

PROBLEMA DEL USO DE FERTILIZANTE EN GRANJAS DE PRODUCCIÓN DE TOMATES.

En el siguiente ejercicio se tratará de exponer, de forma didáctica, el proceso de solución de un problema de regresión simple.

Problema: Uso de fertilizantes para la siembra de tomates.

Este ejercicio o problema ha sido elaborado de forma académica, no representa un experimento real y los valores expresados son unicamente referenciales, aunque sigue un planteamiento metodológico que puede ser utilizado en la práctica profesional. Las medias poblacionales tampoco son reales, se incluyen para ilustrar el proceso de inferencia estadística implicado en un problema econométrico.

En la agricultura tecnificada actual, está extendido el uso de fertilizantes para aumentar la producción por unidad de superficie cosechada

Tabla 1: Medias Condicionales

	Kilos de fertilizantes por hectárea								
	200	250	300	350	400	450	500	550	600
Toneladas de Tomate por hectárea	10	11	18	20	34	38	33	43	49
	9	14	20	22	29	35	43	46	45
	8	9	15	25	31	33	38	40	38
	12	15	16	26		34	44	46	46
	8	13		30			35		48
Medias Condicionales	9,4	12,4	17,25	24,6	31,33	35	38,6	43,75	45,2
Medias Poblacionales	10	15	20	25	30	35	40	45	50

En la tabla 1 se muestra los resultados obtenidos en 40 granjas en las cuales se sembró tomates, en las granjas se usaron diferentes cantidades de fertilizante. Así tenemos que en cinco granjas se usaron 200 kilos de fertilizantes por hectáreas, en otras cinco 250 kilos de fertilizantes, en cuatro granjas se utilizaron 300 kilos de fertilizantes y así hasta 600 kilos de fertilizantes por hectáreas, es decir se agruparon las granjas de acuerdo a la cantidad de fertilizante utilizado. Para el grupo donde se utilizó 200 kilos de fertilizantes (5 granjas), la producción de tomates por hectáreas varió desde 8 hasta 12 toneladas, en el grupo donde se utilizó 250 kilos de fertilizantes la producción por hectárea varió desde 9 toneladas hasta 15 toneladas y así sucesivamente. Es de hacer notar que algunas granjas del grupo donde se utilizó 250 kilos de fertilizantes, mostraron una producción menor que otras granjas en el grupo donde se utilizó 200 kilos de fertilizantes, esto es debido que hay otras variables que también intervienen en la producción de tomates, como por ejemplo el tipo de suelo, el clima, etc.

En este ejemplo se muestra que los valores de la variable independiente (X) quedan fijos para las distintas muestras, es decir, el valor de X no cambia (por ejemplo 200 kilos de fertilizante) pero los valores de la variable dependiente (Y) si cambia (para 200 kilos de fertilizante los valores de Y son 10, 9, 8, 12 y 8 toneladas de tomates). Surge ahora la pregunta, ¿cuál producción utilizar? La Respuesta es: la producción promedio, de donde surge el concepto de medias condicionales, en este caso la

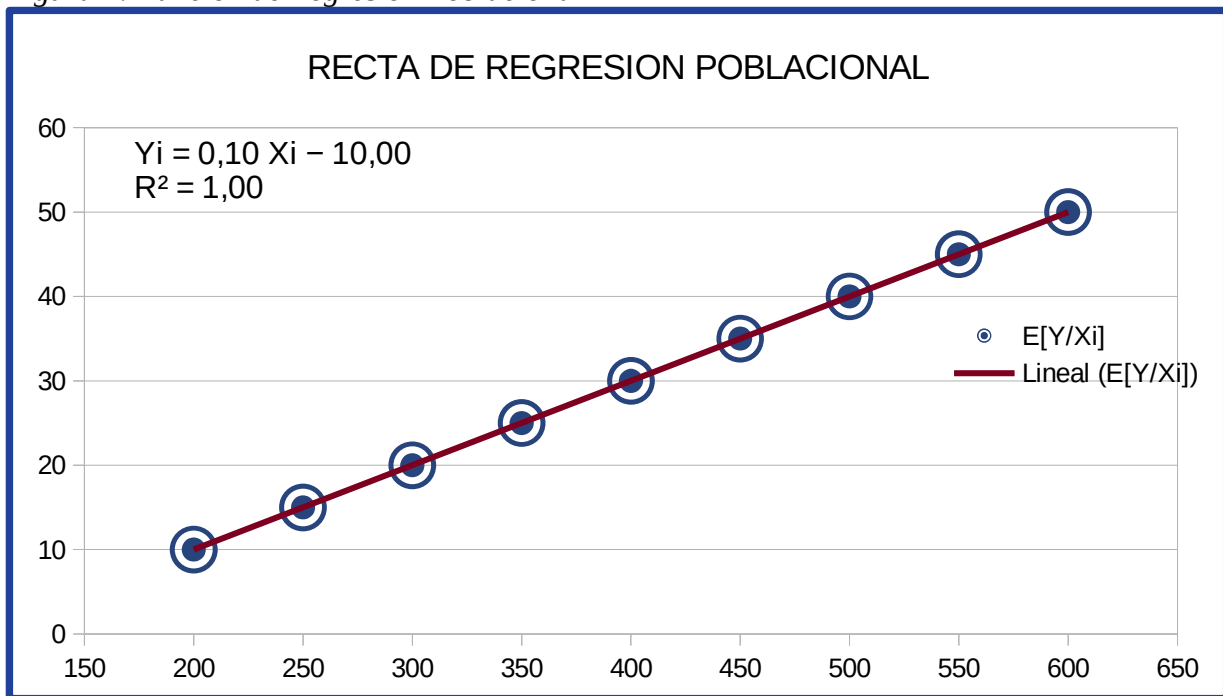
producción de tomates en esta muestra de cinco granjas, está condicionada al uso de 200 kilos de fertilizantes. Así para cada grupo o muestra de granjas.

