UNIVERSIDAD TECNOLÓGICA DEL NORTE DE COAHUILA.

CORRELACION Y REGRESIÓN LINEAL

Nombre: Mario Alberto Caldera Loera.

Carrera: Procesos de Producción.

Grupo: 3“A”

Materia: Matemáticas lll

Maestro: Ing. Luis Arturo García Navarro.

Fecha de entrega: 18-agosto-2009

INDICE

1. Objetivo:

Que el alumno aprenda a realizar predicciones sobres sucesos futuros usando las herramientas estadísticas, parta aplicarlas a procesos de producción.
Enseñar al alumno a utilizar correctamente los diagramas de dispersión teniendo como apoyo el software llamado MINITAB 14.

2. Antecedentes:

Variable

Una **variable** es un símbolo que representa un elemento no especificado de un conjunto dado. Dicho conjunto es llamado **conjunto universal** de la variable, **universo** o **dominio** de la variable, y cada elemento del conjunto es un **valor** de la variable. Sea *x* una variable cuyo universo es el conjunto {1,3,5,7,9,11,13}; entonces *x* puede tener cualquiera de esos valores: 1,3,5,7,9,11,13. En otras palabras *x* puede reemplazarse por cualquier entero positivo impar menor que 14. Por esta razón, a menudo se dice que una variable es un *reemplazo* de cualquier elemento de su universo.

Una **variable** es un elemento de una fórmula, proposición o algoritmo que puede adquirir o ser sustituido por un valor cualquiera (siempre dentro de su universo). Los valores que una variable es capaz de recibir, pueden estar definidos dentro de un rango, y/o estar limitados por criterios o condiciones de pertenencia, al universo que les corresponde (en estos casos, el universo de la variable pasa a ser un subconjunto de un universo mayor, el que tendría sin las restricciones).

Medición

Es comparar la cantidad desconocida que queremos determinar y una cantidad conocida de la misma magnitud, que elegimos como unidad. Teniendo como punto de referencia dos cosas: un objeto (lo que se quiere medir) y una unidad de medida ya establecida ya sea en Sistema Ingles, Sistema Internacional, o Sistema Decimal.

Al resultado de medir lo llamamos Medida.

Cuando medimos algo se debe hacer con gran cuidado, para evitar alterar el sistema que observamos. Por otro lado, no hemos de perder de vista que las medidas se realizan con algún tipo de error, debido a imperfecciones del instrumental o a limitaciones del medidor, errores experimentales, por eso, se ha de realizar la medida de forma que la alteración producida sea mucho menor que el error experimental que se pueda cometer.

La medida o medición es directa, cuando disponemos de un instrumento de medida que la obtiene, así si deseamos medir la distancia de un punto "A" a un punto "B", y disponemos del instrumento que nos permite realizar la medición.

Unidades de medida

Al patrón de medir le llamamos también Unidad de medida.

Debe cumplir estas condiciones:

1º.- Ser inalterable, esto es, no ha de cambiar con el tiempo ni en función de quién realice la medida.

2º.- Ser universal, es decir utilizada por todos los países.

3º.- Ha de ser fácilmente reproducible.

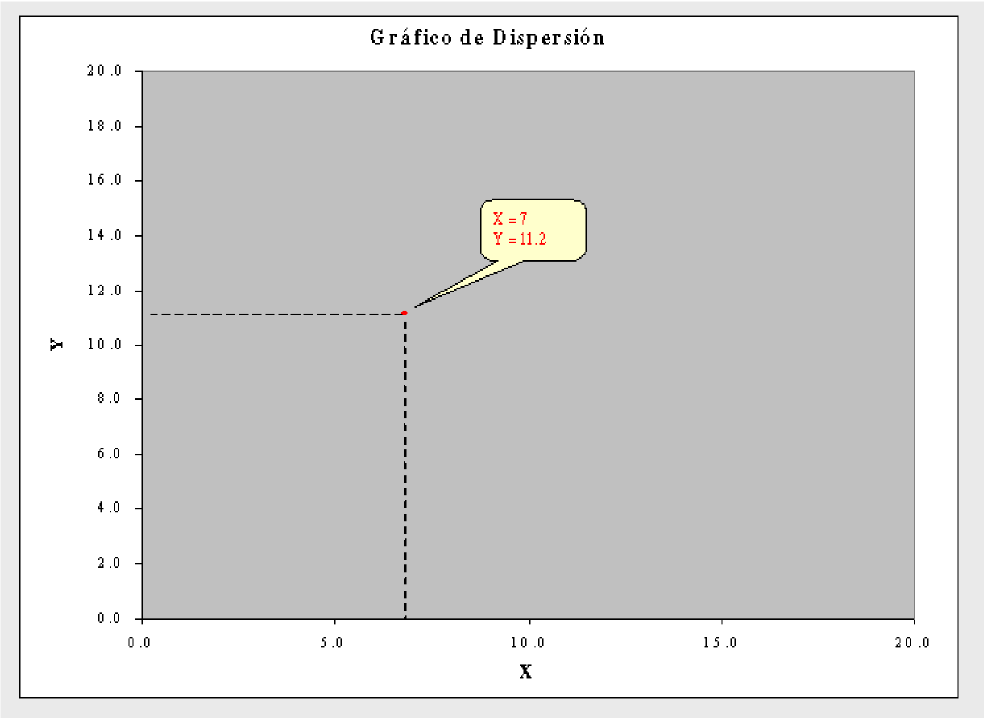
Reuniendo las unidades patrón que los científicos han estimado más convenientes, se han creado los denominados Sistemas de Unidades.

