CORRELACIÓN Y REGRESIÓN CON EXCEL Y GEOGEBRA

Cuando se estudian en forma conjunta dos características (variables estadísticas) de una población o muestra, se dice que estamos analizando una variable estadística bidimensional. La correlación es el grado de relación que existe entre ambas características, y la regresión es la forma de expresar matemáticamente dicha relación.

COEFICIENTE DE CORRELACIÓN DE KARL PEARSON

Llamando también coeficiente de correlación producto-momento.

a) Para datos no agrupados se calcula aplicando la siguiente ecuación:

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

r = Coeficiente producto-momento de correlación lineal; $x = X - \bar{X}$; $y = Y - \bar{Y}$

Ejemplo ilustrativo: Con los datos sobre las temperaturas en dos días diferentes en una ciudad, determinar el tipo de correlación que existe entre ellas mediante el coeficiente de PEARSON.

X	18	17	15	16	14	12	9	15	16	14	16	18	ΣX =180
Y	13	15	14	13	9	10	8	13	12	13	10	8	ΣY= 138

Solución:

Se calcula la media aritmética

$$\bar{x} = \frac{\sum x_i}{n}$$

Para X:

$$\bar{X}_X = \frac{180}{12} = 15$$

Para Y:

$$\bar{Y}_Y = \frac{138}{12} = 11{,}5$$

Se llena la siguiente tabla:

X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	xy	y^2
18	13	3	1,5	9	4,5	2,25
17	15	2	3,5	4	7	12,25
15	14	0	2,5	0	0	6,25
16	13	1	1,5	1	1,5	2,25
14	9	-1	-2,5	1	2,5	6,25
12	10	-3	-1,5	9	4,5	2,25
9	8	-6	-3,5	36	21	12,25
15	13	0	1,5	0	0	2,25
16	12	1	0,5	1	0,5	0,25
14	13	-1	1,5	1	-1,5	2,25
16	10	1	-1,5	1	-1,5	2,25
18	8	3	-3,5	9	-10,5	12,25
180	**138**			**72**	**28**	**63**

Se aplica la fórmula:

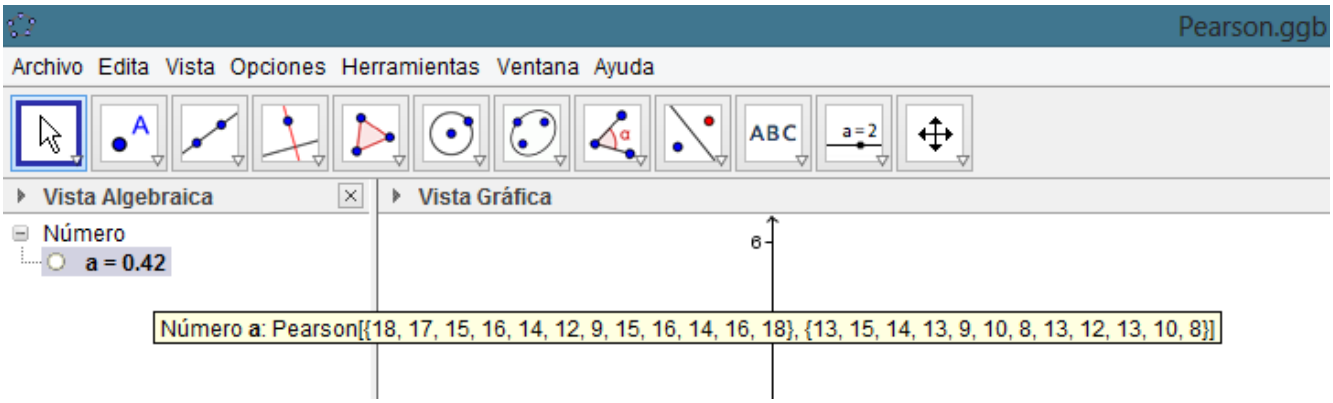
$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} = \frac{28}{\sqrt{(72)(63)}} = 0{,}416$$

Existe una correlación moderada

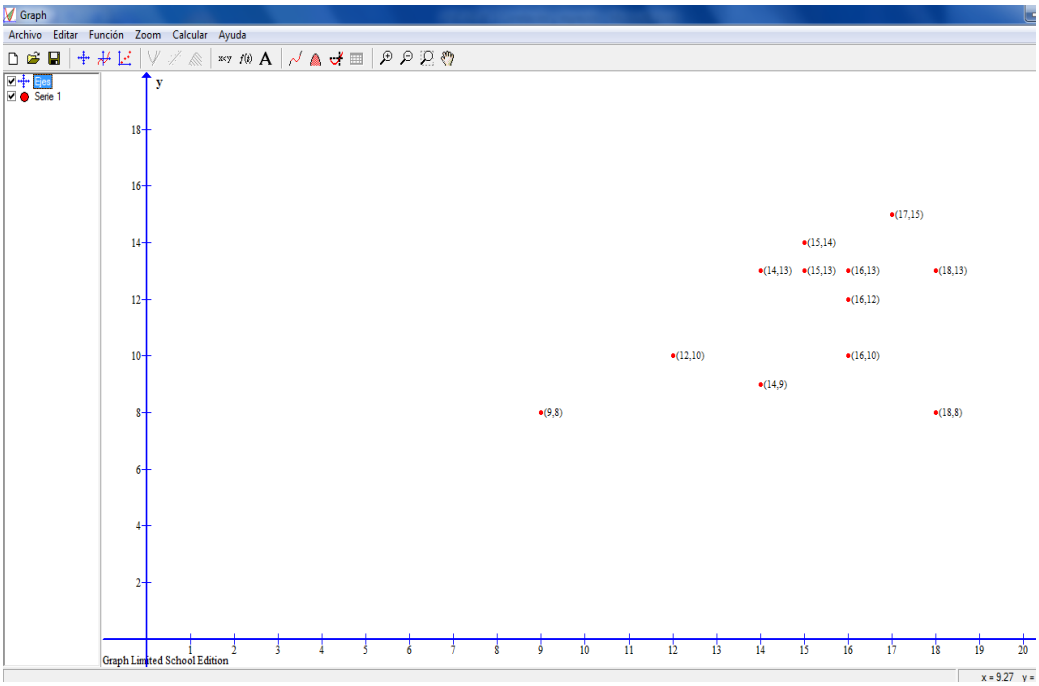
En Excel:

	A	B	C	D	E	F
1	**X**	**Y**				
2	18	13				
3	17	15				
4	15	14				
5	16	13				
6	14	9				
7	12	10				
8	9	8				
9	15	13				
10	16	12				
11	14	13				
12	16	10				
13	18	8				
14						
15	r	0,416	=COEF.DE.CORREL(A2:A13;B2:B13)			

En GeoGebra:



El Diagrama de dispersión en Graph:



b) Para datos agrupados, el coeficiente de Correlación de Pearson se calcula aplicando la siguiente fórmula:

$$r = \frac{n \cdot \sum f \cdot dx \cdot dy - (\sum fx \cdot dx)\,(\sum fy \cdot dy)}{\sqrt{[n \cdot \sum fx \cdot dx^2 - (\sum fx \cdot dx)^2][n \cdot \sum fy \cdot dy^2 - (\sum fy \cdot dy)^2]}}$$

Donde:
n = número de datos; f = frecuencia de celda; fx = frecuencia de la variable X; fy = frecuencia de la variable Y; dx = valores codificados o cambiados para los intervalos de la variable X, procurando que al intervalo central le corresponda $dx = 0$, para que se hagan más fáciles los cálculos; dy = valores codificados o cambiados para los intervalos de la variable X, procurando que al intervalo central le corresponda $dy = 0$, para que se hagan más fáciles los cálculos.

Ejemplo ilustrativo: Con los siguientes datos sobre los Coeficientes Intelectuales (X) y de las calificaciones en una prueba de conocimiento (Y) de 50 estudiantes:

N° de estudiante	X	Y	N° de estudiante	X	Y
1	76	28	26	88	40
2	77	24	27	88	31
3	78	18	28	88	35
4	79	41	29	88	26
5	79	43	30	89	30
6	80	45	31	89	24
7	80	34	32	90	18
8	80	18	33	90	11
9	82	40	34	90	15
10	82	35	35	91	38
11	83	30	36	92	34
12	83	21	37	92	31
13	83	22	38	93	33
14	83	23	39	93	35
15	84	25	40	93	24
16	84	11	41	94	40
17	84	15	42	96	35
18	85	31	43	97	36
19	85	35	44	98	40
20	86	26	45	99	33
21	86	30	46	100	51
22	86	24	47	101	54
23	86	16	48	101	55
24	87	20	49	102	41
25	88	36	50	102	45

1) Elaborar una tabla de dos variables
2) Calcular el coeficiente de correlación

Solución: En la *tabla de frecuencias de dos variables*, cada recuadro de esta tabla se llama una *celda* y corresponde a un par de intervalos, y el número indicado en cada celda se llama *frecuencia de celda*. Todos los totales indicados en la última fila y en la última columna se llaman *totales marginales o frecuencias marginales*, y corresponden, respectivamente, a las frecuencias de intervalo de las distribuciones de frecuencia separadas de la variable X y Y.

Para elaborar la tabla se recomienda:
- Agrupar las variables X y Y en un igual número de intervalos.
- Los intervalos de la variable X se ubican en la parte superior de manera horizontal (fila) y en orden ascendente.