Para este problema la recta de regresión poblacional, si tomamos los datos correspondiente a las medias poblacionales de la tabla 1, es la siguiente:

$$Y_i = -10 + 0,10 X_i + \mu_i \quad (1)$$

Con un $R^2 = 1$. Si hacemos la grafica de esta ecuación nos queda como se muestra en la figura 1

Figura 1: Función de Regresión Poblacional



Si aplicamos la formula para calcular b_2 , y luego calcular b_1 , según se muestra a continuación:

$$1.- b_1 = \bar{y} - b_2 \bar{x}$$

$$2.- b_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (2)$$

Ambas formulas contienen sumas o sumatorias, primero calculamos las medias tanto de X como de Y, luego calculamos los desvíos con respecto a la media de Y como de X. Estos cálculos se muestran en las tablas 2 y 3.

Tabla 2: Datos para Cálculo de Coeficientes.

n	x	y	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})(Y_i - \bar{Y})$
1	200,00	9,40	(200,00)	40.000,00	(19,21)	3.842,96
2	250,00	12,40	(150,00)	22.500,00	(16,21)	2.432,22
3	300,00	17,25	(100,00)	10.000,00	(11,36)	1.136,48
4	350,00	24,60	(50,00)	2.500,00	(4,01)	200,74
5	400,00	31,33	-	-	2,72	-
6	450,00	35,00	50,00	2.500,00	6,39	319,26
7	500,00	38,60	100,00	10.000,00	9,99	998,52
8	550,00	43,75	150,00	22.500,00	15,14	2.270,28
9	600,00	45,20	200,00	40.000,00	16,59	3.317,04
Σ	3.600,00	257,53	0,00	150.000,00	-	14.517,50

$$\bar{X} = 400$$

$$b_1 = -10,0985$$

$$\bar{Y} = 28,615$$

$$b_2 = 0,0968$$

Al sustituir los valores de las sumatorias de la tabla 2 en la correspondiente formula para calcular b_2 , nos queda:

$$b_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{14.517,50}{150.000} = 0,0968 \quad (3)$$

El valor obtenido de b_2 se utiliza para calcular b_1 :

$$b_1 = 28,615 - 0,0968 \times 400 = -10,0985$$

Con estos valores formulamos la función de regresión muestral:

$$Y_i = -10,0985 + 0,0968 X_i + e_i \quad (4)$$

Con esta ecuación, podemos calcular los valores estimados de Y, para luego proceder a calcular los errores. Los valores estimados de Y se calculan sustituyendo los valores de X en la función de regresión muestral sin los errores, este procedimiento se muestra a continuación:

$$\hat{Y}_1 = -10,0985 + 0,0968x(200) = 9,26$$

$$\hat{Y}_2 = -10,0985 + 0,0968x(250) = 14,1$$

·
·
·

$$\hat{Y}_9 = -10,0985 + 0,0968x(200) = 47,97$$

Para el calcular los errores $e_i = Y_i - \hat{Y}$, se resta el valor de estimado de Y (\hat{Y}) calculado en el paso previo del valor observado de Y, esto lo mostramos a continuación:

$$e_1 = 9,40 - 9,26 = 0,14$$

$$e_2 = 12,40 - 14,1 = -1,7$$

·
·
·

$$e_9 = 45,20 - 47,97 = -2,77$$

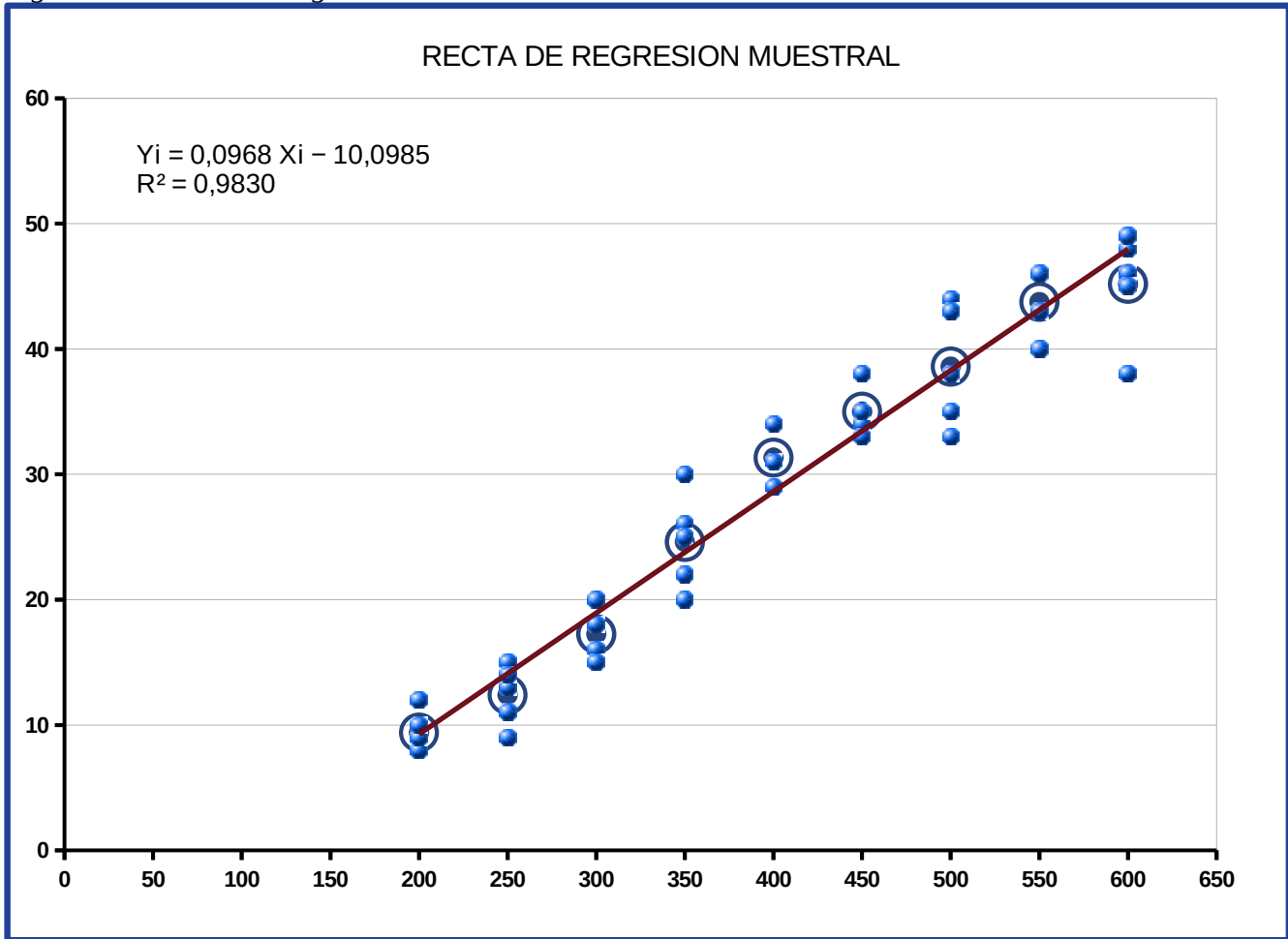
Como la suma de los errores da cero, debemos elevarlos al cuadrado, adicionalmente debemos calcular los desvíos con respecto a la media de Y (\bar{Y}), estos calculos se llevan a una tabla, como se muestra en la tabla 3

Tabla 3: Datos para Cálculo de Coeficientes. Continuación...

n	X	Y	X ²	(\hat{Y}_i)	($Y_i - \hat{Y}_i$)	($Y_i - \hat{Y}_i$) ²	($\hat{Y}_i - \bar{Y}$)	($\hat{Y}_i - \bar{Y}$) ²	($Y_i - \bar{Y}$) ²
1	200,00	9,40	40.000,00	9,26	0,14	0,02	-19,36	374,68	369,21
2	250,00	12,40	62.500,00	14,1	-1,7	2,88	-14,52	210,76	262,92
3	300,00	17,25	90.000,00	18,94	-1,69	2,84	-9,68	93,67	129,16
4	350,00	24,60	122.500,00	23,78	0,82	0,68	-4,84	23,42	16,12
5	400,00	31,33	160.000,00	28,61	2,72	7,39	0	0	7,39
6	450,00	35,00	202.500,00	33,45	1,55	2,39	4,84	23,42	40,77
7	500,00	38,60	250.000,00	38,29	0,31	0,09	9,68	93,67	99,7
8	550,00	43,75	302.500,00	43,13	0,62	0,38	14,52	210,76	229,07
9	600,00	45,20	360.000,00	47,97	-2,77	7,68	19,36	374,68	275,07
Σ	3.600,00	257,53	1.590.000,00	257,53	0,00	24,36	0,00	1.405,05	1.429,41

