n t_i' \\ &= \bar{x} - b - b_1 \{ \bar{t} - a \} \end{aligned}$$

Así pues, con $b_0' = \bar{x} - b - b_1 \{ \bar{t} - a \}$
 $b_1' = b_1 \{ \text{original} \}$

resulta que

$$\begin{aligned} \hat{x}_{t_i}' &= b_0' + b_1' t_i' = \bar{x} - b - b_1 \{ \bar{t} - a \} + b_1 (t_i - a) \\ &= \bar{x} - b - b_1 \{ \bar{t} - a \} - t_i + a \\ &= \bar{x} - b - b_1 \{ \bar{t} - t_i \} \end{aligned}$$

Como es sobre el plan $(t-a, x-b)$ entonces $\hat{x}_{t_i}' = \hat{x}_{t_i} - b$

$$\text{Así, } \hat{x}_{t_i}' = \bar{x} - b - b_1 \{ \bar{t} - t_i \}$$

$$\Leftrightarrow \hat{x}_{t_i} - b = \bar{x} - b - b_1 \{ \bar{t} - t_i \}$$

$$\Leftrightarrow \hat{x}_{t_i} = \bar{x} - b_1 \{ \bar{t} - t_i \}$$

$$\Leftrightarrow \hat{x}_{t_i} = \underbrace{\bar{x} - b}_{{}=b_0} \bar{t} + b_1 t_i$$

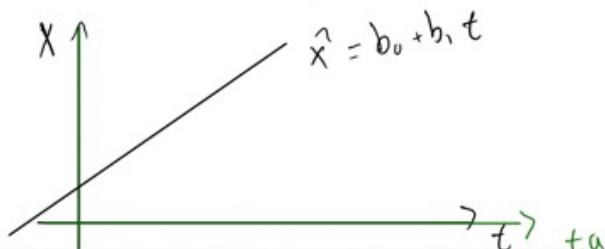
$$\Leftrightarrow \hat{x}_{t_i} = b_0 + b_1 t_i$$

Por lo que no varía bajo cambio de origen //

(2) No se altera respecto a la unidad de medida de tiempo

Consideremos a $t' = ta$, $a \in \mathbb{R}$

Deducimos que si t es medida de tiempo entonces t' es otra medida de tiempo. La estimación de b_0 y b_1 de \hat{x}_t' se da en el plano $(t', x) = (ta, x)$, $a \in \mathbb{R}$



Hacemos las estimaciones en $(t', x) \sim (at, x)$

Así pues,

$$b_1 = \frac{n \sum \epsilon_i x_i - \bar{\epsilon} \sum x_i}{n \sum t_i^2 - (\sum \epsilon_i)^2}$$

$$= \frac{n \sum a t_i x_i - \bar{a} \bar{t} \sum x_i}{n \sum (a t_i)^2 - (\sum a t_i)^2}$$

$$= \frac{n a \sum \epsilon_i x_i - a(n \bar{\epsilon})(n \bar{x})}{n a^2 \sum t_i^2 - a^2 n^2 \bar{x}^2}$$

$$= \frac{\sqrt{a} \left\{ \sum \epsilon_i x_i - \bar{\epsilon} \bar{x} \right\}}{\sqrt{a} \left\{ a \sum \epsilon_i^2 - a n \bar{x}^2 \right\}}$$

$$\begin{aligned}
 \text{Así, } b_0' &= \bar{x} - b_1' \bar{\epsilon}' \\
 &= \bar{x} - \frac{\sum \epsilon_i x_i - \bar{\epsilon} \bar{x}}{a \left\{ \sum t_i^2 - n \bar{x}^2 \right\}} \quad \frac{1}{n} \sum t_i' \\
 &= \bar{x} - \frac{\sum \epsilon_i x_i - \bar{\epsilon} \bar{x}}{\cancel{a} \left\{ \sum t_i^2 - n \bar{x}^2 \right\}} \quad \frac{1}{n} \sum \cancel{a} t_i \\
 &= \bar{x} - \frac{\sum \epsilon_i x_i - \bar{\epsilon} \bar{x}}{\sum t_i^2 - n \bar{x}^2} \quad \bar{\epsilon} \\
 &= \bar{x} - b_1 \bar{\epsilon} \\
 &\quad \underbrace{\text{original}}_{\text{original}} \quad \therefore b_0' = b_0 \quad \{ \text{original}
 \end{aligned}$$