- Los intervalos de la variable Y se ubican en la parte izquierda de manera vertical (columna) y en orden descendente.
Para elaborar los intervalos se procede a realizar los cálculos respectivos:
En la variable X:
Calculando el Rango se obtiene:
$R = x_{máx} - x_{mín} = 102 - 76 = 26$

Calculando el número de intervalos se obtiene:
$n_i = 1 + 3{,}32 \cdot log(n) = 1 + 3{,}32 \cdot log 50 = 6{,}6 = 7$

Calculando el ancho se obtiene:
$i = \frac{R}{ni} = \frac{26}{6{,}6} = 3{,}93 = 4$

En la variable Y:
Calculando el Rango se obtiene:
$R = y_{máx} - y_{mín} = 55 - 11 = 44$

Calculando el número de intervalos se obtiene:
$n_i = 1 + 3{,}32 \cdot log(n) = 1 + 3{,}32 \cdot log 50 = 6{,}64 = 7$

Calculando el ancho se obtiene:
$i = \frac{R}{ni} = \frac{44}{6{,}64} = 6{,}62 = 7$

Nota: Para la variable X se tomará un ancho de intervalo igual a 4 y para la variable Y un ancho de intervalo igual a 7. Debe quedar igual número de intervalos para cada variable, que en este ejemplo es igual a 7.
Contando las frecuencias de celda para cada par de intervalos de las variables X y Y se obtiene la siguiente tabla de frecuencias de dos variables:

		Coeficientes Intelectuales (X)							
		76-79	80-83	84-87	88-91	92-95	96-99	100-103	fy
Calificaciones (Y)	53-59							2	2
	46-52							1	1
	39-45	2	2		1	1	1	2	9
	32-38		2	1	3	3	3		12
	25-31	1	1	4	3	1			10
	18-24	2	4	2	2	1			11
	11-17			3	2				5
	fx	5	9	10	11	6	4	5	50

Interpretación:
- El número 2 es la frecuencia de la celda correspondiente al par de intervalos 76-79 en Coeficiente Intelectual y 39-45 en Calificación obtenida en la prueba de conocimiento.
- El número 5 en la fila de fx es el total marginal o frecuencia marginal del intervalo 76-79 en Coeficiente Intelectual.
- El número 2 en la columna de fy es el total marginal o frecuencia marginal del intervalo 53-59 en Calificación obtenida en la prueba de conocimiento.
- El número 50 es total de frecuencias marginales y representa al número total de estudiantes.

2) Realizando los cálculos respectivos se obtiene la siguiente tabla:

			Coeficientes Intelectuales (X)										
			76-79	80-83	84-87	88-91	92-95	96-99	100-103				
		dx / dy	-3	-2	-1	0	1	2	3	fy	$fy \cdot dy$	$fy \cdot dy^2$	$f \cdot dx \cdot dy$
Calificaciones (Y)	53-59	3							2 18	2	6	18	18
	46-52	2							1 6	1	2	4	6
	39-45	1	2 -6	2 -4		1 0	1 1	1 2	2 6	9	9	9	-1
	32-38	0		2 0	1 0	3 0	3 0	3 0		12	0	0	0
	25-31	-1	1 3	1 2	4 4	3 0	1 -1			10	-10	10	8
	18-24	-2	2 12	4 16	2 4	2 0	1 -2			11	-22	44	30
	11-17	-3			3 9	2 0				5	-15	45	9
	fx		5	9	10	11	6	4	5	50	-30	130	70
	$fx \cdot dx$		-15	-18	-10	0	6	8	15	-14			
	$fx \cdot dx^2$		45	36	10	0	6	16	45	158			
	$f \cdot dx \cdot dy$		9	14	17	0	-2	2	30	70			

Nota:
Los números de las esquinas de cada celda en la anterior tabla representan el producto $f \cdot dx \cdot dy$, así por ejemplo, para obtener el número el número -6 de los intervalos 76-79 en X y 39-45 en Y se obtiene multiplicando 2·(-3)·1 = -6. Para obtener el número 18 de los intervalos 100-103 en X y 53-59 en Y se obtiene multiplicando 2·3·3 = 18

-Los números de la última columna (18, 6, -1, 0, 8, 30 y 9) se obtienen sumando los números de las esquinas en cada fila, así por ejemplo, para obtener el número -1 se suma (-6) + (-4) + 0 + 1 +2 + 6 = -1
-Los números de la última fila (9, 14, 17, 0, -2, 2 y 30) se obtienen sumando los números de las esquinas en cada columna, así por ejemplo, para obtener el número 9 se suma (-6) + 3 + 12 = 9.
-Para obtener el número -30 de la antepenúlmina columna se obtiene sumando los resultados de $fy \cdot dy$, es decir, representa la $\sum fy \cdot dy$
-Para obtener el número -14 de la antepenúlmina fila se obtiene sumando los resultados de $fx \cdot dx$, es decir, representa la $\sum fx \cdot dx$
-Para obtener el número 130 de la penúltima columna se obtiene sumando los resultados de $fy \cdot dy^2$, es decir, representa $\sum fy \cdot dy^2$
-Para obtener el número 158 de la penúltima fila se obtiene sumando los resultados de $fx \cdot dx^2$, es decir, representa $\sum fx \cdot dx^2$
-Para obtener último número 70 de la última columna se obtiene sumando los resultados de la última columna 18 +6 +(-1) +0 + 8 + 30 +9 = 70, es decir, representa $\sum fx \cdot dx \cdot dy$
-Para obtener último número 70 de la última fila se obtiene sumando los resultados de la última fila 9 + 14 + 17 + 0 +(-2) +2 + 30 = 70 , es decir, representa $\sum fx \cdot dx \cdot dy$. Por lo tanto tiene que ser igual al último número de la última columna como comprobación que los cálculos de la tabla han sido correctos.

Observando los datos en la tabla anterior se remplaza los valores en la ecuación del Coeficiente de Correlación de Pearson para datos agrupados, obteniéndose:

$$r = \frac{n \cdot \sum f \cdot dx \cdot dy - (\sum fx \cdot dx)\,(\sum fy \cdot dy)}{\sqrt{[n \cdot \sum fx \cdot dx^2 - (\sum fx \cdot dx)^2][n \cdot \sum fy \cdot dy^2 - (\sum fy \cdot dy)^2]}}$$

$$r = \frac{50 \cdot 70 - (-14)(-30)}{\sqrt{[50 \cdot 158 - (-14)^2][50 \cdot 130 - (-30)^2]}} = \frac{3500 - 420}{\sqrt{[7900 - 196][6500 - 900]}} = \frac{3080}{\sqrt{[7704][5600]}}$$

$$r = \frac{3080}{\sqrt{43142400}} = \frac{3080}{6568{,}287448} = 0{,}469$$

Existe una correlación positiva moderada

COEFICIENTE DE CORRELACIÓN POR RANGOS DE SPEARMAN

Este coeficiente se emplea cuando una o ambas escalas de medidas de las variables son ordinales, es decir, cuando una o ambas escalas de medida son posiciones. Ejemplo: Orden de llegada en una carrera y peso de los atletas. Se calcula aplicando la siguiente ecuación:

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

r_s = Coeficiente de correlación por rangos de Spearman;d = Diferencia entre los rangos (X menos Y)
n = número de datos

Ejemplo ilustrativo N° 1: La siguiente tabla muestra el rango u orden obtenido en la primera evaluación (X) y el rango o puesto obtenido en la segunda evaluación (Y) de 8 estudiantes universitarios en la asignatura de Estadística. Calcular el coeficiente de correlación por rangos de Spearman.

Estudiante	X	Y
Dyanita	1	3
Elizabeth	2	4
Mario	3	1
Orlando	4	5
Mathías	5	6
Josué	6	2
Emily	7	8
Monserrath	8	7

Para calcular el coeficiente de correlación por rangos de Spearman de se llena la siguiente tabla:

Estudiante	X	Y	$d = X - Y$	$d^2 = (X - Y)^2$
Dyanita	1	3	-2	4
Elizabeth	2	4	-2	4
Mario	3	1	2	4
Orlando	4	5	-1	1
Mathías	5	6	-1	1
Josué	6	2	4	16
Emily	7	8	-1	1
Monserrath	8	7	1	1
				$\Sigma d^2 = 32$

Se aplica la fórmula:

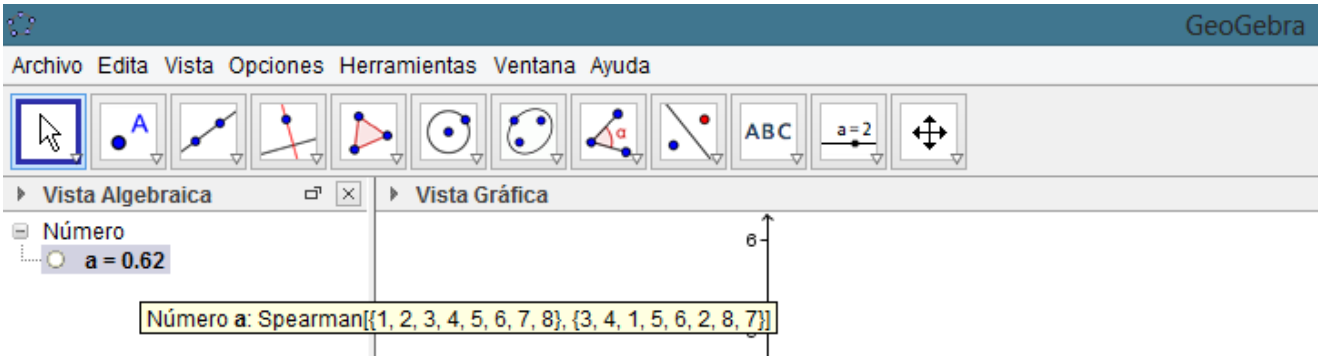
$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \cdot 32}{8(8^2 - 1)} = 1 - \frac{192}{504} = \frac{504 - 192}{504} = \frac{312}{504} = 0{,}619$$

Por lo tanto existe una correlación positiva moderada entre la primera y segunda evaluación de los 8 estudiantes.

En Excel:

	A	B	C	D	E	F	G
1	Estudiante	X	Y				
2	Dyanita	1	3				
3	Elizabeth	2	4				
4	Mario	3	1				
5	Orlando	4	5				
6	Mathías	5	6				
7	Josué	6	2				
8	Emily	7	8				
9	Monserrat	8	7				
10				0,619	=COEF.DE.CORREL(B2:B9;C2:C9)		

En GeoGebra:



Ejemplo ilustrativo N° 2: La siguiente tabla muestra las calificaciones de 8 estudiantes universitarios en las asignaturas de Matemática y Estadística. Calcular el coeficiente de correlación por rangos de Spearman.