Sistema Internacional (S.I.)

Este nombre se adoptó en el año 1960 en la XI Conferencia General de Pesos y Medidas, celebrada en París buscando en él un sistema universal, unificado y coherente que toma como Magnitudes fundamentales: Longitud, Masa, Tiempo, Intensidad de corriente eléctrica, Temperatura termodinámica, Cantidad de sustancia, Intensidad luminosa. Toma además como magnitudes complementarias: Angulo plano y Angulo sólido.

Diagramas de Dispersión

Los Diagramas de Dispersión o Gráficos de Correlación permiten estudiar la relación entre 2 variables. Dadas 2 variables X e Y, se dice que existe una correlación entre ambas si cada vez que aumenta el valor de X aumenta proporcionalmente el valor de Y (Correlación positiva) o si cada vez que aumenta el valor de X disminuye en igual proporción el valor de Y (Correlación negativa).
En un gráfico de correlación representamos cada par X, Y como un punto donde se cortan las coordenadas de X e Y:



Coeficiente de correlación de Pearson

El **coeficiente de correlación de Pearson** es un índice estadístico que mide la relación lineal entre dos variables cuantitativas. A diferencia de la covarianza, la correlación de Pearson es independiente de la escala de medida de las variables.

El cálculo del coeficiente de correlación lineal se realiza dividiendo la covarianza por el producto de las desviaciones estándar de ambas variables:

$$r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$$

Siendo:

σ_{XY} la covarianza de (X,Y)

σ_X y σ_Y las desviaciones típicas de las distribuciones marginales.

El valor del índice de correlación varía en el intervalo [-1, +1]:

- Si r = 0, no existe relación lineal. Pero esto no necesariamente implica una independencia total entre las dos variables, es decir, que la variación de una de ellas puede influir en el valor que pueda tomar la otra. Pudiendo haber relaciones no lineales entre las dos variables. Estas pueden calcularse con la razón de correlación.
- Si r = 1, existe una correlación positiva perfecta. El índice indica una dependencia total entre las dos variables denominada *relación directa*: cuando una de ellas aumenta, la otra también lo hace en idéntica proporción.
- Si 0 < r < 1, existe una correlación positiva.
- Si r = -1, existe una correlación negativa perfecta. El índice indica una dependencia total entre las dos variables llamada *relación inversa*: cuando una de ellas aumenta, la otra disminuye en idéntica proporción.
- Si -1 < r < 0, existe una correlación negativa.

Regresión lineal

Ejemplo de una regresión lineal con una variable dependiente y una variable independiente.

En estadística la **regresión lineal** o **ajuste lineal** es un método matemático que modeliza la relación entre una variable dependiente *Y*, las variables independientes X_i y un término aleatorio ε. Este modelo puede ser expresado como:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

