Regresion Lineal

Clase 5

Analizar la relación de dos variables (ASOCIACIÓN)

- Recta poblacional

$$\boxed{Y = a + b.x + \varepsilon}$$

donde

- Y: var. dependiente
- X: var. independiente
- $a; b$: coeficientes
- ε: Error aleatorio

- Recta Estimada

$$\boxed{\hat{Y} = a + b.x}$$

Ejercicio 1

ELIPSE VIRTUAL
Asesorías Universitarias

Hallar la Recta estimada "$\hat{Y} = a + b.x$"

	X	Y		
	Longitud	Periodo		
N° puntos	L (m)	T (s)	X*Y	X²
1	0.95	2.03	1.9285	0.9025
2	0.90	1.99	1.791	0.8100
3	0.85	1.95	1.6575	0.7225
4	0.80	1.88	1.504	0.6400
5	0.70	1.78	1.246	0.4900
6	0.65	1.7	1.105	0.4225
Total	4.85	11.33	9.23	3.99

DATOS

- $n = 6$
- $\sum x.y = 9{,}23$
- $\sum x^2 = 3{,}99$
- $\bar{x} = \dfrac{\sum x}{n} = \dfrac{4{,}85}{6} = 0{,}81$
- $\bar{y} = \dfrac{\sum y}{n} = \dfrac{11{,}33}{6} = 1{,}89$

Recta de Regresión Estimada: $\hat{Y} = a + b.X$

Calculemos los Coef. por Minimos Cuadrados Ordinarios (M.C.O.)

$$\boxed{b = \frac{\sum x.y - n.\bar{x}.\bar{y}}{\sum x^2 - n.\bar{x}^2}} \rightarrow b = \frac{9{,}23 - 6.0{,}81.1{,}89}{3{,}99 - 6.(0{,}81)^2} = 1{,}10$$

$$\boxed{a = \bar{Y} - b.\bar{x}} \rightarrow a = 1{,}89 - 1{,}10.0{,}81 = 1{,}00$$

$$\therefore \hat{Y} = 1{,}00 + 1{,}10.X$$

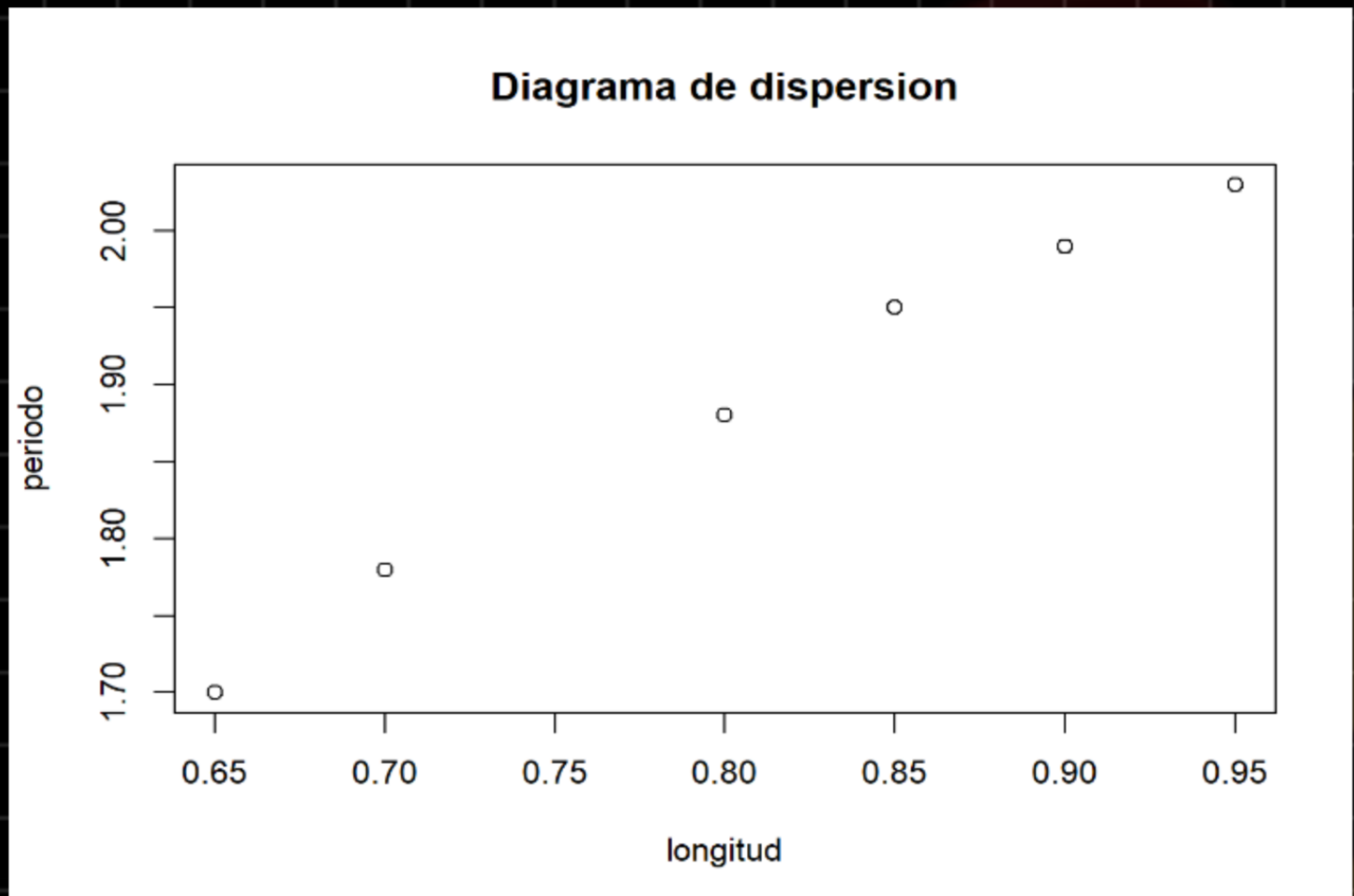
"a": no tiene interpretación

Interpretación de los Coeficientes

- a: Cuando la longitud es cero, entonces el periodo es 1,00 seg
- b: Cuando la longitud aumenta en 1 metro, entonces el periodo aumenta en 1,10 seg.

