Análisis de diferentes representaciones en la regresión lineal simple

5ª. Jornada sobre el Uso de los Sistemas de Cómputo Simbólico en la Enseñanza de las Matemáticas

M.C. Gudelia Figueroa Preciado

M. C. Irma Nancy Larios Rodríguez

Introducción

- En este trabajo se presentarán dos situaciones donde se ejemplifican diferentes representaciones que facilitan la enseñanza del análisis de regresión lineal simple. Para ello se utiliza tanto la tecnología (software estadístico, applets en Internet) como hojas de trabajo a desarrollar en el salón de clase.

Situación Número 1: “Introducción al Método de Mínimos Cuadrados”

Objetivo:

Que el estudiante comprenda que la recta obtenida por el método de mínimos cuadrados es la que minimiza la suma de los cuadrados de las distancias de estos puntos a la recta.

Estrategia

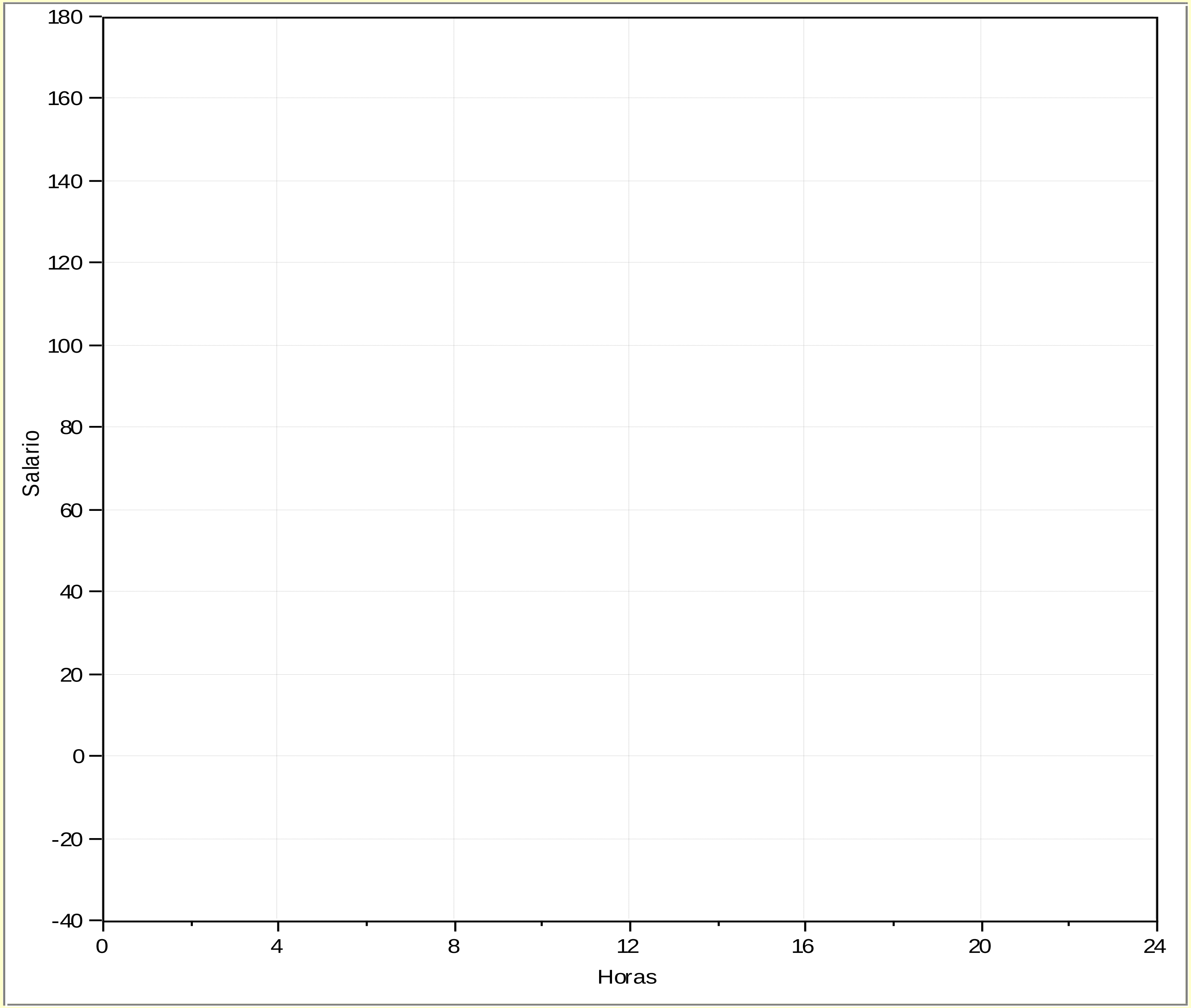
- Uso de hoja de trabajo, en el salón de clases, con el problema a desarrollar.
- Desarrollo de la actividad por parte de los estudiantes.
- Comparación y análisis grupal, de los resultados obtenidos con la solución estimada por mínimos cuadrados.
- Interpretación de los resultados.