N°	Estudiante	Matemática	Estadística
1	Dyana	10	8
2	Elizabeth	9	6
3	Mario	8	10
4	Orlando	7	9
5	Mathías	7	8
6	Josué	6	7
7	Emily	6	6
8	Monserrath	4	9

Solución:

Para calcular el coeficiente de correlación por rangos de Spearman se procede a clasificar u ordenar los datos en rangos (X para Matemática y Y para Estadística) tomando en cuenta las siguientes observaciones:

En la asignatura de Matemática se observa:

- Dyana tiene la más alta calificación, ocupando el primer puesto, por lo que su rango es 1
- Elizabeth ocupa el segundo puesto, por lo que su rango es 2
- Mario se encuentra ubicado en el tercer lugar, por lo que su rango es 3
- Orlando y Mathías ocupan el cuarto y quinto puesto, por lo que su rango es la media aritmética de 4 y 5 que da por resultado 4,5
- Josué y Emily ocupan el sexto y séptimo lugar, por lo que su rango es la media aritmética de 6 y 7 que da por resultado 6,5
- Monserrath se encuentra ubicada en el octavo lugar, por lo que su rango es 8

En la asignatura de Estadística se observa:

- Mario tiene la más alta calificación, ocupando el primer puesto, por lo que su rango es 1
- Orlando y Monserrath ocupan el segundo y tercer puesto, por lo que su rango es la media aritmética de 2 y 3 que da por resultado 2,5

- Dyana y Mathías ocupan el cuarto y quinto puesto, por lo que su rango es la media aritmética de 4 y 5 que da por resultado 4,5
- Josué se encuentra ubicado en el sexto lugar, por lo que su rango es 6
- Elizabeth y Emily ocupan el séptimo y octavo lugar, por lo que su rango es la media aritmética de 7 y 8 que da por resultado 7,5

Los rangos X y Y se presentan en la siguiente tabla:

N°	Estudiante	Matemática	Estadística	X	Y
1	Dyana	10	8	1	4,5
2	Elizabeth	9	6	2	7,5
3	Mario	8	10	3	1
4	Orlando	7	9	4,5	2,5
5	Mathías	7	8	4,5	4,5
6	Josué	6	7	6,5	6
7	Emily	6	6	6,5	7,5
8	Monserrath	4	9	8	2,5

Calculando d, d^2 y Σd^2 se obtiene los siguientes resultados:

N°	Estudiante	Matemática	Estadística	X	Y	$d = X - Y$	$d^2 = (X - Y)^2$
1	Dyana	10	8	1	4,5	-3,5	12,25
2	Elizabeth	9	6	2	7,5	-5,5	30,25
3	Mario	8	10	3	1	2	4
4	Orlando	7	9	4,5	2,5	2	4
5	Mathías	7	8	4,5	4,5	0	0
6	Josué	6	7	6,5	6	0,5	0,25
7	Emily	6	6	6,5	7,5	-1	1
8	Monserrath	4	9	8	2,5	5,5	30,25
							$\Sigma d^2 = 82$

Aplicando la fórmula se obtiene:

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \cdot 82}{8(8^2-1)} = 1 - \frac{492}{504} = \frac{504 - 492}{504} = \frac{12}{504} = 0{,}024$$

COEFICIENTE DE DETERMINACIÓN

Revela qué porcentaje del cambio en Y se explica por un cambio en X. Se calcula elevando al cuadrado el coeficiente de correlación.

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

$x = X - \bar{X}$; $y = Y - \bar{Y}$; r = Coeficiente de correlación de Pearson; r^2 = Coeficiente de determinación

La ecuación del coeficiente producto-momento (Coeficiente de Pearson) $r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$ puede escribirse en la forma equivalente:

$$Coeficiente\ de\ Pearson = r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}}$$

De donde coeficiente de determinación = $r^2 = (\text{Coeficiente de Pearson})^2$

Ejemplo ilustrativo: Con los datos de la siguiente tabla sobre las temperaturas, calcular el coeficiente de determinación empleando la ecuación obtenida de la forma equivalente del coeficiente de Pearson.

X	18	17	15	16	14	12	9	15	16	14	16	18
Y	13	15	14	13	9	10	8	13	12	13	10	8

Solución:
Se calcula el coeficiente de Pearson llenando la siguiente tabla:

X	Y	XY	X^2	Y^2
18	13	234	324	169
17	15	255	289	225
15	14	210	225	196
16	13	208	256	169
14	9	126	196	81
12	10	120	144	100
9	8	72	81	64
15	13	195	225	169
16	12	192	256	144
14	13	182	196	169
16	10	160	256	100
18	8	144	324	64
$\Sigma X = 180$	$\Sigma Y = 138$	$\Sigma XY = 2098$	$\Sigma X^2 = 2772$	$\Sigma Y^2 = 1650$

Se aplica la ecuación para calcular el coeficiente de Pearson.

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}} = \frac{12 \cdot 2098 - 180 \cdot 138}{\sqrt{[12 \cdot 2772 - (180)^2][12 \cdot 1650 - (138)^2]}}$$

$$r = \frac{25176 - 24840}{\sqrt{[33264 - 32400][19800 - 19044]}} = \frac{336}{\sqrt{[864][756]}} = \frac{336}{\sqrt{653184}} = \frac{336}{808{,}198} = 0{,}4157$$

Elevando al cuadrado coeficiente de Pearson queda calculado el coeficiente de determinación.

Coeficiente de determinación = $r^2 = (0{,}4157)^2 = 0{,}1728$

Esto establece que 17,28% del cambio en Y se explica mediante un cambio en X.

Nota: El r^2 tiene significado sólo para las relaciones lineales. Dos variables pueden tener $r^2 = 0$ y sin embargo estar relacionadas en sentido curvilíneo. El valor de r^2 no se interpreta como si la variable Y fuera causado por un cambio de la variable X, ya que la correlación no significa causa.

ANÁLISIS DE REGRESIÓN

La regresión examina la relación entre dos variables, pero restringiendo una de ellas con el objeto de estudiar las variaciones de una variable cuando la otra permanece constante. En otras palabras, la regresión es un método que se emplea para predecir el valor de una variable en función de valores dados a la otra variable.

a) LA RECTA DE LOS MÍNIMOS CUADRADOS

Se llama línea de mejor ajuste y se define como la línea que hace mínima la suma de los cuadrados de las desviaciones respecto a ella de todos los puntos que corresponden a la información recogida.

La recta de los mínimos cuadrados que aproxima el conjunto de puntos $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \ldots \ldots (X_N, Y_N)$ tomando en cuenta a Y como variable dependiente tiene por ecuación

$$Y = a_0 + a_1 X$$

A esta ecuación suele llamarse recta de regresión de Y sobre X, y se usa para estimar los valores de Y para valores dados de X

Si a la recta de regresión $Y = a_0 + a_1X$ se le suma en ambos lados $\sum Y = \sum(a_0 + a_1X)$ se obtiene $\sum Y = a_0N + a_1 \sum X$

Si a la recta de regresión $Y = a_0 + a_1X$ se multiplica por X a ambos lados y luego se suma $\sum XY = \sum X(a_0 + a_1X)$ se obtiene $\sum XY = a_0 \sum X + a_1 \sum X^2$

Las constantes a_0 y a_1 quedan fijadas al resolver simultáneamente las ecuaciones anteriormente encontradas, es decir, al resolver el siguiente sistema de ecuaciones:

$$\begin{cases} \Sigma Y = a_0N + a_1\Sigma X \\ \Sigma XY = a_0\Sigma X + a_1\Sigma X^2 \end{cases}$$

Que se llaman las ecuaciones normales para la recta de mínimos cuadrados.

Las constantes a_0 y a_1 de las anteriores ecuaciones también se pueden calcular empleando las siguientes fórmulas:

$$a_0 = \frac{\sum Y \cdot \sum X^2 - \sum X \cdot \sum XY}{N\sum X^2 - (\sum X)^2} \qquad a_1 = \frac{N\sum XY - \sum X \cdot \sum Y}{N\sum X^2 - (\sum X)^2}$$

Otra ecuación para los mínimos cuadrados para $x = X - \bar{X}$, $y = Y - \bar{Y}$ de la recta de regresión de Y sobre X es:

$$y = \left(\frac{\sum xy}{\sum \mathrm{x}^2}\right)x$$

La recta de los mínimos cuadrados que aproxima el conjunto de puntos $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \ldots \ldots (X_N, Y_N)$ tomando en cuenta a X como variable dependiente tiene por ecuación:

$$X = b_0 + b_1Y$$

A esta ecuación suele llamarse recta de regresión de X sobre Y, y se usa para estimar los valores de X para valores dados de Y. Las constantes b_0 y b_1 quedan fijadas al resolver el siguiente sistema de ecuaciones:

$$\begin{cases} \Sigma X = b_0N + b_1\Sigma Y \\ \Sigma XY = b_0\Sigma Y + b_1\Sigma Y^2 \end{cases}$$

Las constantes b_0 y b_1 del sistema de ecuaciones anterior se pueden calcular empleando las siguientes fórmulas:

$$b_0 = \frac{\sum X \cdot \sum Y^2 - \sum Y \cdot \sum XY}{N\sum Y^2 - (\sum Y)^2} \qquad b_1 = \frac{N\sum XY - \sum X \cdot \sum Y}{N\sum Y^2 - (\sum Y)^2}$$

Otra ecuación para los mínimos cuadrados para $x = X - \bar{X}$, $y = Y - \bar{Y}$ es:

$$x = \left(\frac{\sum xy}{\sum \mathrm{y}^2}\right)y$$

El punto de intersección entre las rectas $Y = a_0 + a_1X$ con $X = b_0 + b_1Y$ se simboliza $(\bar{X}, \bar{Y})$ y se llama centroide o centro de gravedad

Ejemplo ilustrativo: Con los datos de la siguiente tabla sobre la altura en centímetros (X) y los pesos en kilogramos (Y) de una muestra de 8 estudiantes varones tomada al azar del segundo semestre de una universidad.