Así pues, tenemos que

$$\hat{x}_{at} = b_0 + b_1 t \quad (= \hat{x}_t)$$

\therefore No cambia respecto a unidades de
tiempo

Esta respuesta está equivocada, pues ϵ

- (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) Correlación entre \tilde{X} y Y

$$\tilde{X}_t = X_t - \hat{X}_t$$

$$t \rightarrow (0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11)$$

dónde $t_0 = 2005$

Primero calculamos b_0, b_1

$$\sum X_{it} i = 3421 \quad \sum t_i = 66$$

$$\sum X_i = 659 \quad \sum t_i^2 = 506$$

$$b_1 = \frac{3421 - (659 \cdot \frac{66}{12})}{506 - (\frac{66^2}{12})} = -1.42307692307692$$

$$b_0 = \frac{659}{12} - b_1 \cdot \frac{66}{12} = 62.7435897435897$$

$$\sum Y_{it} i = 1965 \quad \sum t_i = 66$$

$$\sum Y_i = 333 \quad \sum t_i^2 = 506$$

$$S_{\tilde{X}Y} = \frac{\sum \tilde{X}_i Y_i - \sum \tilde{X}_i \sum Y_i / n}{n-1}$$

$$= \frac{9.730769231 - (7.10543E-14) \left(\frac{333}{12} \right)}{12-1}$$

$$= 0.8836$$

$$\text{Var } y = 3.49 \quad \text{Var } \tilde{x} = 14.12$$

$$\bar{x} = 3.757658846$$

$$\therefore \text{Cor}(X, Y) = 0.06738692 \quad \text{Bien}$$

ii) Correlación entre \tilde{x} y \tilde{y}

$$\tilde{x}_t = x_t - \hat{x}_t$$

$$\hat{\gamma}_t = \gamma_t - \hat{\gamma}_t$$

Primero calculamos a_0, a_1, \dots

$$q_1 = \frac{1965 - (333 \cdot \frac{66}{12})}{506 - \left(\frac{66^2}{12}\right)} = 0.933566433566434$$

$$q_0 = \frac{333}{12} - q_1 \cdot \frac{66}{12} = 22.6153846153846$$

$$\bar{SXY} = \frac{9.730769231 - (7.10543E-14) \left(\frac{-2.13163E-14}{12} \right)}{12 - 1}$$

$$= 0.884$$

$$\text{Var } \tilde{y} = 0.8744$$

$$\Gamma \tilde{\gamma} = 0.93509356$$

$$\therefore \text{Corr}(\tilde{x}, \tilde{y}) = 0.7517506$$

y sabemos que $\text{cor}(x,y) = -0.7375285$

Es decir, $\hat{X}_t - \bar{X}_t$ correlacionado con \bar{Y} disminuye porque X_t contiene a todas las y_i 's

Dado que las y_i 's son no correlacionadas, resulta que la correlación es casi cercana a cero

Ahora bien, para la correlación entre \bar{X} y \bar{Y} vuelve a aumentar porque ahora estimamos tanto \bar{X} como \bar{Y}

Aquí sucede que involucramos en segunda ocasión a la muestra pero ahora en \bar{Y} . Nuevamente, cada una modela el error abs. para X_t y \bar{Y}_t

Tenemos involucradas a b_0 y b_1 los par. dc est. de ambas var. aleat. Aquí la muestra se involucra en ambos estimaciones \hat{X}_t y \hat{Y}_t

Aquí entra en juego las b_0 y b_1 que al correlacionar sus componentes tenemos correlaciones mayores a donde solo estimamos a \hat{X}_t

Pero en términos abs. es menor a la correlación de X con Y porque las correlaciones de \hat{X}_t y \hat{Y}_t consideran las diferencias entre los datos observados y la estimación

Mientras que $\text{Cor}(X, Y)$ solo considera los datos

Esta conclusión no es correcta, lo que pasa es que en el primer caso se obtuvo una correlació

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.