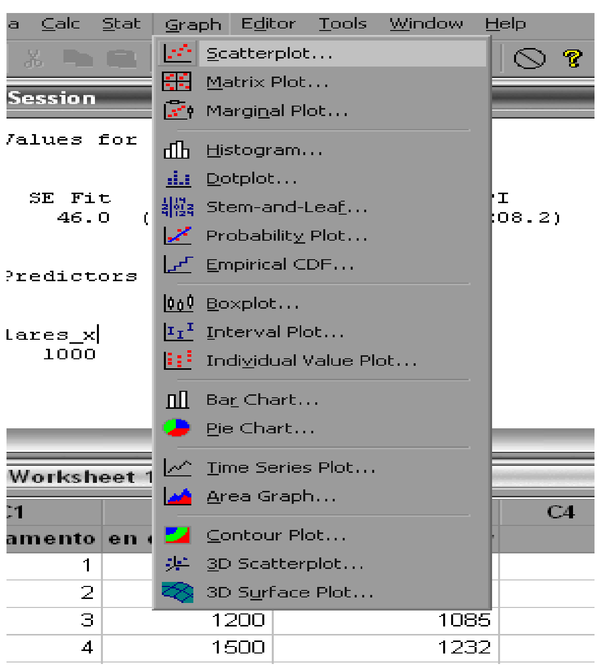
donde β_0 es la intersección o término "constante", las β_i son los parámetros respectivos a cada variable independiente, y *p* es el número de parámetros independientes a tener en cuenta en la regresión. La regresión lineal puede ser contrastada con la regresión no lineal.

3. ejemplos teóricos.

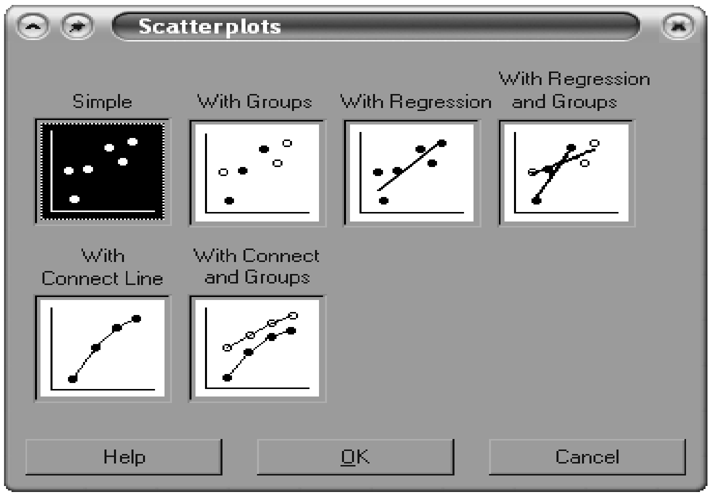
I. Una compañía de bienes raíces residenciales en una ciudad grande desea poder predecir los costos mensuales de rentas para departamentos, basado en el tamaño de los mismos definidos por los pies cuadrados de espacios. Selecciona una muestra.

Renta mensual		Tamaño	Renta mensual		Tamaño
Departamento	**En Dólares**	**Pies cuadrados**	**Departamento**	**En dólares**	**Pies cuadrados**
1	950	850	14	1800	1369
2	1600	1450	15	1400	1175
3	1200	1085	16	1450	1225
4	1500	1232	17	1100	1245
5	950	718	18	1700	1259
6	1700	1485	19	1200	1150
7	1650	1136	20	1150	896
8	935	726	21	1600	1361
9	875	700	22	1650	1040
10	1150	956	23	1200	755
11	1400	1100	24	800	1000
12	1650	1285	25	1750	1200
13	2300	1985			

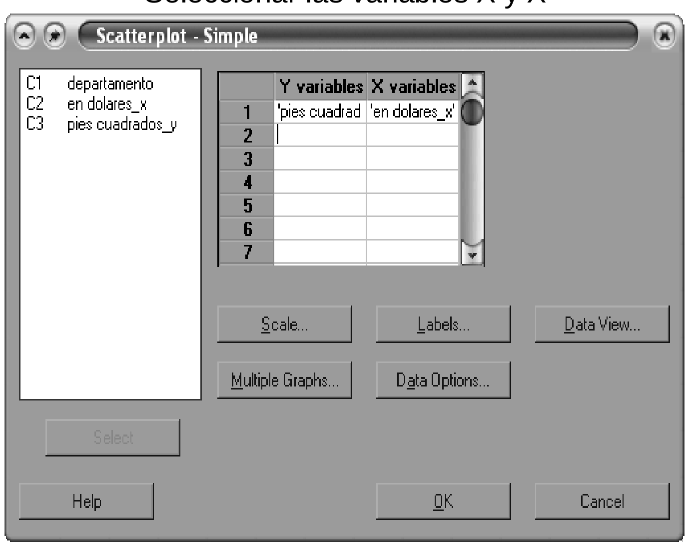
a) Grafique el diagrama de dispersión (en minitab 14). Menú graph seleccionar scatter plot.