X	152	157	162	167	173	178	182	188
Y	56	61	67	72	70	72	83	92

1) Ajustar la recta de mínimos cuadrados para Y como variable dependiente resolviendo el sistema:

$$\begin{cases} \Sigma Y = a_0 N + a_1 \Sigma X \\ \Sigma XY = a_0 \Sigma X + a_1 \Sigma X^2 \end{cases}$$

2) Ajustar la recta de mínimos cuadrados para Y como variable dependiente empleando las fórmulas:

$$a_0 = \frac{\sum Y \cdot \sum X^2 - \sum X \cdot \sum XY}{N \sum X^2 - (\sum X)^2} \qquad a_1 = \frac{N \sum XY - \sum X \cdot \sum Y}{N \sum X^2 - (\sum X)^2}$$

3) Ajustar la recta de mínimos cuadrados para Y como variable dependiente empleando la fórmula:

$$y = \left(\frac{\sum xy}{\sum x^2}\right) x$$

4) Ajustar la recta de mínimos cuadrados para X como variable dependiente resolviendo el sistema:

$$\begin{cases} \Sigma X = b_0 N + b_1 \Sigma Y \\ \Sigma XY = b_0 \Sigma Y + b_1 \Sigma Y^2 \end{cases}$$

5) Calcular el punto centroide.

6) Elaborar el diagrama de dispersión. Y en el mismo diagrama graficar las dos rectas de mínimos cuadrados obtenidas en los pasos anteriores.

7) Estimar el valor de Y cuando X = 200 en el diagrama de dispersión de Y como variable dependiente.

8) Estimar el valor de X cuando Y= 100 en el diagrama de dispersión X como variable dependiente.

Solución: Se llena la siguiente tabla:

X	Y	XY	X^2	Y^2
152	56	8512	23104	3136
157	61	9577	24649	3721
162	67	10854	26244	4489
167	72	12024	27889	5184
173	70	12110	29929	4900
178	72	12816	31684	5184
182	83	15106	33124	6889
188	92	17296	35344	8464
$\Sigma X = 1359$	$\Sigma Y = 573$	$\Sigma XY = 98295$	$\Sigma X^2 = 231967$	$\Sigma Y^2 = 41967$

1) Remplazando valores en el sistema se tiene:

$$\begin{cases} \Sigma Y = a_0 N + a_1 \Sigma X \\ \Sigma XY = a_0 \Sigma X + a_1 \Sigma X^2 \end{cases} \Rightarrow \begin{matrix} 573 = a_0 \cdot 8 + a_1 \cdot 1359 \\ 98295 = a_0 \cdot 1359 + a_1 \cdot 231967 \end{matrix} \Rightarrow \begin{cases} 8a_0 + 1359a_1 = 573 \\ 1359a_0 + 231967a_1 = 98295 \end{cases}$$

Resolviendo el sistema por determinantes (regla de Cramer) se obtiene:

$$a_0 = \frac{\Delta a_0}{\Delta} = \frac{\begin{vmatrix} 573 & 1359 \\ 98295 & 231967 \end{vmatrix}}{\begin{vmatrix} 8 & 1359 \\ 1359 & 231967 \end{vmatrix}} = \frac{573 \cdot 231967 - 98295 \cdot 1359}{8 \cdot 231967 - 1359 \cdot 1359} = \frac{-665814}{8855} = -75{,}191$$

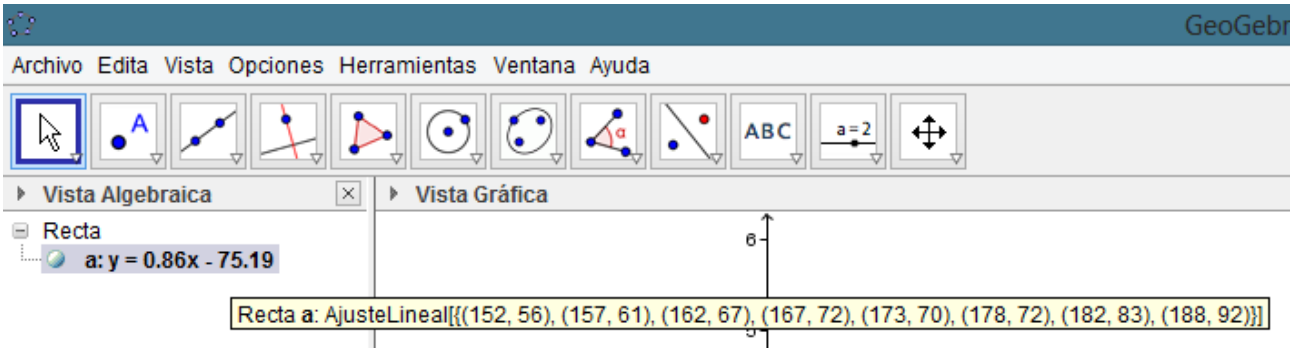
$$a_1 = \frac{\Delta a_1}{\Delta} = \frac{\begin{vmatrix} 8 & 573 \\ 1359 & 98295 \end{vmatrix}}{8855} = \frac{8 \cdot 98295 - 1359 \cdot 573}{8855} = \frac{7653}{8855} = 0{,}864$$

Para calcular los valores de a_1 y a_0 en Excel se calcula de la siguiente manera:

B11 | {=ESTIMACION.LINEAL(B2:B9;A2:A9)}

	A	B	C	D	E
1	X	Y			
2	152	56			
3	157	61			
4	162	67			
5	167	72			
6	173	70			
7	178	72			
8	182	83			
9	188	92			
10					
11		0,8642575	-75,19074		

Los cálculos en GeoGebra se muestran en la siguiente figura:



Remplazando valores en la ecuación respectiva se obtiene:
$Y = a_0 + a_1 X \Rightarrow Y = -75{,}191 + 0{,}864X$

Interpretación:
- El valor $a_1 = 0{,}864$ indica que la recta tiene una pendiente positiva aumentando a razón de 0,864
- El valor de $a_0 = -75{,}191$ indica el punto en donde la recta interseca al eje Y cuanto X = 0

2) Con los datos de la tabla anterior se substituye valores en las siguientes ecuaciones:

$$a_0 = \frac{\sum Y \cdot \sum X^2 - \sum X \cdot \sum XY}{N\sum X^2 - (\sum X)^2} = \frac{573 \cdot 231967 - 1359 \cdot 98295}{8 \cdot 231967 - (1359)^2} = \frac{-665814}{8855} - 75{,}191$$

$$a_1 = \frac{N\sum XY - \sum X \cdot \sum Y}{N\sum X^2 - (\sum X)^2} = \frac{8 \cdot 98295 - 1359 \cdot 573}{8 \cdot 231967 - (1359)^2} = \frac{7653}{8855} = 0{,}864$$

Remplazando valores en la ecuación respectiva se obtiene:
$Y = a_0 + a_1 X \Rightarrow Y = -75{,}191 + 0{,}864X$

3) Se calcula las medias aritméticas de X y Y para llenar la siguiente tabla:

$$\bar{X} = \frac{1359}{8} = 169{,}875 \; ; \bar{Y} = \frac{573}{8} = 71{,}625$$

X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	xy	x^2	y^2
152	56	-17,88	-15,625	279,297	319,516	244,141
157	61	-12,88	-10,625	136,797	165,766	112,891
162	67	-7,875	-4,625	36,422	62,016	21,391
167	72	-2,875	0,375	-1,078	8,266	0,141
173	70	3,125	-1,625	-5,078	9,766	2,641
178	72	8,125	0,375	3,047	66,016	0,141
182	83	12,125	11,375	137,922	147,016	129,391
188	92	18,125	20,375	369,297	328,516	415,141
$\Sigma X = 1359$	$\Sigma Y = 573$			$\Sigma xy = 956,625$	$\Sigma x^2 = 1106,875$	$\Sigma y^2 = 925,875$

Remplazando valores en la fórmula respectiva se obtiene:

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x \Rightarrow y = \frac{956,625}{1106,875}x \Rightarrow Y - \bar{Y} = \frac{956,625}{1106,875}(X - \bar{X})$$

$$Y - 71,625 = \frac{956,625}{1106,875}(X - 169,875) \Rightarrow 1106,875(Y - 71,625) = 956,625(X - 169,875)$$

$$1106,875Y - 79280,20838 = 956,625X - 162510,4984$$

$$1106,875Y = 956,625X - 162510,4984 + 79280,20838$$

$$1106,875Y = 956,625X - 83230,29$$

$$Y = \frac{956,625X - 83230,29}{1106,875} \Rightarrow Y = \frac{956,625X}{1106,875} - \frac{83230,29}{1106,875} \Rightarrow Y = 0,864X - 75,19$$

$$Y = -75,19 + 0,864X$$

4) Remplazando valores en sistema respectivo se obtiene:

$$\begin{cases} \Sigma X = b_0 N + b_1 \Sigma Y \\ \Sigma XY = b_0 \Sigma Y + b_1 \Sigma Y^2 \end{cases} \Rightarrow \begin{matrix} 1359 = b_0 \cdot 8 + b_1 \cdot 573 \\ 98295 = b_0 \cdot 573 + b_1 \cdot 41967 \end{matrix} \Rightarrow \begin{matrix} 8b_0 + 573b_1 = 1359 \\ 573b_0 + 41967b_1 = 98295 \end{matrix}$$