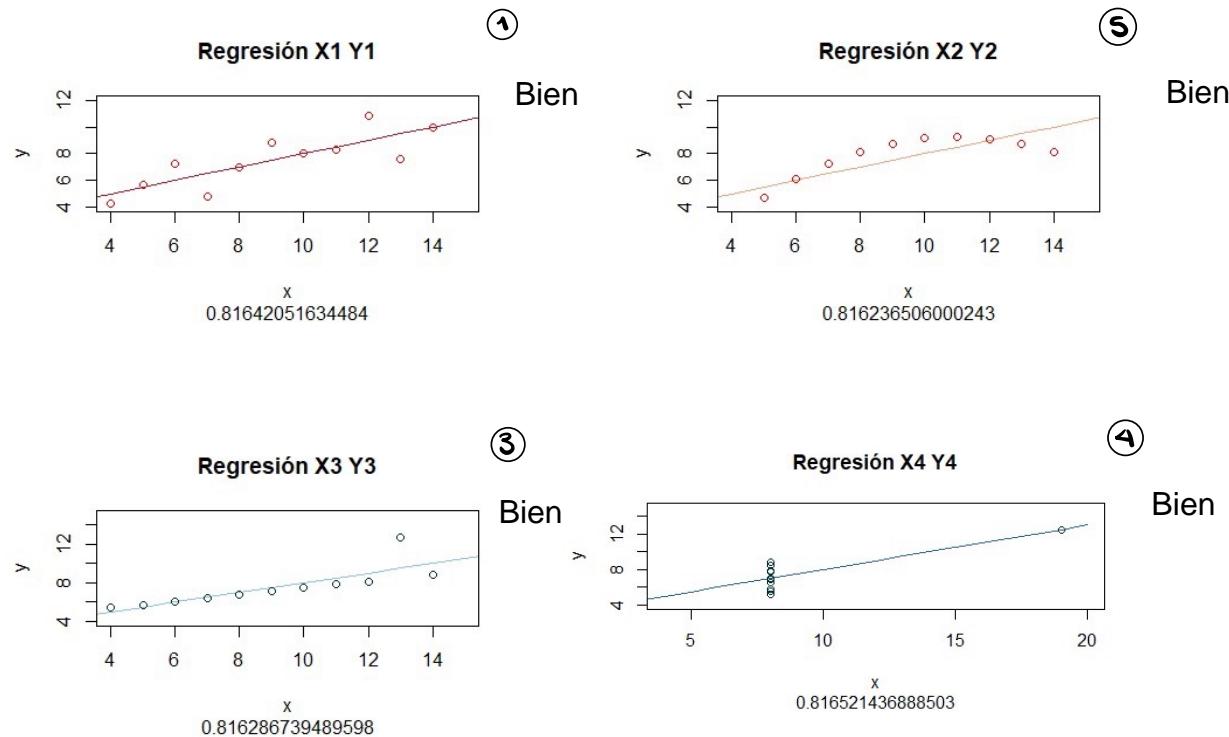
Coeficientes de correlación:

$$\text{COR}(X_1, Y_1) = 0.81642052$$

$$\text{COR}(X_2, Y_2) = 0.8162365$$

$$\text{COR}(X_3, Y_3) = 0.8162867$$

$$\text{COR}(X_4, Y_4) = 0.8165214$$



Se puede concluir tanto los parámetros como sus correlaciones del modelo son casi iguales. Pero los puntos son distintos, por lo que esto nos enseña que puedes tener puntos diferentes y sin embargo, una correlación y parámetros parecidos. Diferentes datos pueden darnos una misma estimación de recta de regresión. Los datos de la gráfica 1 y 3 se adaptan a una MRLS, los de la 2 parecen ser más cuadráticos.