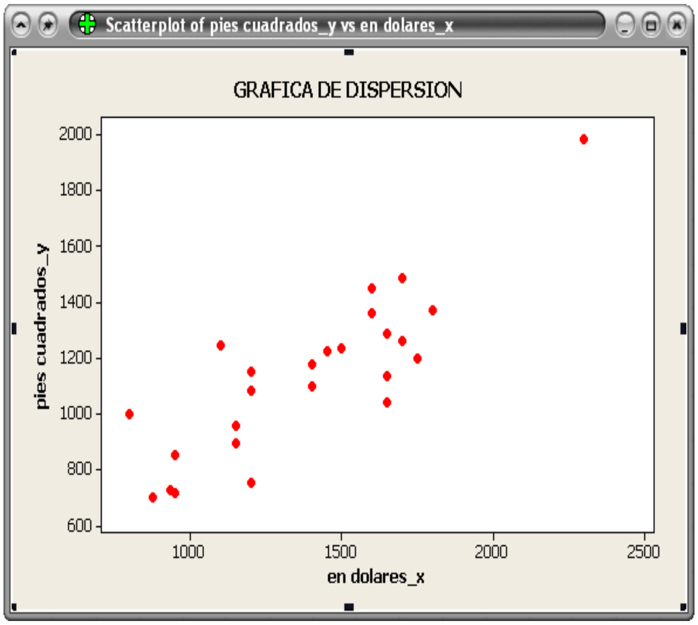


Seleccionar el tipio de grafica deseada en este caso simple.



Seleccionar las variables X y X





b) Utilice el método de mínimos cuadrados para encontrar los coeficientes de regresión a y b.

en dolares_x	pies cuadrados_y	(x-x)	(x-x)2	(Y-Y)	(Y-Y)2	(x-x)(Y-Y)
950	850	-436.4	190444.96	-285.32	81407.5024	124513.648
1600	1450	213.6	45624.96	314.68	99023.5024	67215.648
1200	1085	-186.4	34744.96	-50.32	2532.1024	9379.648
1500	1232	113.6	12904.96	96.68	9347.0224	10982.848
950	718	-436.4	190444.96	-417.32	174155.9824	182118.448
1700	1485	313.6	98344.96	349.68	122276.1024	109659.648
1650	1136	263.6	69484.96	0.68	0.4624	179.248
935	726	-451.4	203761.96	-409.32	167542.8624	184767.048
875	700	-511.4	261529.96	-435.32	189503.5024	222622.648
1150	956	-236.4	55884.96	-179.32	32155.6624	42391.248
1400	1100	13.6	184.96	-35.32	1247.5024	-480.352
1650	1285	263.6	69484.96	149.68	22404.1024	39455.648
2300	1985	913.6	834664.96	849.68	721956.1024	776267.648
1800	1369	413.6	171064.96	233.68	54606.3424	96650.048
1400	1175	13.6	184.96	39.68	1574.5024	539.648
1450	1225	63.6	4044.96	89.68	8042.5024	5703.648
1100	1245	-286.4	82024.96	109.68	12029.7024	-31412.352
1700	1259	313.6	98344.96	123.68	15296.7424	38786.048
1200	1150	-186.4	34744.96	14.68	215.5024	-2736.352
1150	896	-236.4	55884.96	-239.32	57274.0624	56575.248
1600	1361	213.6	45624.96	225.68	50931.4624	48205.248
1650	1040	263.6	69484.96	-95.32	9085.9024	-25126.352
1200	755	-186.4	34744.96	-380.32	144643.3024	70891.648
800	1000	-586.4	343864.96	-135.32	18311.5024	79351.648
1750	1200	363.6	132204.96	64.68	4183.5024	23517.648
34660	28383	-2.27374E-12	3139726	1.59162E-12	1999747.44	2130018.8

Sacamos las medias

$$X = \frac{\sum x}{n} \quad = \quad 1386.4$$

$= \frac{34660}{25}$

$Y = \frac{\sum y}{n} = \frac{28383}{25} = 1135.3$

Sacamos las desviaciones estandares.

$Sx^2 = \frac{\sum(x-x)^2}{n-1} = \frac{3139726}{24} = 130821.41$

$Sx = (130821.41)^{½} = 361.69$

$Sy^2 = \frac{\sum(y-y)^2}{n-1} = \frac{3139726}{24} = 83322.81$

$Sy = (83322.81)^{½} = 288.65$

Sacamos la covarianza.