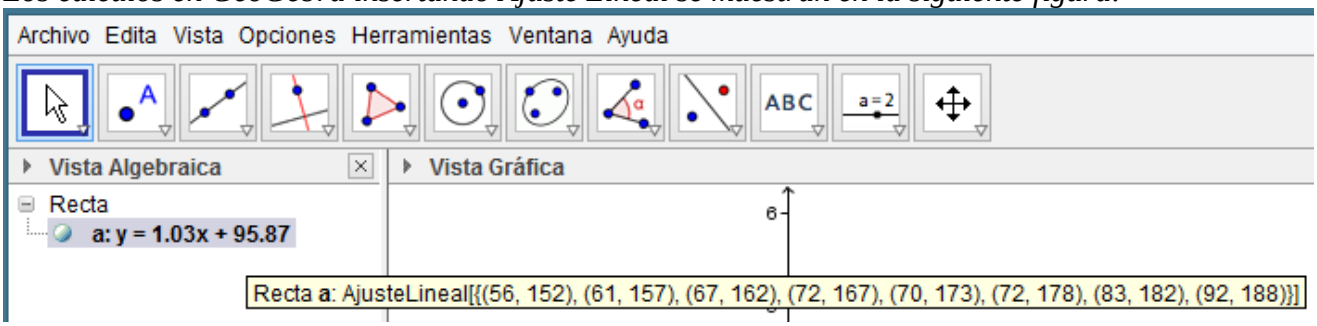
Resolviendo el sistema se obtiene:

$b_0 = 95,871;\ b_1 = 1,033$

Remplazando valores en la ecuación de la recta de mínimos cuadrados se obtiene:

$X = b_0 + b_1 Y \Longrightarrow X = 95,871 + 1,033Y$

Los cálculos en GeoGebra insertando Ajuste Lineal se muestran en la siguiente figura:



Interpretación:

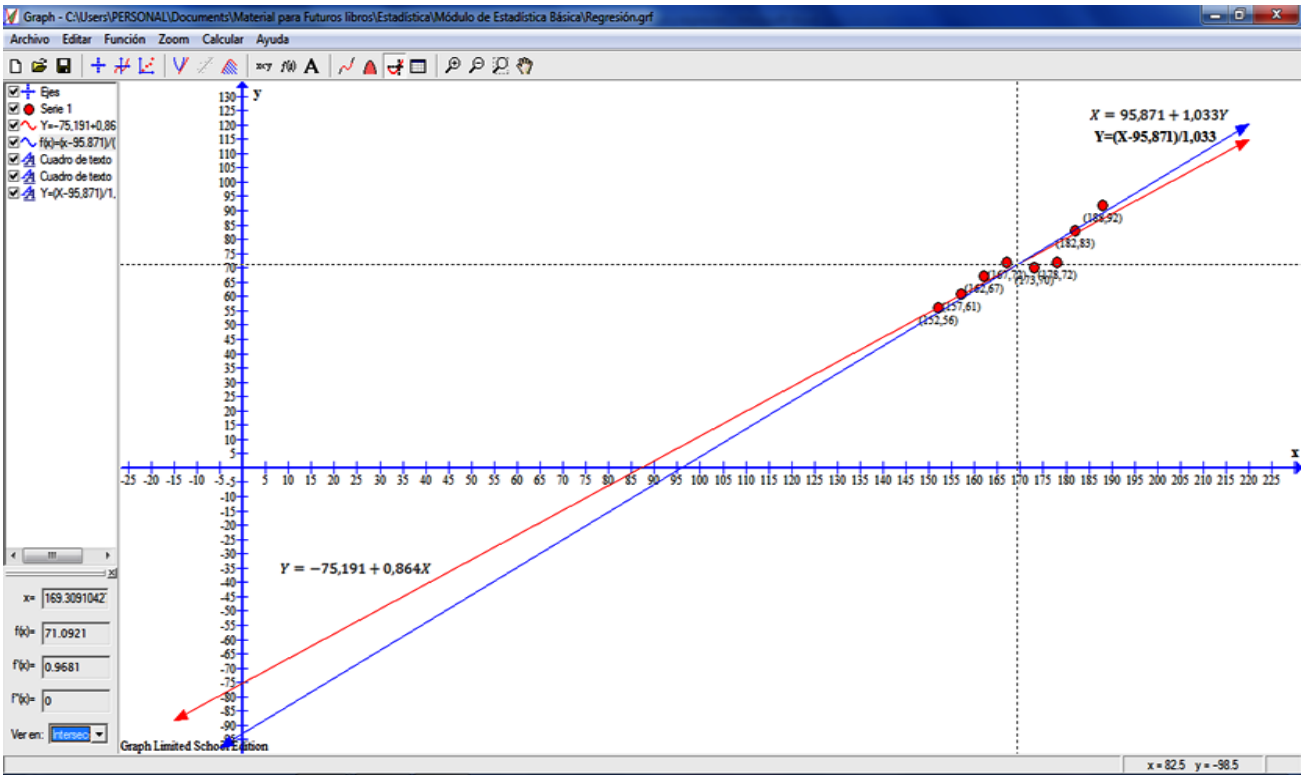
- El valor $b_1 = 1,033$ indica que la recta tiene una pendiente positiva aumentando a razón de 1,033
- El valor de $b_0 = 95,871$ indica el punto en donde la recta interseca al eje X cuanto Y = 0

5) Para calcular el centroide $(\bar{X}, \bar{Y})$ se resuelve el sistema formado por las dos rectas de los mínimos cuadrados en donde X es $\bar{X}$ y Y es $\bar{Y}$.

$$\begin{cases} Y = -75,191 + 0,864X \\ X = 95,871 + 1,033Y \end{cases}$$

Al resolver el sistema se obtiene el centroide: X = 169,3 y Y = 71,092

6) *Empleando el programa Graph se obtiene la siguiente figura:*



7) Remplazando X = 200 en la ecuación solicitada se obtiene:
$Y = -75{,}191 + 0{,}864X = -75{,}191 + 0{,}864 \cdot 200 = -75{,}191 + 172{,}8 = 97{,}609$

8) Remplazando Y = 100 en la ecuación solicitada se obtiene:
$X = 95{,}871 + 1{,}033Y = X = 95{,}871 + 1{,}033 \cdot 100 = X = 95{,}871 + 103{,}3 = 199{,}171$

b) LA PARÁBOLA DE LOS MÍNIMOS CUADRADOS

La parábola de mínimos cuadrados que aproxima el conjunto de puntos $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \ldots (Y_N, Y_N)$ tiene ecuación dada por $Y = a_0 + a_1X + a_2X^2$, donde las constantes a_0, a_1 y a_2 se determinan al resolver simultáneamente el sistema de ecuaciones que se forma al multiplicar la ecuación $Y = a_0 + a_1X + a_2X^2$ por $1, X, Y$ sucesivamente, y sumando después.

$$\begin{cases} \Sigma Y = a_0N + a_1\Sigma X + a_2\Sigma X^2 \\ \Sigma XY = a_0\Sigma X + a_1\Sigma X^2 + a_2\Sigma X^3 \\ \Sigma X^2Y = a_0\Sigma X^2 + a_1\Sigma X^3 + a_2\Sigma X^4 \end{cases}$$

Ejemplo ilustrativo: La siguiente tabla muestra la población de un país en los años 1960-2010 en intervalos de 5 años.

Año	1960	1965	1970	1975	1980	1985	1990	1995	2000	2005	2010
Población (millones)	4,52	5,18	6,25	7,42	8,16	9,12	10,92	11,62	12,68	13,12	13,97

1) Ajustar una parábola de mínimos cuadrados de la forma $Y = a_0 + a_1X + a_2X^2$
2) Calcular los valores de tendencia para los años dados.
3) Estimar la población para los años 2015 y 2020.
4) Elaborar un diagrama de dispersión, y en el mismo diagrama graficar la parábola de los mínimos cuadrados.

Nota: Se recomienda codificar o cambiar la numeración de los años, tratando que X = 0 esté ubicado en lo posible en el centro.

Solución: Para ajustar una parábola de mínimos cuadrados se llena la siguiente tabla:

Año	X	Y	X^2	X^3	X^4	XY	X^2Y
1960	-5	4,52	25	-125	625	-22,6	113
1965	-4	5,18	16	-64	256	-20,72	82,88
1970	-3	6,25	9	-27	81	-18,75	56,25
1975	-2	7,42	4	-8	16	-14,84	29,68
1980	-1	8,16	1	-1	1	-8,16	8,16
1985	0	9,12	0	0	0	0	0
1990	1	10,92	1	1	1	10,92	10,92
1995	2	11,62	4	8	16	23,24	46,48
2000	3	12,68	9	27	81	38,04	114,12
2005	4	13,12	16	64	256	52,48	209,92
2010	5	13,97	25	125	625	69,85	349,25
Σ	0	102,96	110	0	1958	109,46	1020,66

Se remplaza valores en el sistema y se obtiene:

$$\begin{cases} \Sigma Y = a_0 N + a_1 \Sigma X + a_2 \Sigma X^2 \\ \Sigma XY = a_0 \Sigma X + a_1 \Sigma X^2 + a_2 \Sigma X^3 \\ \Sigma X^2 Y = a_0 \Sigma X^2 + a_1 \Sigma X^3 + a_2 \Sigma X^4 \end{cases}$$

$$\begin{cases} 102{,}96 = a_0 \cdot 11 + a_1 \cdot 0 + a_2 \cdot 110 \\ 109{,}46 = a_0 \cdot 0 + a_1 \cdot 110 + a_2 \cdot 0 \\ 1020{,}66 = a_0 \cdot 110 + a_1 \cdot 0 + a_2 \cdot 1958 \end{cases} \Rightarrow \begin{cases} 11a_0 + 0a_1 + 110a_2 = 102{,}96 \\ 0a_0 + 110a_1 + 0a_2 = 109{,}46 \\ 110a_0 + 0a_1 + 1958a_2 = 1020{,}66 \end{cases}$$

Resolviendo el sistema empleando determinantes (regla de Cramer) se obtiene:

$$a_o = \frac{\Delta a_0}{\Delta} = \frac{\begin{vmatrix} 102{,}96 & 0 & 110 \\ 109{,}46 & 110 & 0 \\ 1020{,}66 & 0 & 1958 \end{vmatrix}}{\begin{vmatrix} 11 & 0 & 110 \\ 0 & 110 & 0 \\ 110 & 0 & 1958 \end{vmatrix}} = \frac{\begin{vmatrix} 102{,}96 & 0 & 110 \\ 109{,}46 & 110 & 0 \\ 1020{,}66 & 0 & 1958 \\ 102{,}96 & 0 & 110 \\ 109{,}46 & 110 & 0 \end{vmatrix}}{\begin{vmatrix} 11 & 0 & 110 \\ 0 & 110 & 0 \\ 110 & 0 & 1958 \\ 11 & 0 & 110 \\ 0 & 110 & 0 \end{vmatrix}}$$

$$a_0 = \frac{22175524{,}8 + 0 + 0 - 12349986 - 0 - 0}{2369180 + 0 + 0 - 1331000 - 0 - 0} = \frac{9825538{,}8}{1038180} = 9{,}464$$