La grafica de la función de regresión muestral se presenta en la figura 2, los puntos dentro de círculos son las medias condicionales de cada grupo de granja, y la recta de regresión se ajusta a estas medias condicionales dados los valores de X.

Figura 2: Función de Regresión Muestral



Análisis de Varianzas:

Para el análisis de varianza debemos verificar el cumplimiento de la conocida identidad:

$$SCT = SCE + SCR$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (\hat{y}_i - y_i)^2 \tag{5}$$

En la tabla 3 encontramos los valores correspondientes a cada una de las sumas de la ecuación 6, de donde obtenemos:

- SCT = 1.429,41
- SCE = 1.405,05
- SCR = 24,36

Al dividir estas sumas de cuadrados entre sus respectivos grados de libertad se obtienen las varianzas o cuadrados medios, que luego nos permitirán hacer inferencia estadística mediante pruebas de hipótesis e intervalos de confianza.

$$cme = \frac{1.405,05}{1} = 1.405,05$$

$$cme = \frac{24,36}{7} = 3,4803 = S^2(e)$$

$$S(e) = \sqrt{3,4803} = 1,8656$$

$$F = \frac{1.405,05}{3,4803} = 403,7151$$

$$R^2 = \frac{1.405,05}{1.429,41} = 0,983 = 98,3 \%$$

El cálculo de las varianzas de los estimadores se muestra a continuación:

Para b_1 :

$$S_{b_1}^2 = \frac{S_e^2}{n} * \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{3,4803}{9} * \frac{1.590.000}{150.000} = 4,0990 \quad (6)$$

Para b_2 :

$$S_{b_2}^2 = \frac{S_e^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{3,4803}{150.000} = 0,0000232 \quad (7)$$

Con las varianzas, calculamos las desviaciones estándar de los estimadores b_1 y b_2 :

$$S(b_1) = \sqrt{4,0990} = 2,0246$$

$$S(b_2) = \sqrt{0,0000232} = 0,0048$$

Y con estos valores obtenemos los t calculados para las pruebas de hipótesis.

$$t_c(b_1) = \frac{-10,0985}{2,0246} = -4,9881$$

$$t_c(b_2) = \frac{0,0968}{0,0048} = 20,0927$$

Con estos valores se realizan las pruebas de hipótesis para la validación individual, tanto para β_1 como para β_2 .

Validación individual para los coeficientes del modelo estimado:

Prueba de hipótesis del $(1 - \alpha)$ % para el parámetro β_1 .

1) Hipótesis:

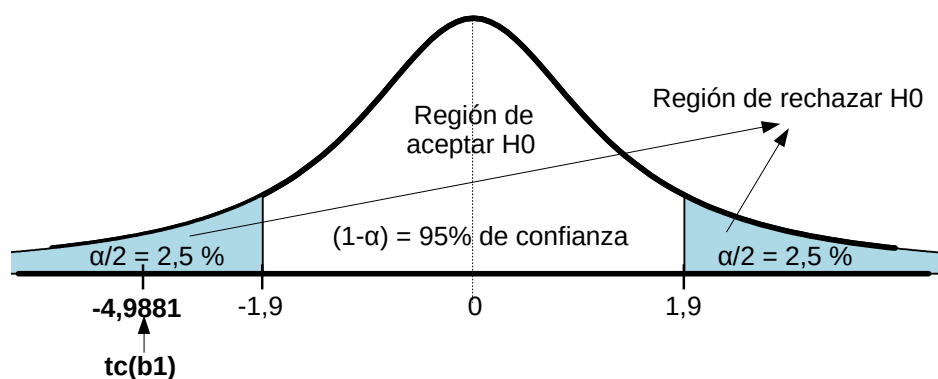
$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

2) Estadístico de prueba:

$$t_c(b_1) = \frac{-10,0985}{2,0246} = -4,9881 \quad \text{El valor crítico, de acuerdo a la tabla t, es: } t(0.95, 7) = 1.9$$

3) Criterio de aceptación o rechazo de la hipótesis nula β_1 :



4) Conclusión:

Con un 95% de confianza, la muestra de tamaño $n = 9$, no nos da evidencia para aceptar la hipótesis nula de que el parámetro $\beta_1 = 0$ es igual a cero, por consiguiente se valida como distinto de cero este parámetro del modelo, es decir, el valor t calculado es desde el punto de vista estadístico significativo.