$Sxy = \frac{\sum(x-x)(y-y)}{n-1} = \frac{2130018.8}{24} = 88750.78$

Sacamos la correlación.

$r = \frac{Sxy}{SxSy} = \frac{88750.78}{361.69\ (288.65)} = 0.850$

$b = r \frac{(Sy)}{(Sx)} = 0.850 \frac{(288.65)}{(361.69)} = .6783$

a = y- bx = 1135.32-.6783 (1386.4) = 194.9249

c) establezca la ecuación de regresión.

Ecuación de regresión:

y = a + bx

d) pronostique la renta mensual promedio para un departamento que tiene 1,000 pies cuadrados.

y = 194.9249+ 0.6783 (1000)
y = 873

2. La materia prima que se usa en la elaboración de una fibra sintética se almacena en un local que no tiene control de humedad. Las mediciones de la humedad relativa en el local y del contenido de humedad de una muestra de la materia prima (ambos en porcentajes) durante 12 días, dieron los siguientes resultados:

Humedad, X	Contenido de Humedad, Y
42	12
35	8

50	14
43	9
48	11
62	16
31	7
36	9
44	12
39	10
55	13
48	11

Ajuste una línea recta y determine el contenido de humedad cuando la humedad del local de almacenamiento es de 40%.

X	Y	(x-x)	(x-x)2	(Y-Y)	(Y-Y)2	(x-x)(Y-Y)
42	12	-2.42	5.8564	1	1	-2.42
35	8	-9.42	88.7364	-3	9	28.26
50	14	5.58	31.1364	3	9	16.74
43	9	-1.42	2.0164	-2	4	2.84
48	11	3.58	12.8164	0	0	0
62	16	17.58	309.0564	5	25	87.9
31	7	-13.42	180.0964	-4	16	53.68
36	9	-8.42	70.8964	-2	4	16.84
44	12	-0.42	0.1764	1	1	-0.42
39	10	-5.42	29.3764	-1	1	5.42
55	13	10.58	111.9364	2	4	21.16
48	11	3.58	12.8164	0	0	0
533	132	-0.04	854.9168	0	74	230

Sacamos las medias

$$X = \frac{\sum x}{n} = \frac{533}{12} = 44.42$$

$$Y = \frac{\sum y}{n} = \frac{132}{12} = 11$$

Sacamos las desviaciones estándares.

$$Sx^2 = \frac{\sum (x-x)^2}{n-1} = \frac{854.91}{11} = 77.71$$

$$S_x^2 \quad \text{n-1} \qquad 11$$

$S_x = (77.71)^{½} = 8.81$

$S_y^2 = \frac{\sum(y-\bar{y})^2}{\text{n-1}} = \frac{74}{11} = 6.72$ $\qquad S_y = (6.72)^{½} = 2.59$

Sacamos la covarianza.

$$S_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1} = \frac{230}{11} = 20.90$$

Sacamos la correlación.

$$r = \frac{S_{xy}}{S_xS_y} = \frac{20.90}{8.81\,(2.59)} = 0.915$$

$$b = r\,\frac{(S_y)}{(S_x)} = 0.915\ \frac{(2.59)}{(8.81)} = 0.268$$

$a = y- bx = 11 – 0.268\,(44.42) = - 0.90$

Ecuación de regresión:

$Y = a + bx$

$Y = -0.90 + .268(40) = 9.82$

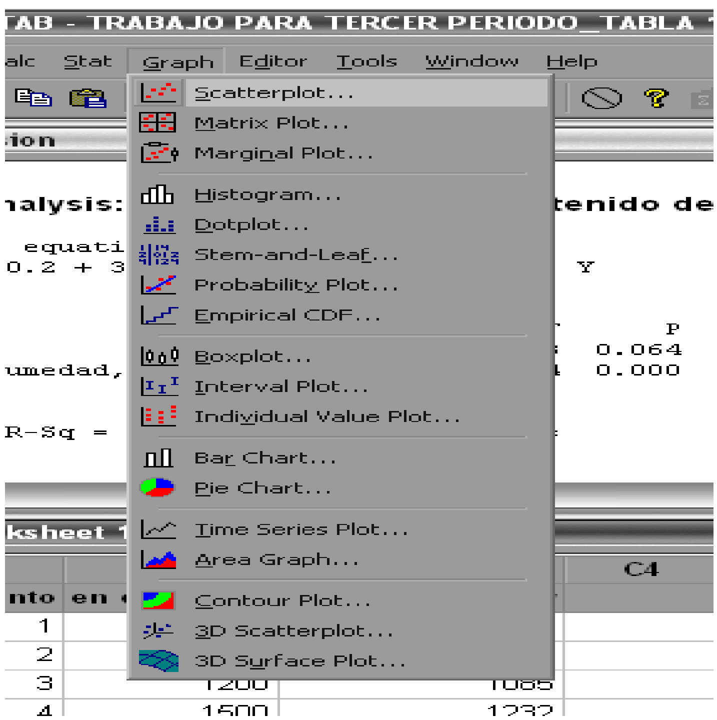
3. Un negocio de ventas por catalogo de artículos para computadoras personales, software y hardware, tiene un almacén centralizado para la distribución de los productos que se le ordenan. La administración examina el proceso de distribución desde el almacén y está interesado en estudiar los factores que afecten los costos de distribución. En la actualidad, se agrega un pequeño cargo por envío independiente del monto de la orden. Se recolectan datos durante los últimos 24 meses que indican los costos de distribución y el número de órdenes recibidas. Los resultados son los siguientes:

	Costo de Distribución			Costo de distribución	
Mes	**Miles de Dólares**	**Número de Ordenes**	**Mes**	**Miles de Dólares**	**Número de Ordenes**
1	52.95	4,015	13	62.98	3,977
2	71.66	3,806	14	72.30	4,428
3	85.58	5,308	15	58.99	3,964
4	63.69	4,262	16	79.38	4,582
5	72.81	4,269	17	94.44	5,582
6	68.44	4,097	18	59.74	3,450
7	52.46	3,213	19	90.50	5,079
8	70.77	4,809	20	93.24	5,735
9	82.03	5,237	21	69.33	4,269
10	74.39	4,732	22	53.71	3,708

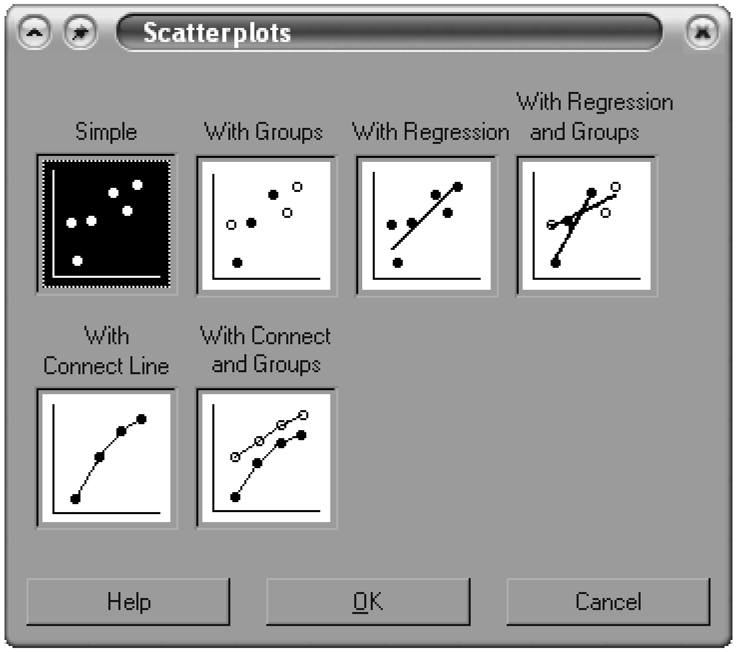
11	70.84	4,413	23	89.18	5,387
12	54.08	2,291	24	66.80	4,161

a) Establezca el diagrama de dispersión.