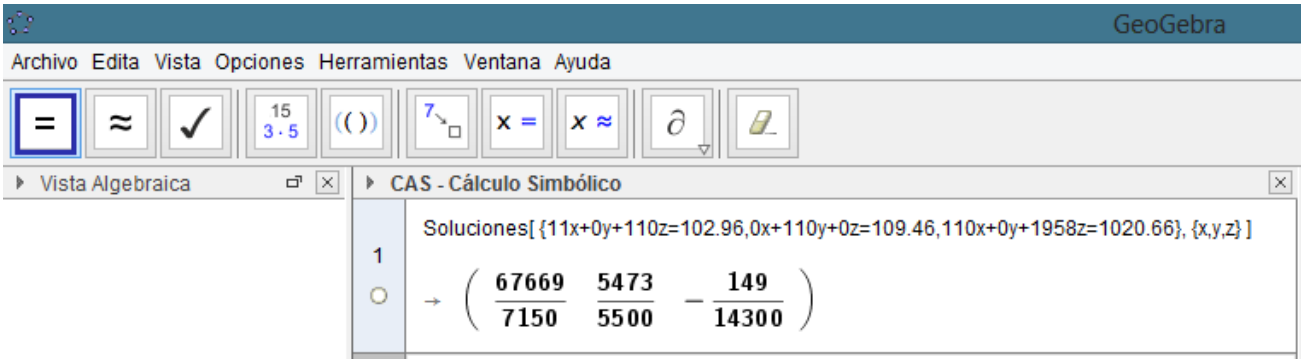
$$a_1 = \frac{\Delta a_1}{\Delta} = \frac{\begin{vmatrix} 11 & 102{,}96 & 110 \\ 0 & 109{,}46 & 0 \\ 110 & 1020{,}66 & 1958 \end{vmatrix}}{1038180} = \frac{\begin{vmatrix} 11 & 102{,}96 & 110 \\ 0 & 109{,}46 & 0 \\ 110 & 1020{,}66 & 1958 \\ 11 & 102{,}96 & 110 \\ 0 & 109{,}46 & 0 \end{vmatrix}}{1038180}$$

$$a_1 = \frac{23577549{,}48 + 0 + 0 - 1324466 - 0 - 0}{1038180} = \frac{2357549{,}48}{1038180} = 0{,}995$$

$$a_2 = \frac{\Delta a_2}{\Delta} = \frac{\begin{vmatrix} 11 & 0 & 102{,}96 \\ 0 & 110 & 109{,}46 \\ 110 & 0 & 1020{,}66 \end{vmatrix}}{1038180} = \frac{\begin{vmatrix} 11 & 0 & 102{,}96 \\ 0 & 110 & 109{,}46 \\ 110 & 0 & 1020{,}66 \\ 11 & 0 & 102{,}96 \\ 0 & 110 & 109{,}46 \end{vmatrix}}{1038180}$$

$$a_2 = \frac{1234998{,}6 + 0 + 0 - 1245816 - 0 - 0}{1038180} = \frac{-10817{,}4}{1038180} = -0{,}01$$

El sistema de ecuaciones resuelto en GeoGebra:



$$\frac{67669}{7150} = 9{,}464\,; \frac{5473}{5500} = 0{,}995\,; -\frac{149}{14300} = -0{,}01$$

Remplazando los valores encontrados se obtiene la ecuación de la parábola de mínimos cuadrados:
$Y = a_0 + a_1 X + a_2 X^2 \Rightarrow Y = 9{,}464 + 0{,}995X - 0{,}01X^2$

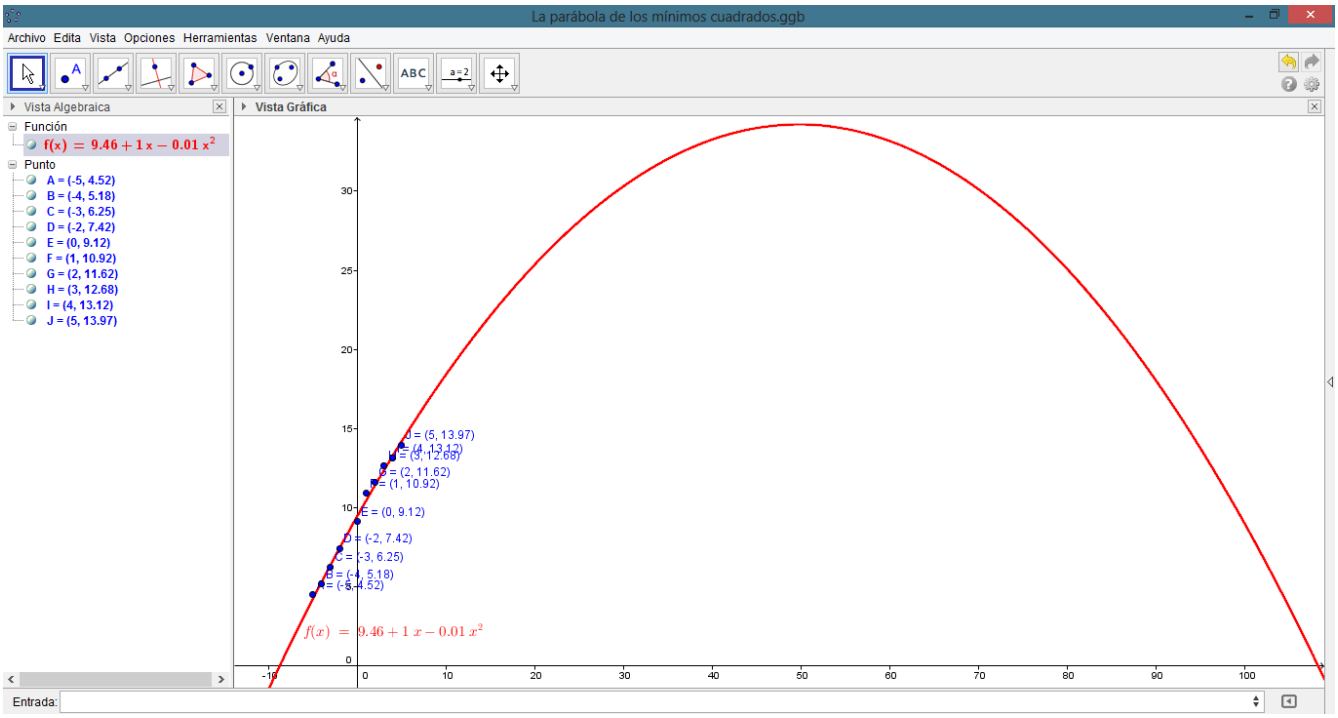
2) Los valores de tendencia se obtienen al remplazar los valores de X en la ecuación de la parábola de mínimos cuadrados, los cuales se presenta en la siguiente tabla:

Año	X	Y	Valores de tendencia $Y = 9{,}464 + 0{,}995X - 0{,}01X^2$
1960	-5	4,52	4,24
1965	-4	5,18	5,32
1970	-3	6,25	6,39
1975	-2	7,42	7,43
1980	-1	8,16	8,46
1985	0	9,12	9,46
1990	1	10,92	10,45
1995	2	11,62	11,41
2000	3	12,68	12,36
2005	4	13,12	13,28
2010	5	13,97	14,19

3) Para estimar la población de los años 2015 y 2020 se transforma estos años a X siguiendo la secuencia de la tabla anterior, siendo X = 6 para el año 2015 y X= 7 para el 2020

Entonces para el 2015 se tiene:
$Y = 9{,}464 + 0{,}995X - 0{,}01X^2 = 9{,}464 + 0{,}995(6) - 0{,}01(6)^2 = 9{,}464 + 5{,}97 - 0{,}36 = 15{,}074$
Para el 2020 se tiene:
$Y = 9{,}464 + 0{,}995X - 0{,}01X^2 = 9{,}464 + 0{,}995(7) - 0{,}01(7)^2 = 9{,}464 + 6{,}965 - 0{,}49 = 15{,}939$

4) *El diagrama de dispersión y la parábola de los mínimos cuadrados en GeoGebra:*



REGRESIÓN EXPONENCIAL

Cuando la curva de regresión de y sobre x es exponencial, es decir para cualquier x considerada, la media de la distribución está dada por la siguiente ecuación predictora:

$$Y = \alpha \cdot \beta^{X}$$

Tomando logaritmos en ambos miembros:

$$logY = log\alpha + X \cdot log\beta$$

Y se puede estimar ahora log Y y log β, y de ahí obtener α y β, aplicando *los métodos de los mínimos cuadrados*.

Donde las constantes α y β quedan fijadas al resolver simultáneamente las ecuaciones:

$$\begin{cases} \Sigma logY = log\alpha \cdot N + log\beta \cdot \Sigma X \\ \Sigma X \cdot \log Y = \log\alpha \cdot \Sigma X + \log\beta \cdot \Sigma X^2 \end{cases}$$

Ejemplo ilustrativo: Las cifras siguientes son datos sobre el porcentaje de llantas radiales producidas por cierto fabricante que aún pueden usarse después de recorrer cierto número de millas:

Miles de Millas recorridas (X)	1	2	5	15	25	30	35	40
Porcentaje útil (Y)	99	95	85	55	30	24	20	15

1) Ajustar una curva exponencial aplicando el método de mínimos cuadrados.
2) Calcular la ecuación predictora.
3) Graficar la ecuación predictora.
4) Estimar qué porcentaje de las llantas radiales del fabricante durarán 50000 millas.