Prueba de hipótesis del $(1 - \alpha)$ % para el parámetro β_2 .

1) Hipótesis:

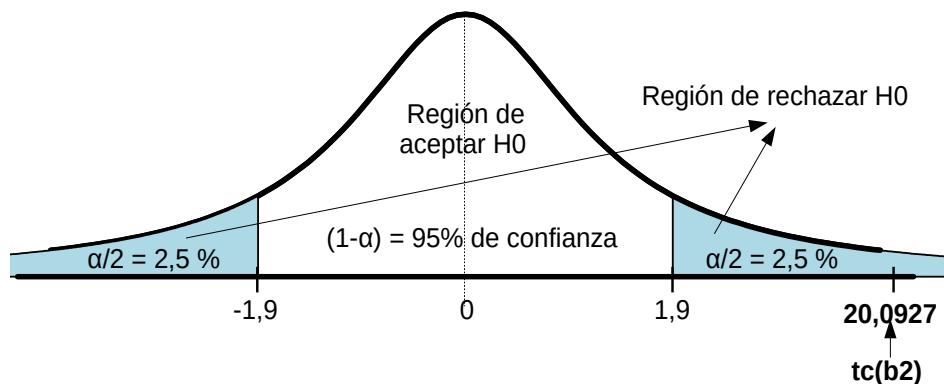
$$H_0: \beta_2 = 0$$

$$H_a: \beta_2 \neq 0$$

2) Estadístico de prueba:

$$t_c(b_2) = \frac{0,0968}{0,0048} = 20,0927$$

3) Criterio de aceptación o rechazo de la hipótesis nula β_2 :



4) Conclusión:

Con un 95% de confianza, la muestra de tamaño $n = 9$, no nos da evidencia para aceptar la hipótesis nula de que el parámetro $\beta_2 = 0$ es igual a cero, por consiguiente se valida como distinto de cero este parámetro del modelo, es decir, el valor t calculado es desde el punto de vista estadístico significativo.

Validación global o conjunta para los coeficientes del modelo estimado:

Prueba de hipótesis del $(1 - \alpha)$ % para los parámetros β_1 y β_2 .

1) Hipótesis:

$$H_0: \beta_1 \text{ y } \beta_2 = 0 \text{ en forma conjunta.}$$

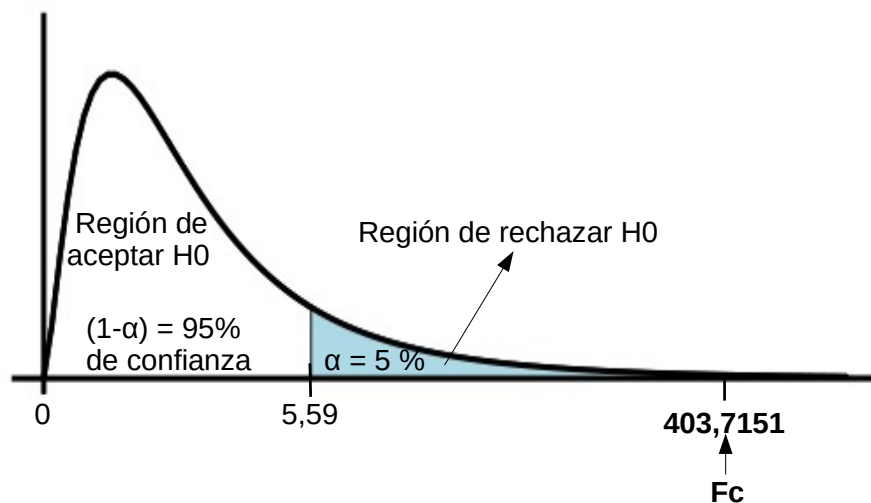
$$H_a: \beta_1 \text{ y } \beta_2 \neq 0$$

2) Estadístico de prueba:

$$F_c = \frac{1.405,05}{3,4803} = 403,7151$$

Valor crítico para la distribución F : $F(0.95, 1, 7) = 5.59$

3) Criterio de aceptación o rechazo de la hipótesis nula β_1 y $\beta_2 = 0$ en forma conjunta



4) Conclusión:

Con un 95% de confianza, la muestra de tamaño $n = 9$, no nos da evidencia para aceptar la hipótesis nula de que los parámetros β_1 y β_2 son iguales a cero en forma conjunta o global, por consiguiente se valida como distintos de cero estos parámetros del modelo, es decir, la prueba F es, desde el punto de vista estadístico significativa. Se concluye que con un 95 % de confianza, el modelo es apto para predecir

Intervalos del $(1 - \alpha)$ % de confianza.

Intervalos para la predicción media y para la predicción puntual.