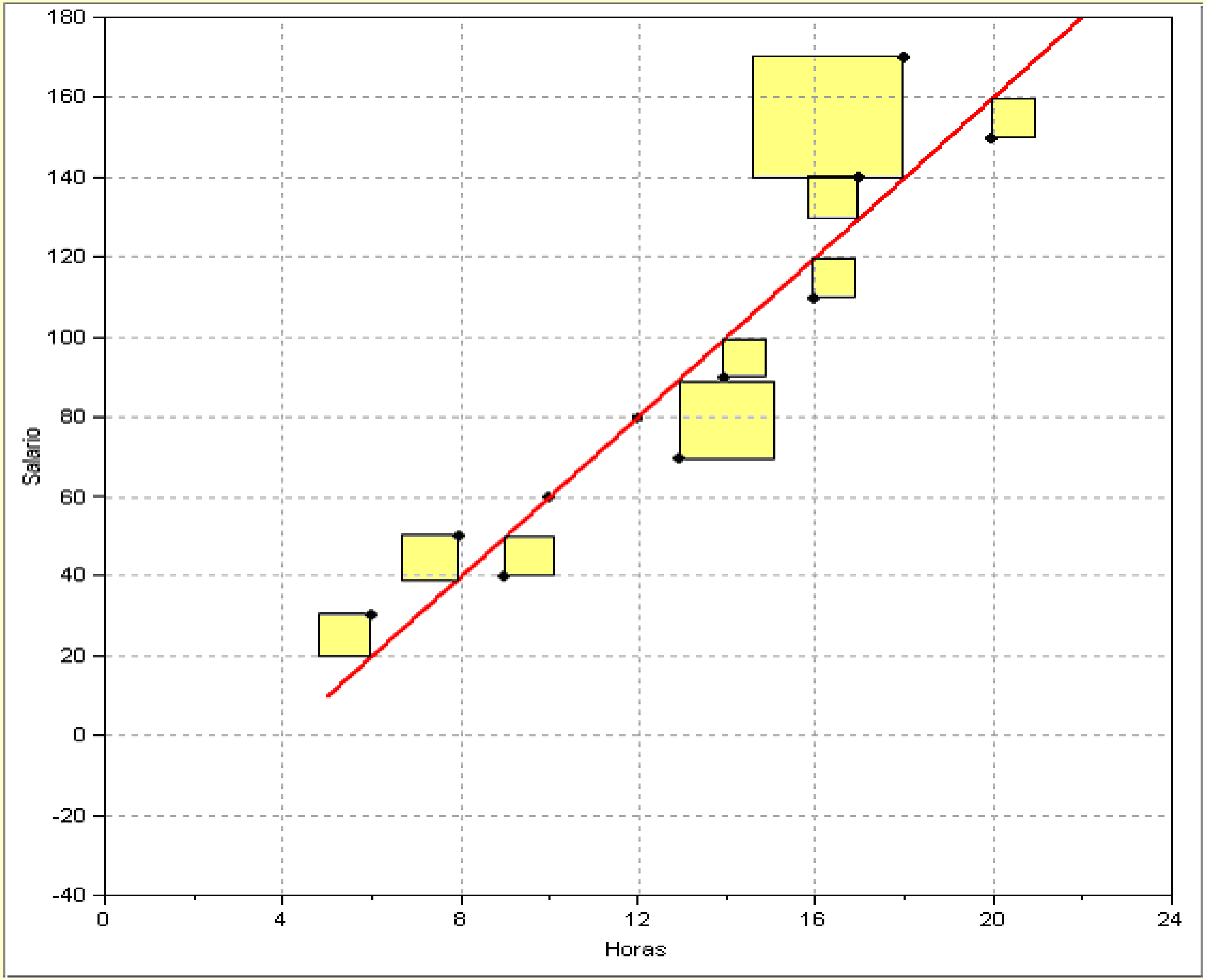
Hoja de trabajo

Los datos que a continuación se proporcionan corresponden a las horas trabajadas y el sueldo obtenido por once empleados de una empresa.