Solución:

1) Se llena la siguiente tabla:

X	Y	$logY$	X^2	$X \cdot logY$
1	99	1,996	1	1,996
2	95	1,978	4	3,955
5	85	1,929	25	9,647
15	55	1,740	225	26,105
25	30	1,477	625	36,928
30	24	1,380	900	41,406
35	20	1,301	1225	45,536
40	15	1,176	1600	47,044
$\Sigma X = 153$		$\Sigma logY = 12{,}97759$	$\Sigma X^2 = 4605$	$\Sigma X \cdot logY = 212{,}61769$

Remplazando valores en el sistema se obtiene:

$$\begin{cases} \Sigma logY = log\alpha \cdot N + log\beta \cdot \Sigma X \\ \Sigma X \cdot \log Y = \log\alpha \cdot \Sigma X + \log\beta \cdot \Sigma X^2 \end{cases}$$

$$\begin{cases} 12{,}97759 = \log\alpha \cdot 8 + \log\beta \cdot 153 \\ 212{,}61769 = \log\alpha \cdot 153 + \log\beta \cdot 4605 \end{cases} \Rightarrow \begin{cases} 8\log\alpha + 153\log\beta = 12{,}97759 \\ 153\log\alpha + 4605\log\beta = 212{,}61769 \end{cases}$$

Al resolver el sistema se obtiene:

$log\alpha = 2{,}027495747; log\beta = -0{,}02119180389$

Remplazando valores se obtiene:

$logY = log\alpha + X \cdot log\beta \Longrightarrow logY = 2{,}027496 - 0{,}02119X$

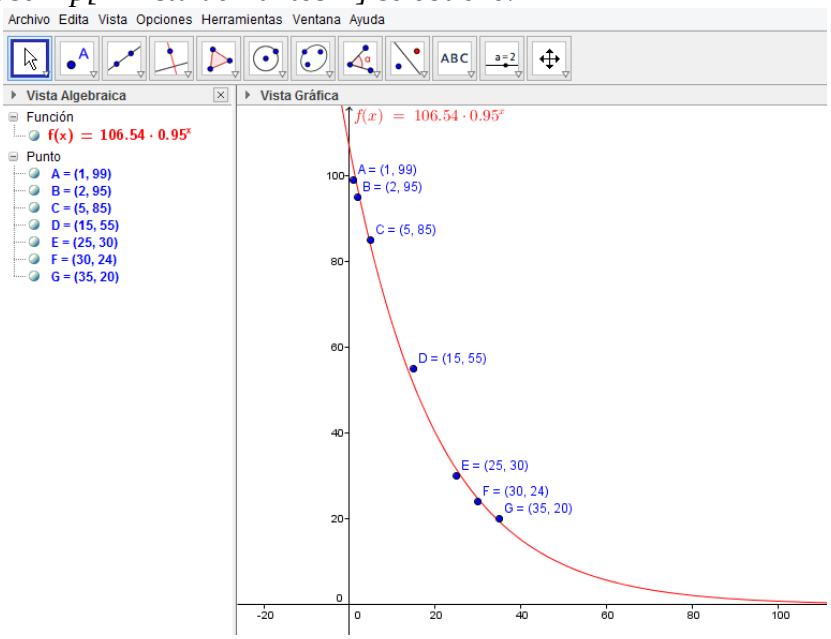
Aplicando el antilogaritmo se obtiene:

$\alpha = \text{anti log } 2{,}027495747 = 106{,}536; \quad \beta = \text{anti log}(-0{,}02119180389) = 0{,}952$

2) Remplazando en la ecuación predictora se obtiene:

$Y = \alpha \cdot \beta^X \Longrightarrow Y = 106{,}536 \cdot 0{,}952^X$

3) Realizando el diagrama de dispersión y los cálculos de la ecuación predictora de GeoGebra insertando AjusteBaseExp[<Lista de Puntos>] se obtiene:



4) La estimación del porcentaje de llantas radiales que durarán 50000 millas se obtiene remplazando en la ecuación predictora el valor de X = 50

$Y = 106{,}536 \cdot 0{,}952^X \Longrightarrow Y = 106{,}536 \cdot 0{,}952^{50} = 9{,}106$

Entonces el porcentaje sería de 9,106%

REGRESIÓN POTENCIAL

La regresión potencial tiene por ecuación predictora:

$$Y = \alpha \cdot X^{\beta}$$

Y la regresión recíproca es:

$$Y = \frac{1}{\alpha + \beta \cdot X}$$

Para el primer caso los valores siguen una ley potencial. Si la ecuación predictora está dada por: $Y = \alpha \cdot X^{\beta}$ tomando logaritmos en ambos miembros, queda:

$$logY = log\alpha + \beta \cdot logX$$

Donde las constantes α y β quedan fijadas al resolver simultáneamente las ecuaciones:

$$\begin{cases} \Sigma logY = log\alpha \cdot N + \beta \cdot \Sigma \text{logX} \\ \Sigma \text{logX} \cdot \text{logY} = \text{log}\alpha \cdot \Sigma \text{logX} + \beta \cdot \Sigma (logX)^2 \end{cases}$$

Para el segundo caso, si la ecuación predictora está dada por $Y = \frac{1}{\alpha+\beta \cdot X}$, entonces invirtiendo, la misma expresión se puede escribir $\frac{1}{Y} = \frac{\alpha+\beta \cdot X}{1}$, o sea:

$$Y = \frac{1}{\alpha + \beta \cdot X} \Rightarrow \frac{1}{Y} = \alpha + \beta \cdot X$$

Donde las constantes α y β quedan fijadas al resolver simultáneamente las ecuaciones:

$$\begin{cases} \Sigma \frac{1}{Y} = \alpha \cdot N + \beta \cdot \Sigma \text{X} \\ \Sigma \text{X} \cdot \frac{1}{Y} = \alpha \cdot \Sigma \text{X} + \beta \cdot \Sigma X^2 \end{cases}$$

Ejemplos ilustrativo N° 1: Sea el siguiente conjunto de valores, las lecturas de un experimento donde X es el volumen (variable independiente) e Y es la presión de una masa dada de gas (variable resultante).

X	1	2	3	4	5	6	7
Y	7	30	90	170	290	450	650

1.1) Ajustar una curva exponencial aplicando el método de mínimos cuadrados.
1.2) Calcular la ecuación predictora.
1.3) Graficar la ecuación predictora.
1.4) Estimar la presión de la masa de gas de volumen 9.

Solución:

1.1) Para ajustar una curva exponencial aplicando el método de mínimos cuadrados:

X	Y	$logX$	$logY$	$logX \cdot logY$	$(logX)^2$
1	7	0,0000	0,8451	0,0000	0,0000
2	30	0,3010	1,4771	0,4447	0,0906
3	90	0,4771	1,9542	0,9324	0,2276
4	170	0,6021	2,2304	1,3429	0,3625
5	290	0,6990	2,4624	1,7211	0,4886
6	450	0,7782	2,6532	2,0646	0,6055
7	650	0,8451	2,8129	2,3772	0,7142
$\Sigma X = 28$		$\Sigma logX = 3{,}7024$	$\Sigma logY = 14{,}4354$	$\Sigma logX \cdot logY = 8{,}8829$	$\Sigma (logX)^2 = 2{,}4890$

Remplazando valores en el sistema de ecuaciones se obtiene:

$$\begin{cases} \Sigma logY = log\alpha \cdot N + \beta \cdot \Sigma \text{logX} \\ \Sigma \text{logX} \cdot \text{logY} = \text{log}\alpha \cdot \Sigma \text{logX} + \beta \cdot \Sigma (logX)^2 \end{cases}$$

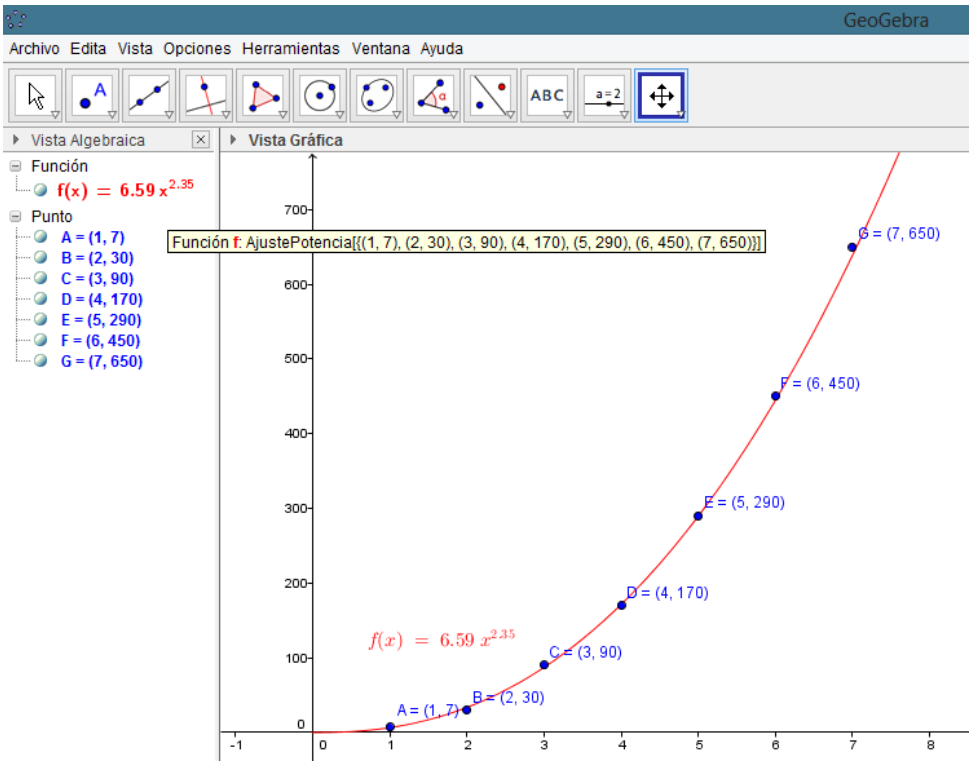
$$\begin{cases} 14{,}4354 = log\alpha \cdot 7 + \beta \cdot 3{,}7024 \\ 8{,}8829 = log\alpha \cdot 3{,}7024 + \beta \cdot 2{,}4890 \end{cases} \Rightarrow \begin{cases} 7log\alpha + 3{,}7024\beta = 14{,}4354 \\ 3{,}7024log\alpha + 2{,}4890\beta = 8{,}8829 \end{cases}$$