En R-studio

```
# Ejercicio 1
longitud<- c(0.95,0.90,0.85,0.80,0.70,0.65)
periodo<- c(2.03,1.99,1.95,1.88,1.78,1.70)
# graficar
plot(longitud,periodo,main = 'Diagrama de dispersion')
```




```
# Hallar el modelo
lm(periodo ~ longitud)
```

```
Coefficients:
(Intercept)     longitud
      1.002        1.097
>
```

a = 1.002, b = 1.097

Recta Estimada

$$\hat{Y} = a + b.X$$

$$\therefore \hat{Y} = 1{,}002 + 1{,}097.X$$

Interpretación

Se observa una RELACIÓN lineal y positiva entre las variables longitud y periodo.

OBS:

X	Y	$\hat{Y}$	$Y-\hat{Y}$
0,95	2,03	2,04	0,01

$1{,}002 + 1{,}097.(0{,}95)$

Ejercicio 2

Del ejercicio anterior calcule el coeficiente de determinación e interprete.

	X	Y					
	Longitud	Periodo					
N° puntos	L (m)	T (s)	X*Y	X²	Y estimado	(Yi - Yi est)²	(Yi - ȳ)²
1	0.95	2.03	1.9285	0.9025	2.043726708	0.000188	0.020069
2	0.90	1.99	1.791	0.8100	1.988881988	0.000001	0.010336
3	0.85	1.95	1.6575	0.7225	1.934037267	0.000255	0.003803
4	0.80	1.88	1.504	0.6400	1.879192547	0.000001	0.000069
5	0.70	1.78	1.246	0.4900	1.769503106	0.000110	0.011736
6	0.65	1.7	1.105	0.4225	1.714658385	0.000215	0.035469
Total	**4.85**	**11.33**	**9.23**	**3.99**		**0.000770**	**0.081483**

$$\boxed{R^2 = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2}}$$

- $R^2 = 1 - \dfrac{0{,}000770}{0{,}081483} = 0{,}990548 \approx 99{,}05\%$

Interpretación

La variabilidad del periodo es explicado por la longitud del pendulo en un 99,05%

En R-studio

```r
# con el comando summary
summary(lm(periodo ~ longitud))
```

coeficientes

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.00168    0.04368   22.93 2.14e-05 ***
longitud     1.09689    0.05357   20.47 3.36e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01388 on 4 degrees of freedom
Multiple R-squared:  0.9905,    Adjusted R-squared:  0.9882
F-statistic: 419.2 on 1 and 4 DF,  p-value: 3.361e-05

>
```

r2

- $R^2 = 0{,}9905 = 99{,}05\%$

Interpretación

El "99,05%" de la variabilidad del periodo es explicado por la longitud y el "0,95%" no es explicado por el modelo.

OBS

1. El R^2 es llamado bondad del modelo.
2. Cuando tenemos varios modelos para explicar una misma variable Y. El MEJOR MODELO es aquel con MAYOR R^2.

Pregunta 1 (4.0 puntos)

La base de datos `publicidad.csv` contiene las siguientes variables medidas en 200 campañas de venta de un producto de una empresa:

Variable	Descripción
youtube	Inversión en publicidad en youtube en miles de US$
facebook	Inversión en publicidad en facebook en miles de US$
periodicos	Inversión en publicidad en periódicos en miles US$.
ventas	Ventas en miles de unidades

a) (1.0 punto) Estudie la posible relación lineal de cada tipo de inversión en publicidad con las ventas del producto. Indique cuál de ellas tiene una mayor asociación lineal con las ventas. Justifique su respuesta mediante gráficos y medidas apropiadas.

b) (1.0 punto) Estime un modelo de regresión para predecir las ventas en función de la variable escogida en el ítem anterior. Interprete los coeficientes de regresión estimados.

c) (1.0 punto) Considere ahora que un analista sugiere que se use como variable para predecir las ventas la suma total de las inversiones en publicidad. Estime este modelo y compárelo con el modelo hallado en el ítem anterior. Indique que modelo sería el más adecuado. Justifique su respuesta en base a sus resultados.

d) (1.0 punto) Considerando el mejor modelo encontrado encontrado en c), estime el número de unidades vendidas del producto si se planea invertir 30 mil dólares en cada tipo de publicidad.

```
# subimos el archivo
library(readr)
publicidad <- read_csv("D:/ASESORIA/ELIPSE/2023/ADELANTO 2023 -
View(publicidad)

# modelo 1
# x:youtube ; y:ventas
modelo1<- lm(publicidad$ventas ~ publicidad$youtube)
summary(modelo1)
```

```
Residuals:
   Min     1Q Median     3Q    Max
-5.720 -3.027  0.076  2.975  6.083

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        8.558834   0.479181   17.86   <2e-16 ***
publicidad$youtube 0.046720   0.002347   19.91   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.411 on 198 degrees of freedom
Multiple R-squared:  0.6669,    Adjusted R-squared:  0.6652
F-statistic: 396.4 on 1 and 198 DF,  p-value: < 2.2e-16

>
```

coeficientes del modelo 1

r2 del modelo 1