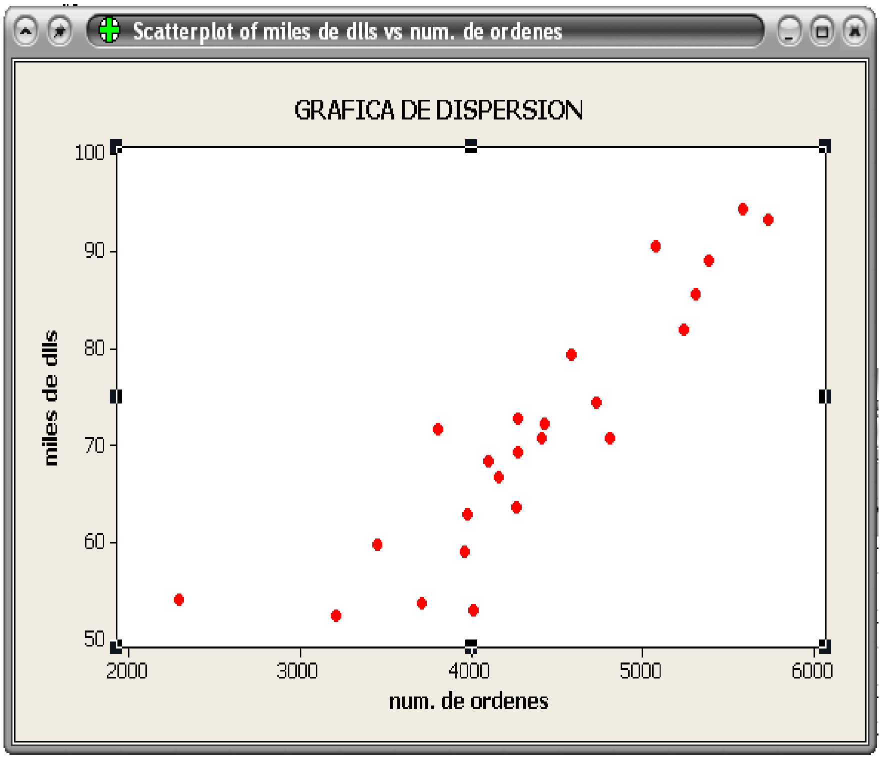
Menú graph seleccionar scatter plot.



Seleccionar el tipio de grafica deseada en este caso simple.



Seleccionar las variables X y X



b) Suponga una relación lineal y utilice el método de mínimos cuadrados para encontrar los coeficientes de regresión a y b.

miles de dlls	num. de ordenes	(x-x)	(x-x)2	(Y-Y)	(Y-Y)2	(x-x)(Y-Y)
52.95	4015	-18.31	335.2561	-351	123201	6426.81
71.66	3806	0.4	0.16	-560	313600	-224
85.58	5308	14.32	205.0624	942	887364	13489.44
63.69	4262	-7.57	57.3049	-104	10816	787.28
72.81	4269	1.55	2.4025	-97	9409	-150.35
68.44	4097	-2.82	7.9524	-269	72361	758.58
52.46	3213	-18.8	353.44	-1153	1329409	21676.4
70.77	4809	-0.49	0.2401	443	196249	-217.07
82.03	5237	10.77	115.9929	871	758641	9380.67
74.39	4732	3.13	9.7969	366	133956	1145.58
70.84	4413	-0.42	0.1764	47	2209	-19.74
54.08	2291	-17.18	295.1524	-2075	4305625	35648.5
62.98	3977	-8.28	68.5584	-389	151321	3220.92
72.3	4428	1.04	1.0816	62	3844	64.48
58.99	3964	-12.27	150.5529	-402	161604	4932.54
79.38	4582	8.12	65.9344	216	46656	1753.92
94.44	5582	23.18	537.3124	1216	1478656	28186.88
59.74	3450	-11.52	132.7104	-916	839056	10552.32
90.5	5079	19.24	370.1776	713	508369	13718.12
93.24	5735	21.98	483.1204	1369	1874161	30090.62
69.33	4269	-1.93	3.7249	-97	9409	187.21
53.71	3708	-17.55	308.0025	-658	432964	11547.9
89.18	5387	17.92	321.1264	1021	1042441	18296.32
66.8	4161	-4.46	19.8916	-205	42025	914.3
1710.29	104774	0.05	3845.1305	-10	14733346	212167.63

Sacamos las medias

$X = \frac{\sum x}{n} = \frac{1710.29}{24} = 71.26$

$\frac{\sum y}{n} = \frac{104774}{24} = 4365.58$

Sacamos las desviaciones estandares.

$Sx^2 = \frac{\sum(x-x)^2}{n-1} = \frac{3845.13}{23} = 167.17$

$Sx = (160.21)^{½} = 12.92$

$Sy^2 = \frac{\sum(y-y)^2}{n-1} = \frac{14733346}{23} = 640580.26$

$Sy = (613889.41)^{½} = 800.36$

Sacamos la covarianza.