Horas	6	8	9	10	12	13	14	16	17	18	20
Salario	30	50	40	60	80	70	90	110	140	170	150

- Grafique estos puntos en el plano que se le proporciona
- Trace la recta que considera mejor se ajusta a estos puntos.
- Dibuje las distancias de estos puntos a la recta y forme cuadrados.
- Estime la suma de las áreas de los cuadrados dibujados.


Salario
180
160
140
120
100
80
60
40
20
0
-20
-40
0
4
8
12
16
20
24
Horas



180
160
140
120
100
80
60
40
20
0
-20
-40
Salario
0
4
8
12
16
20
24
Horas

Resultado obtenido por un estudiante:

Recta de regresión Lineal.

Problema 1. Los datos que a continuación se proporcionan corresponden a las horas trabajadas y el sueldo devengado por once empleados de una empresa.

a. Grafique estos puntos en el plano que se le proporciona.

Horas	6	8	9	10	12	13	14	16	17	18	20
Salario	30	50	40	60	80	70	90	110	140	170	150

$l_1 = \frac{7}{12} \cdot 20 = \frac{35}{3}$

$l_2 = \frac{6}{12} \cdot 20 = 10$

$l_3 = \frac{5}{12} \cdot 20 = \frac{25}{3}$

$l_4 = \frac{13}{12} 20 = \frac{65}{3}$

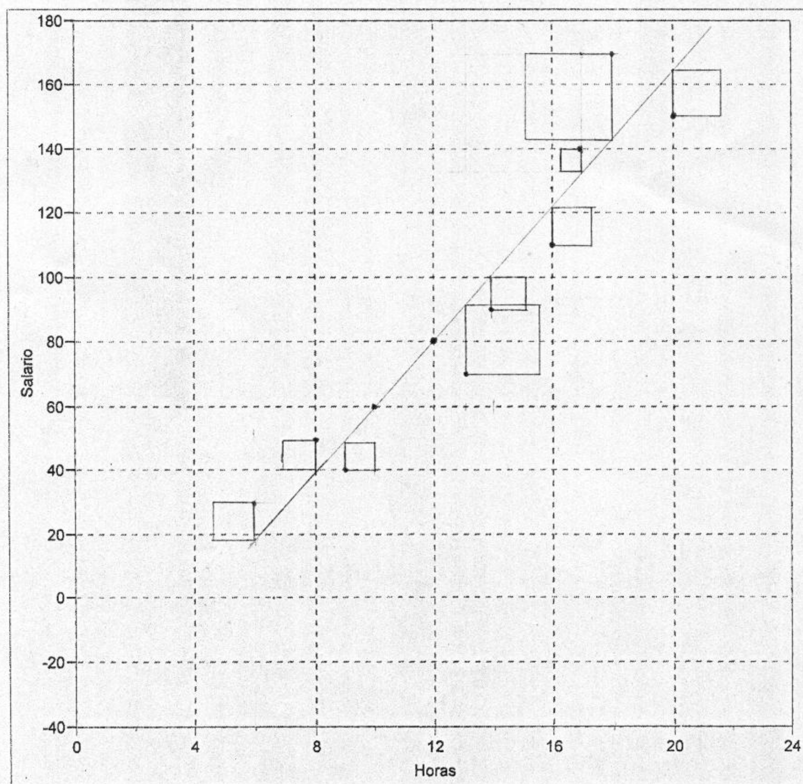
$l_5 = \frac{6}{12} 20 = 10$

$l_6 = \frac{7}{12} 20 = \frac{35}{3}$

$l_7 = \frac{8}{12} 20 = \frac{25}{3}$

$l_8 = \frac{16}{12} \cdot 20 = \frac{80}{3}$

$l_9 = \frac{9}{12} \cdot 20 = 15$



$A_1 = 136.11$

$A_2 = 100$

$A_3 = 69.44$

$A_4 = 469.44$

$A_5 = 100$

$A_6 = 136.11$

$A_7 = 69.44$

$A_8 = 711.11$

$A_9 = 225$

$\sum_{i=1}^{9} A_i = 2016.65$

b. Trace la recta que considera mejor se ajusta a estos puntos.

c. Dibuje las distancias de estos puntos a la recta y forme cuadrados.

d. Estime la suma de las áreas de los cuadrados dibujados.