Problema: ¿Cuánto será la producción de tomates si se utilizan 415 kilos de fertilizantes? Construya e interprete un intervalo del 95 % de confianza tanto para la predicción media como para la predicción individual.

Primero utilizamos la función de regresión muestral para estimar la producción de tomate cuando se utilizan 415 kilos de fertilizantes ($X = 415$)

$\hat{Y}(415) = -10,0989 + 0,0968*(415) = 30,0662$ producción promedio de tomates en toneladas cuando se utilizan 415 de fertilizantes.

Luego procedemos a calcular las varianzas de la estimación media:

$$S^2(\hat{y}_0) = \left[\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n - k} \right] * \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] = 3,4803 \left[\frac{1}{9} + \frac{225}{150000} \right] = 0,3919 \quad (8)$$

Con las varianzas, calculamos las desviaciones estándar:

$$S(\hat{Y}_0) = \sqrt{0,3919} = 0,6260$$

La varianza de la estimación puntual es:

$$S^2(y_0 - \hat{y}_0) = \left[\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n - k} \right] * \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] = 3,4803 \left[1 + \frac{1}{9} + \frac{225}{150000} \right] = 3,8722 \quad (9)$$

La desviación estándar es:

$$S(b_2) = \sqrt{1,9678} = 1,9678$$

Hora construimos el intervalo del 95 % de confianza para la estimación media:

[30,0662 - (1,9)(0,6260) ; 30,0662 + (1,9)(0,6260)] = [**28,8767 ; 31,2557**] Intervalo del 95 % de confianza para la producción media de tomates en toneladas cuando se utilizan 415 kilos de fertilizantes.

Interpretación:

Con un 95 % de confianza, el intervalo [**28,8767 ; 31,2557**], construido a partir de una muestra de tamaño 9, contendrá el verdadero valor de la producción media de tomates en toneladas cuando el uso de fertilizantes es de 415 kilos.

El intervalo del 95 % de confianza para la estimación puntual es:

[30,0662 - (1,9)(1,9678) ; 30,0662 + (1,9)(1,9678)] = [**26,3274 ; 33,8050**] Intervalo del 95 % de confianza para la producción de tomates en toneladas cuando se utilizan 415 kilos de fertilizantes.

Interpretación:

Con un 95 % de confianza, el intervalo [**26,3274 ; 33,8050**], construido a partir de una muestra de tamaño 9, contendrá el verdadero valor de la producción de tomates en toneladas cuando el uso de fertilizantes es de 415 kilos.

Como se observa en los resultados obtenidos, el intervalo de confianza para la estimación puntual, es más grande que el intervalo de confianza para la estimación media. Esto es así porque la media es un valor menos representativo o preciso que la estimación puntual en si.

Intervalo del 95 % de confianza para la varianza (σ^2):

Valores críticos para la distribución chi cuadrado: $\chi^2(0.95, 7) = 14.1$; $\chi^2(0.05, 7) = 2.17$

$$\left[7 * \frac{3,4803}{14,1} ; 7 * \frac{3,4803}{2,17} \right] = [**1,7278 ; 11,2268**] \text{ Intervalo del 95 \% de confianza para la varianza.}$$

Interpretación:

Con un 95 % de confianza, el intervalo [**1,7278 ; 11,2268**], construido a partir de una muestra de tamaño 9, contendrá el verdadero valor de la varianza.

Intervalo del 95 % de confianza para B₁:

[$-10,0989 - (1,9)(2,0246)$; $-10,0989 + (1,9)(2,0246)$] = [**-13,9456 ; -6,2521**] Intervalo del 95 % de confianza para B₁.

Interpretación:

Con un 95 % de confianza, el intervalo [**-13,9456 ; -6,2521**], construido a partir de una muestra de tamaño 9, contendrá el verdadero valor de B₁.

Intervalo del 95 % de confianza para B₂:

[$0,0968 - (1,9)(0,0048)$; $0,0968 + (1,9)(0,0048)$] = [**0,0876 ; 0,1059**] Intervalo del 95 % de confianza para B₂.

Interpretación:

Con un 95 % de confianza, el intervalo [**0,0876 ; 0,1059**], construido a partir de una muestra de tamaño 9, contendrá el verdadero valor de B₂.

Validación económica:

Para la validación económica se tomará en cuenta tanto el signo esperado para el coeficiente como el valor o magnitud de este coeficiente y su relación con las variables del modelo.

Para el coeficiente b₁ (-10,0985): Se espera un signo negativo para este coeficiente, pues representa la producción promedio de tomates (en toneladas) en caso que no se utilice fertilizantes o el valor de la variable explicativa sea igual a cero (X = 0), como el signo dio negativo no tiene una explicación racional o creíble desde el punto de vista económico, aunque alguien podría forzar una interpretación en el sentido que representaría las pérdidas por no utilizar este factor de producción. Por ser este coeficiente el término independiente, en realidad no es muy importante su interpretación, a menos que esté incluido expresamente en la teoría económica que se utilice para especificar el modelo. En la práctica, se mantiene este término independiente aunque no se valide.