En la gráfica 4 los datos de x parece que no tiene variabilidad, más que por un punto el cual no nos ayuda para hacer un pronóstico sobre el comportamiento de y a diferentes valores de x .

En el caso de la gráfica 3 se tiene una observación que es

Anexo: código en R

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins Project: (None)

R demos Tirar dado.R Aguja simulacion.R CitasDr.R Tarea 2.R

```
1 #Sección 1
2 X<- c(60,62,61,55,53,60,63,53,52,48,49,43)
3 Y <-c(23,23,25,25,26,26,29,30,30,32,33,31)
4 Xm <- mean(X)
5 Ym<- mean(Y)
6 cor(X,Y) #correlación X Y
7
8 t<- c(0,1,2,3,4,5,6,7,8,9,10,11) # tiempo 0 = año 2005
9 tm<- mean(t)
10
11 b0tx<= {length(X)*sum(X*t)-sum(X)*sum(t)}/{length(t)*sum(t^2)-{sum(t)}^2}
12 b0tx<- Xm-b0tx*tm
13
14 xt <-b0tx+b0tx*t
15 Xgt<-X-Xt      #X gorrito de t
16
17 b0ty<= {length(Y)*sum(Y*t)-sum(Y)*sum(t)}/{length(t)*sum(t^2)-{sum(t)}^2}
18 b0ty<- Ym-b0ty*tm
19
20 Yt <-b0ty+b0ty*t
21 Ygt<-Y-Yt      #Y gorrito de t
```

1:1 (Top Level) R Script

Console

Escribe aquí para buscar

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins Project: (None)

R demos Tirar dado.R Aguja simulacion.R CitasDr.R Tarea 2.R

```
81     xlim = c(4, 15),
82     main = "Regresión X2 Y2",
83     ylab = "y",
84     sub= cor(x2,y2))
85 points(x2, y2,type = "p", col="red")
86
87 plot(x,y3,type = "l", col= "#92C5DE",
88       ylim = c(4, 15),
89       xlim = c(4, 15),
90       main = "Regresión X3 Y3",
91       ylab = "y",
92       sub= cor(x3,y3))
93 points(x3, y3,type = "p", col="#003333")
94
95 plot(x,y4,type = "l", col = "#2166AC",
96       ylim = c(4, 15),
97       xlim = c(4, 20),
98       main = "Regresión X4 Y4",
99       ylab = "y",
100      sub= cor(x4,y4))
101 points(x4, y4,type = "p" , col="#003333") #Fin
```

101:47 (Top Level) R Script

Console

Escribe aquí para buscar

12:48 p.m. 07/09/2021

RStudio

File Edit Code View Session Build Debug Profile Tools Help

Go to file/function Addins

R demos Tirar dado.R Aguja simulacion.R CitasDr.R Tarea 2.R

Run Source

```
62 #graficos
63 x<-1:20
64
65 y1<-b0x1 + b1x1*x
66 y2<-b0x2 + b1x2*x
67 y3<-b0x3 + b1x3*x
68 y4<-b0x4 + b1x4*x
69 par(mfrow=c(2,2))
70
71 plot(x,y1,type = "l", col ="#B2182B",
72       ylim = c(4, 12),
73       xlim = c(4, 15),
74       main = "Regresión X1 Y1",
75       ylab = "y",
76       sub= cor(x1,Y1))
77 points(X1, Y1,type = "p", col="red")
78
79 plot(x,y2,type = "l", col="#F4A582",
80       ylim = c(4, 12),
81       xlim = c(4, 15) ,
82       main = "Regresión X2 Y2",
83       ylab = "y",
84       sub= cor(x2,Y2))
85
86 plot(x,y3,type = "l", col="#A5C494",
87       ylim = c(4, 12),
88       xlim = c(4, 15) ,
89       main = "Regresión X3 Y3",
90       ylab = "y",
91       sub= cor(x3,Y3))
92
93 plot(x,y4,type = "l", col="#4CAF50",
94       ylim = c(4, 12),
95       xlim = c(4, 15) ,
96       main = "Regresión X4 Y4",
97       ylab = "y",
98       sub= cor(x4,Y4))
```