Al resolver el sistema se obtiene: log α = 0,819; β = 2,351

Remplazando valores en la ecuación predictora expresada en logaritmos se tiene:
$logY = log\alpha + \beta \cdot logX \Longrightarrow logY = 0{,}819 + 2{,}351 \cdot logX$

1.2) Para calcular la ecuación predictora, primero se calcula el valor de α de la siguiente manera:
$log\alpha = 0{,}819 \Rightarrow \alpha = antilog\ 0{,}819 = 6{,}592$
Remplazando en la ecuación predictora se obtiene:
$Y = \alpha \cdot X^{\beta} \Longrightarrow Y = 6{,}592 \cdot X^{2{,}351}$

1.3) Realizando el diagrama de dispersión y calculando la ecuación predictora en GeoGebra:



1.4) Para estimar la presión de la masa de gas de volumen 9 se reemplaza el valor X = 9 en la ecuación predictora
$Y = 6{,}592 \cdot X^{2{,}351} \Longrightarrow Y = 6{,}592 \cdot 9^{2{,}351} = 1154{,}63$

Ejemplo ilustrativo N° 2: Sea el siguiente conjunto de valores, las lecturas de un experimento donde X es la variable independiente e Y la variable resultante.

X	1	2	3	4	5	6	7
Y	1,4	1	0,9	0,7	0,6	0,55	0,5

2.1) Calcular las constantes α y β, aplicando el método de mínimos cuadrados.
2.2) Calcular la ecuación predictora.
2.3) Graficar la ecuación predictora.
2.4) Estimar el valor de Y para X = 9

Solución:

2.1) Para calcular las constantes α y β, aplicando el método de mínimos cuadrados se llena la siguiente tabla:

X	Y	$1/Y$	$X(1/Y)$	X^2
1	1,4	0,7143	0,7143	1
2	1	1,0000	2,0000	4
3	0,9	1,1111	3,3333	9
4	0,7	1,4286	5,7143	16
5	0,6	1,6667	8,3333	25
6	0,55	1,8182	10,9091	36
7	0,5	2,0000	14,0000	49
$\Sigma X = 28$		$\Sigma\,(1/Y) = 9{,}7388$	$\Sigma\,X(1/Y) = 45{,}0043$	$\Sigma X^2 = 140$

Remplazando valores en el siguiente sistema se obtiene:

$$\begin{cases} \Sigma\dfrac{1}{Y} = \alpha \cdot N + \beta \cdot \Sigma X \\ \Sigma X \cdot \dfrac{1}{Y} = \alpha \cdot \Sigma X + \beta \cdot \Sigma X^2 \end{cases} \Rightarrow \begin{cases} 9{,}7388 = \alpha \cdot 7 + \beta \cdot 28 \\ 45{,}0043 = \alpha \cdot 28 + \beta \cdot 140 \end{cases} \Rightarrow \begin{cases} 7\alpha + 28\beta = 9{,}7388 \\ 28\alpha + 140\beta = 45{,}0043 \end{cases}$$

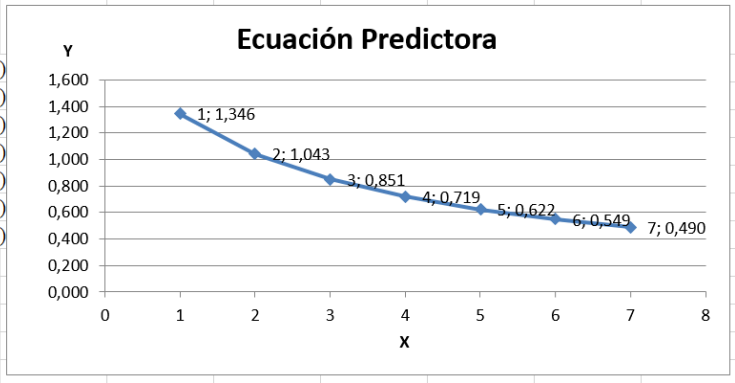
Al resolver el sistema se obtiene:
α = 0,5271; β = 0,2160

2.2) Para calcular la ecuación predictora se remplaza los valores encontrados de α y β, y se obtiene:

$$Y = \frac{1}{\alpha + \beta \cdot X} \Rightarrow Y = \frac{1}{0{,}5271 + 0{,}2160X}$$

2.3) *La gráfica la ecuación predictora elaborada en Excel:*

	A	B	C
1	X	Y = 1/(0,5271+0,2160X)	
2	1	1,346	=1/($B$10+$B$11*A2)
3	2	1,043	=1/($B$10+$B$11*A3)
4	3	0,851	=1/($B$10+$B$11*A4)
5	4	0,719	=1/($B$10+$B$11*A5)
6	5	0,622	=1/($B$10+$B$11*A6)
7	6	0,549	=1/($B$10+$B$11*A7)
8	7	0,490	=1/($B$10+$B$11*A8)
9			
10		0,5271	
11		0,216	
12			
13			



2.4) Para estimar el valor de Y para X = 9 se reemplaza el valor de X en la ecuación predictora.

$$Y = \frac{1}{0{,}5271 + 0{,}2160X} \Rightarrow Y = \frac{1}{0{,}5271 + 0{,}2160 \cdot 9} = 0{,}405$$

ERROR ESTÁNDAR DE ESTIMACIÓN

Es el grado de dispersión de los datos con respecto a la recta de regresión $Y = a_0 + a_1X$
El error estándar de estimación se calcula con la fórmula:

$$s_e = \sqrt{\frac{\sum(Y_i - Y_{est})^2}{N - 2}}$$

Donde:
Y_i = cada valor de Y
Y_{est}= valor estimado de Y a partir de la recta de regresión
N = número de datos
Otras ecuaciones para calcular el error estándar de estimación son:

$$s_e = \sqrt{\frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY}{N-2}} \qquad s_e = \sqrt{\frac{\sum y^2 - a_1 \sum xy}{N-2}}$$

Donde:
a_0 = ordenada en el origen (punto de intersección de la recta con el eje y)
a_1 = pendiente de la recta (tangente del ángulo de inclinación de la recta)
$x = X - \bar{X}$
$y = Y - \bar{Y}$

Ejemplo ilustrativo: Calcular error estándar de estimación empleando las 3 fórmulas dadas, utilizando los datos de la tabla del ejemplo para ajustar la recta de mínimos cuadrados para Y como variable dependiente.

X	152	157	162	167	173	178	182	188
Y	56	61	67	72	70	72	83	92

Solución:
Para comenzar a resolver este ejemplo recordemos que ya se obtuvo los valores respectivos al resolver el ejemplo para ajustar la recta de mínimos cuadrados, los cuales fueron:
$\Sigma X = 1359; \Sigma Y = 573; \Sigma XY = 98295; \Sigma X^2 = 231967; \Sigma Y^2 = 41967; \Sigma xy = 956{,}625$
$\Sigma x^2 = 1106{,}875; \Sigma y^2 = 925{,}875; a_0 = -75{,}191; a_1 = 0{,}864; Y = -75{,}191 + 0{,}864X$

1) Para emplear la primera fórmula se llena la siguiente tabla:

X	Y	$Y_{est} = 75{,}191 + 0{,}86X$	Y_{est}	$(Y - Y_{est})^2$
152	56	-75,191+0,86(152)	55,529	0,222
157	61	-75,191+0,86(157)	59,829	1,371
162	67	-75,191+0,86(162)	64,129	8,243
167	72	-75,191+0,86(167)	68,429	12,752
173	70	-75,191+0,86(173)	73,589	12,881
178	72	-75,191+0,86(178)	77,889	34,680
182	83	-75,191+0,86(182)	81,329	2,792
188	92	-75,191+0,86(188)	86,489	30,371
Σ				103,312

Se remplaza valores en la primera fórmula se obtiene:

$$s_e = \sqrt{\frac{\sum (Y_i - Y_{est})^2}{N-2}} = \sqrt{\frac{103{,}312}{8-2}} = 3{,}842$$

2) Remplazando valores en la segunda fórmula se obtiene:

$$s_e = \sqrt{\frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY}{N-2}}$$

$$s_e = \sqrt{\frac{41967 - (-75{,}191)(573) - 0{,}864(98295)}{8-2}} = \sqrt{\frac{41967 + 43084{,}443 - 84926{,}88}{6}} = 4{,}556$$

3) Remplazando valores en la tercera fórmula se obtiene:

$$s_e = \sqrt{\frac{\sum y^2 - a_1 \sum xy}{N-2}} = \sqrt{\frac{925{,}875 - 0{,}864(956{,}625)}{8-2}} = \sqrt{\frac{99{,}351}{6}} = 4{,}069$$

Empleando exclusivamente Excel para calcular el error estándar de estimación se procede de la siguiente manera:

Se inserta la función ERROR.TIPICO.XY. Se selecciona las celdas respectivas. Pulsar en Aceptar.

	A	B	C	D
1	X	Y		
2	152	56		
3	157	61		
4	162	67		
5	167	72		
6	173	70		
7	178	72		
8	182	83		
9	188	92		
10	4,06417	=ERROR.TIPICO.XY(B2:B9;A2:A9)		

Interpretación: El valor de $s_e = 4{,}064$, significa que los puntos están dispersos a una distancia de 4,064 de la recta de regresión.

Fuente:

Suárez, Mario. & Tapia, Fausto. (2014). *Interaprendizaje de Estadística Básica*. Ibarra, Ecuador: Universidad Técnica de Norte

Suárez, Mario. (2014). *Probabilidades y Estadística empleando las TIC. Ibarra*, Ecuador: Imprenta GRAFICOLOR