Observación**: La suma de los cuadrados de los residuales es 2000. Esto es lo que se conoce como suma de cuadrados del error y el método de mínimos cuadrados está minimizando esta suma.**

Ecuación de regresión: $y = -40 + 10x$

Horas	6	8	9	10	12	13	14	16	17	18	20
Salario	30	50	40	60	80	70	90	110	140	170	150
Valor Predicho	20	40	50	60	80	90	100	120	130	140	160
Residual	10	10	-10	0	0	-20	-10	-10	10	30	10
$(\text{Residual})^2$	100	100	100	0	0	400	100	100	100	900	100

Ejemplificación del procedimiento efectuado, utilizando un applet de Internet que se encuentra en la siguiente liga:

http://www.duxbury.com/authors/mcclellandg/tiein/johnson/reg.htm

Este ejemplo está desarrollado en el libro de Johnson, Robert: *Just the Essentials of Elementary Statistics*, Dusbury Press,1995

Situación Número 2: Integración de diferentes formas de representación en el análisis de regresión lineal simple.

- **Objetivos:**

  Que el estudiante:

- **Decida, sobre la representación tabular de los datos y su diagrama de dispersión, si existe o no una relación lineal en las variables involucradas, en las situaciones planteadas.**
- **A partir de los cálculos efectuados, concluya que es necesario integrar las diferentes formas de representación al analizar un conjunto de datos.**

Estrategia

- Uso de hoja de trabajo, en el salón de clases, con el problema a desarrollar.
- Desarrollo de la actividad por parte de los estudiantes, donde a partir de la información dada:

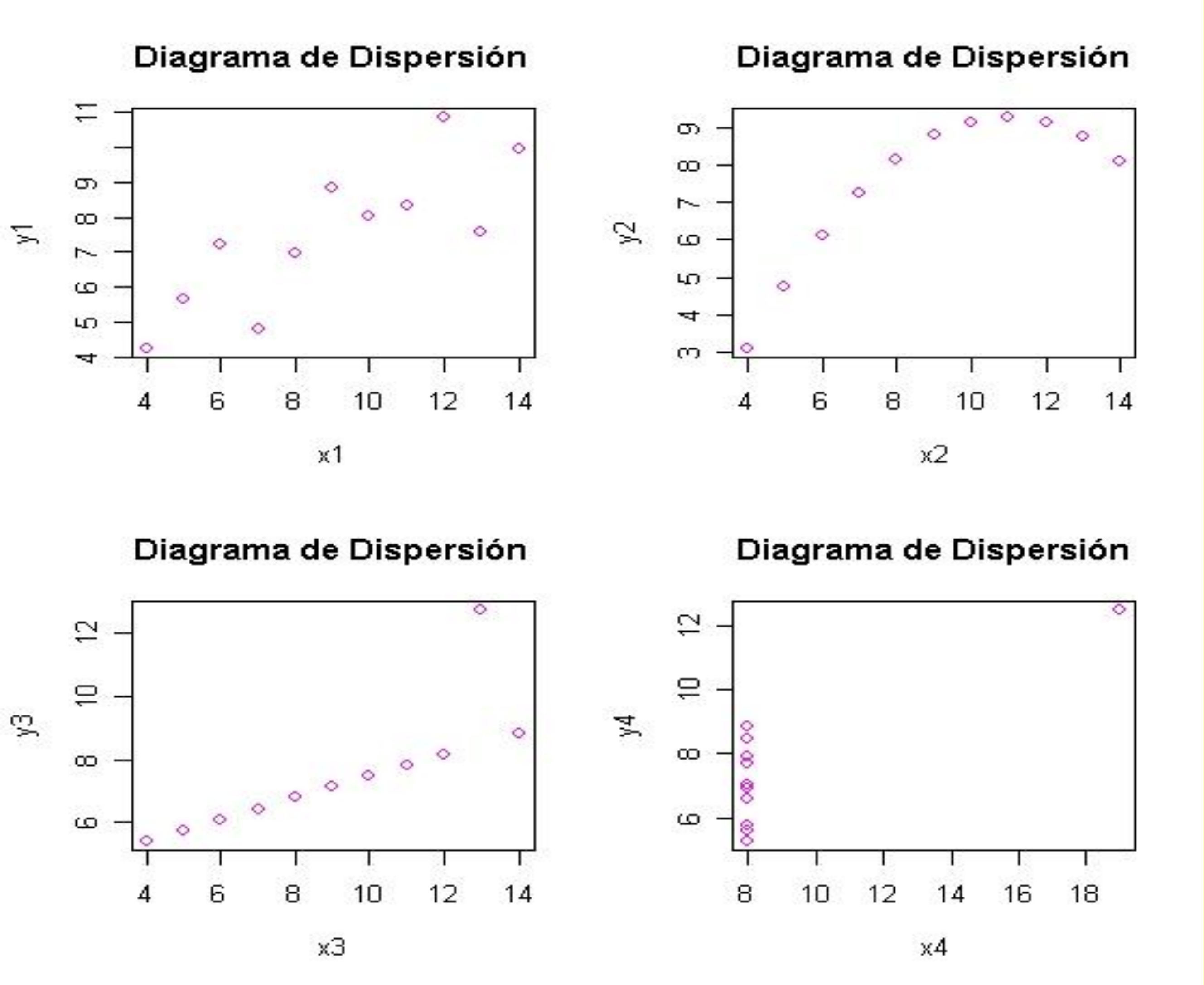
1. *Construyan, utilizando algún software, los diagramas de dispersión respectivos y a partir de ellos propongan si existe o no una relación lineal entre las variables involucradas, en cada una de las situaciones.*
2. *Calculen, utilizando software, la ecuación de regresión, la tabla de análisis de varianza, y el coeficiente de determinación.*
3. *Comparar los cálculos obtenidos con sus respectivos diagramas de dispersión.*
4. *Discusión y análisis grupal sobre los resultados obtenidos.*