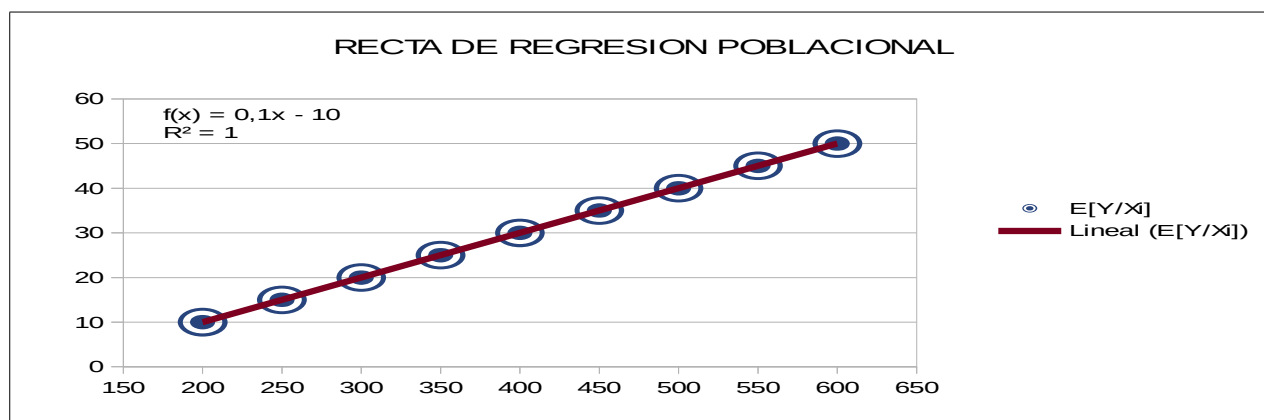
Para el coeficiente b₂ (0,0968): Este coeficiente si tiene el signo esperado desde el punto de vista económico, pues tiene signo positivo. Se entiende que un aumento en el uso de fertilizante hará aumentar la producción de tomate o una disminución en el uso de fertilizante hará disminuir la producción de tomates, es decir, las variaciones van en el mismo sentido por lo que se espera un signo positivo.

PROBLEMA FERTILIZANTE (RESUMEN)

	Kilos de fertilizantes por hectárea								
	200	250	300	350	400	450	500	550	600
Toneladas de Tomate por hectárea	10	11	18	20	34	38	33	43	49
	9	14	20	22	29	35	43	46	45
	8	9	15	25	31	33	38	40	38
	12	15	16	26		34	44	46	46
	8	13		30			35		48
Medias Condicionales	9,4	12,4	17,25	24,6	31,33	35	38,6	43,75	45,2
Medias Poblacionales	10	15	20	25	30	35	40	45	50