1:1 (Top Level) R Script

Console

Escribe aquí para buscar

12:48 p.m. 07/09/2021

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

R demos Tirar dado.R Aguja simulacion.R CitasDr.R Tarea 2.R

Run Source

```
22
23 #Comparacion de correlaciones
24 cor(Xgt,Y)
25 cor(Xgt,Ygt)
26 cor(X,Y)
27
28 #Sección 2
29 #Datos
30 X1 <-c(10,8,13,9,11,14,6,4,12,7,5)
31 Y1<-c(8.04,6.95,7.58,8.81,8.33,9.96,7.24,4.26,10.84,4.82,5.68)
32
33 X2<-c(10,8,13,9,11,14,6,4,12,7,5)
34 Y2<-c(9.14,8.14,8.74,8.77,9.26,8.10,6.13,3.10,9.13,7.26,4.74)
35
36 X3<-c(10,8,13,9,11,14,6,4,12,7,5)
37 Y3<-c(7.46,6.77,12.74,7.11,7.81,8.84,6.08,5.39,8.15,6.42,5.73)
38
39 X4<-c(8,8,8,8,8,8,19,8,8,8)
40 Y4<-c(6.58,5.76,7.71,8.84,8.47,7.04,5.25,12.5,5.56,7.91,6.89)
41
42 #Parametros b1 y b0 de cada par de datos
43
```

1:1 (Top Level) R Script

Console

Escribe aquí para buscar

12:48 p.m. 07/09/2021

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

R demos Tirar dado.R Aguja simulacion.R CitasDr.R Tarea 2.R

Source on Save Run Source

```
42 #Parametros b1 y b0 de cada par de datos
43
44 b1x1<-{length(Y1)*sum(Y1*X1)-sum(Y1)*sum(X1)}/{length(X1)*sum(X1^2)-{sum(X1)}^2}
45 b0x1<- mean(Y1)-b1x1*mean(X1)
46
47 b1x2<-{length(Y2)*sum(Y2*X2)-sum(Y2)*sum(X2)}/{length(X2)*sum(X2^2)-{sum(X2)}^2}
48 b0x2<- mean(Y2)-b1x2*mean(X2)
49
50 b1x3<-{length(Y3)*sum(Y3*X3)-sum(Y3)*sum(X3)}/{length(X3)*sum(X3^2)-{sum(X3)}^2}
51 b0x3<- mean(Y3)-b1x3*mean(X3)
52
53 b1x4<-{length(Y4)*sum(Y4*X4)-sum(Y4)*sum(X4)}/{length(X4)*sum(X4^2)-{sum(X4)}^2}
54 b0x4<- mean(Y4)-b1x4*mean(X4)
55
56 #correlaciones
57 cor(X1,Y1)
58 cor(X2,Y2)
59 cor(X3,Y3)
60 cor(X4,Y4)
61
62 #graficos
```

1:1 (Top Level) ▾

R Script

Console

Escribe aquí para buscar



12:48 p. m.

07/09/2021

values

| | |
|------|--------------------|
| b0tx | 62.7435897435897 |
| b0ty | 22.6153846153846 |
| b0x1 | 3.00009090909091 |
| b0x2 | 3.0009090909091 |
| b0x3 | 3.00245454545454 |
| b0x4 | 3.00172727272727 |
| b1tx | -1.42307692307692 |
| b1ty | 0.933566433566434 |
| b1x1 | 0.500090909090909 |
| b1x2 | 0.4999999999999999 |
| b1x3 | 0.499727272727273 |
| b1x4 | 0.499909090909091 |