X1	Y1	X2	Y2	X3	Y3	X4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Referencia: Anscombe, Francis (1973), *Graphs in Statistical Analysis*, The American Statistician, pp. 195-199.

Programa en R

- X1<-c(10,8,13,9,11,14,6,4,12,7,5)
- Y1<-c(8.04,6.95,7.58,8.81,8.33,9.96,7.24,4.26,10.84,4.82,5.68)
- X2<-c(10,8,13,9,11,14,6,4,12,7,5)
- Y2<-c(9.14,8.14,8.74,8.77,9.26,8.1,6.13,3.1,9.13,7.26,4.74)
- X3<-c(10,8,13,9,11,14,6,4,12,7,5)
- Y3<-c(7.46,6.77,12.74,7.11,7.81,8.84,6.08,5.39,8.15,6.42,5.73)
- X4<-c(8,8,8,8,8,8,8,19,8,8,8)
- Y4<-c(6.58,5.76,7.71,8.84,8.47,7.04,5.25,12.5,5.56,7.91,6.84)
- plot(X1,Y1,main="Diagrama de Dispersión",col="6")
- plot(X2,Y2,main="Diagrama de Dispersión",col="6") ...



¿Existe una relación lineal entre las variables analizadas? En caso de que así sea, aproximadamente, qué coeficiente de determinación le asignaría?

Procedimiento para encontrar las ecuaciones de regresión y las tablas de análisis de varianza utilizando R

- **lm(Y1 ~ X1)**

(Intercept)	*X1*
3.0001	*0.5001*

- **anova(lm.1 <- lm(Y1 ~ X1))**

Response: Y1

	Df	*Sum Sq*	*Mean Sq*	*F value*	*Pr(>F)*
X1	*1*	*27.5100*	*27.5100*	*17.99*	*0.002170 ***
Residuals	*9*	*13.7627*	*1.5292*		

Rsq=27.5/41.27

=0.66

- **lm(Y2 ~ X2)**

(Intercept)	*X2*
3.001	*0.500*

- **anova(lm.2 <- lm(Y2 ~ X2))**

Response: Y2

	Df	*Sum Sq*	*Mean Sq*	*F value*	*Pr(>F)*
X2	*1*	*27.5000*	*27.5000*	*17.966*	*0.002179 ***
Residuals	*9*	*13.7763*	*1.5307*		

Rsq=27.5/41.27

=0.66

- **lm(Y3 ~ X3)**

(Intercept)	*X3*
3.0025	*0.4997*

- **anova(lm.3 <- lm(Y3 ~ X3))**

Response: Y3

	Df	*Sum Sq*	*Mean Sq*	*F value*	*Pr(>F)*
X3	*1*	*27.4700*	*27.4700*	*17.972*	*0.002176 ** *
Residuals	*9*	*13.7562*	*1.5285*		

Rsq=27.5/41.27

=0.66

- **lm(Y4 ~ X4)**

(Intercept)	*X4*
2.9931	*0.5004*

- **anova(lm.4 <- lm(Y4 ~ X4))**

Response: Y4

	Df	*Sum Sq*	*Mean Sq*	*F value*	*Pr(>F)*
X4	*1*	*27.5400*	*27.5400*	*18.019*	*0.002159 ** *
Residuals	*9*	*13.7558*	*1.5284*		

Rsq=27.5/41.27

=0.66

?

¿Conclusiones?