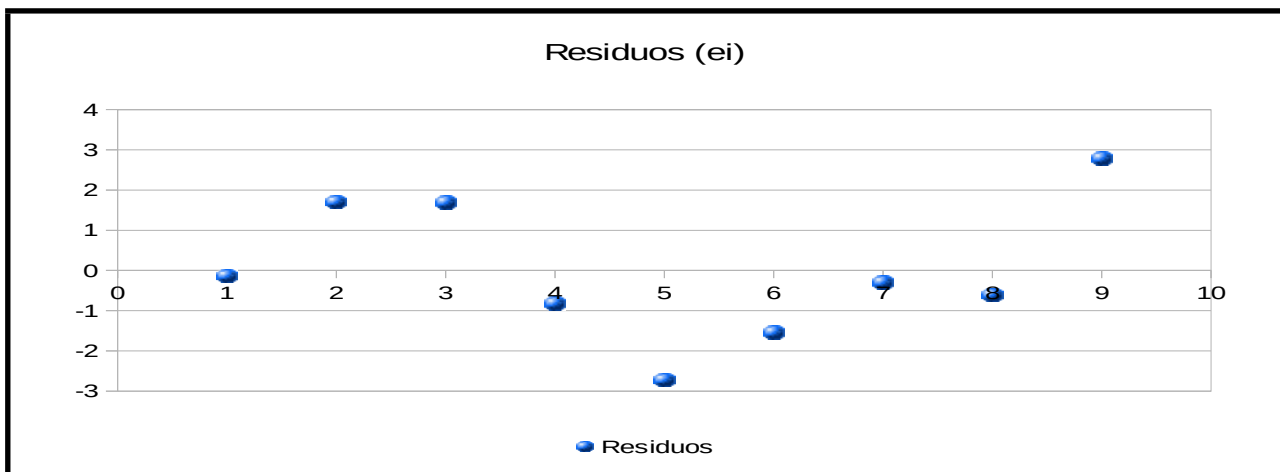
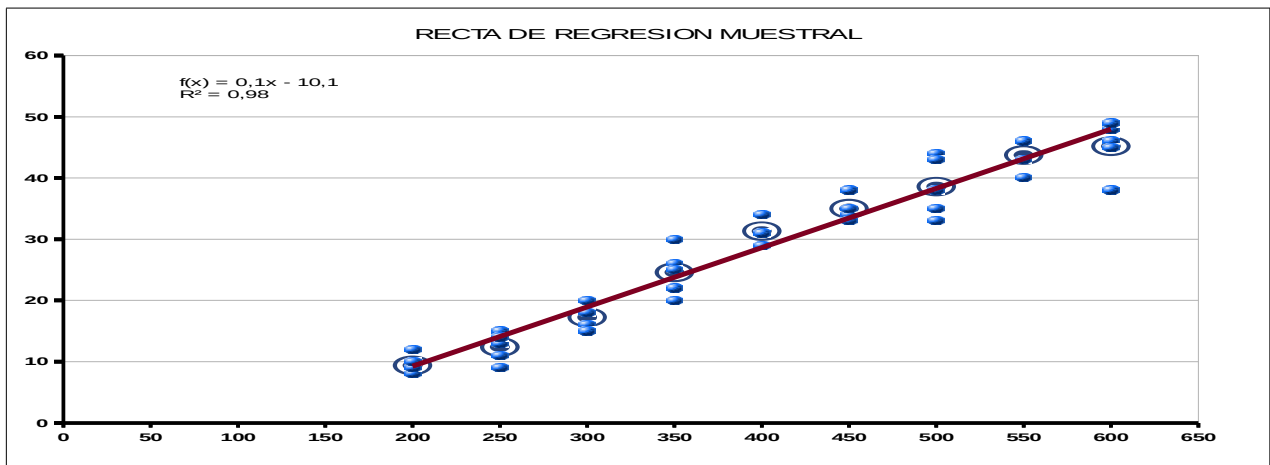
n	X	Y	X - x	(Y - X) ²	Y - Y	(X - X)*(Y - Y)
1	200,00	9,40	(200,00)	40.000,00	(19,21)	3.842,96
2	250,00	12,40	(150,00)	22.500,00	(16,21)	2.432,22
3	300,00	17,25	(100,00)	10.000,00	(11,36)	1.136,48
4	350,00	24,60	(50,00)	2.500,00	(4,01)	200,74
5	400,00	31,33	-	-	2,72	-
6	450,00	35,00	50,00	2.500,00	6,39	319,26
7	500,00	38,60	100,00	10.000,00	9,99	998,52
8	550,00	43,75	150,00	22.500,00	15,14	2.270,28
9	600,00	45,20	200,00	40.000,00	16,59	3.317,04
Σ	3.600,00	257,53	0,00	150.000,00	0,00	14.517,50

X =	400	b =	0,0968
Y =	28,61	a =	-10,1



Problema Fertilizantes. (Continuación)

n	X	Y	X ²	(\hat{Y})	($\hat{Y} - Y$)	($\hat{Y} - Y$) ²	($\hat{Y} - \bar{Y}$)	($\hat{Y} - \bar{Y}$) ²	($\bar{Y} - Y$) ²
1	200,00	9,40	40.000,00	9,26	-0,14	0,02	-19,36	374,68	369,21
2	250,00	12,40	62.500,00	14,1	1,7	2,88	-14,52	210,76	262,92
3	300,00	17,25	90.000,00	18,94	1,69	2,84	-9,68	93,67	129,16
4	350,00	24,60	122.500,00	23,78	-0,82	0,68	-4,84	23,42	16,12
5	400,00	31,33	160.000,00	28,61	-2,72	7,39	0	0	7,39
6	450,00	35,00	202.500,00	33,45	-1,55	2,39	4,84	23,42	40,77
7	500,00	38,60	250.000,00	38,29	-0,31	0,09	9,68	93,67	99,7
8	550,00	43,75	302.500,00	43,13	-0,62	0,38	14,52	210,76	229,07
9	600,00	45,20	360.000,00	47,97	2,77	7,68	19,36	374,68	275,07
Σ	3.600,00	257,53	1.590.000,00	257,53	(0,00)	24,36	0,00	1.405,05	1.429,41



$t(0.95, 7) = 1.9$
 $F(0.95, 1, 7) = 5.59$

$\chi^2(0.95, 7) = 14.1$; $\chi^2(0.05, 7) = 2.17$
 $Z(0.95) = 1.96$ Una cola. $Z(0.95) = \pm 1.645$ Dos colas.