$Sxy = \frac{\sum(x-x)(y-y)}{n-1} = \frac{212167.63}{23} = 9224.37$

Sacamos la correlación.

$r = \frac{Sxy}{SxSy} = \frac{9224.37}{12.92\ (800.36)} = .892$

$b = r \frac{(S_y)}{(S_x)} = 0.892 \frac{(800.36)}{(12.92)} = 55.25$

$a = y - bx = 4365.58 – 55.25 (71.26) = 428.46$

c) Pronostique los costos de distribución de almacén para un mes en el que el número de órdenes es de 4,500.

$Y = a + bx$

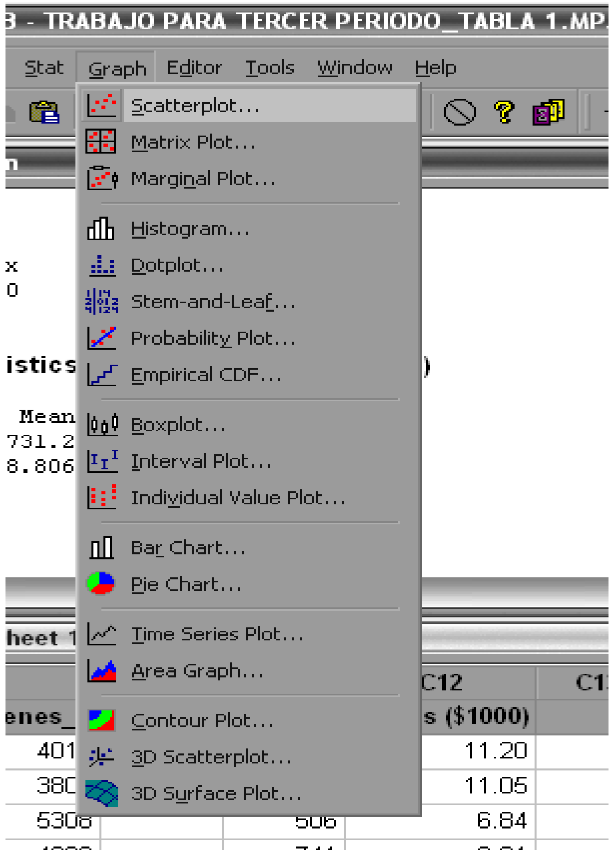
$Y = 428.46 + 55.25(4500) = 249053.46$

4. Suponga que el gerente de una cadena de servicios de entrega de paquetería desea desarrollar un modelo para predecir las ventas semanales (en miles de dólares) para las tiendas individuales basado en el número de clientes que realizan las compras. Se seleccionó una muestra aleatoria entre todas las tiendas de la cadena con los siguientes resultados:

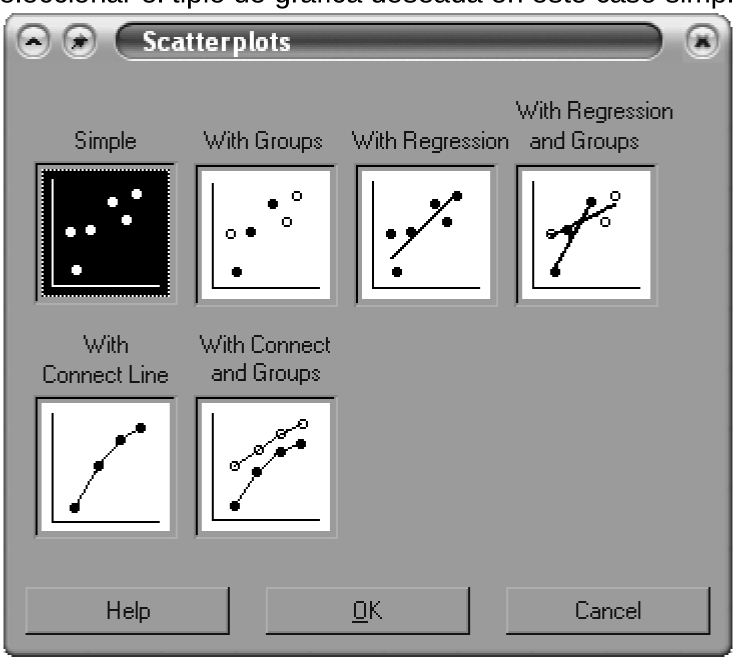
Tienda	Clientes	Ventas ($1000)
1	907	11.20
2	926	11.05
3	506	6.84
4	741	9.21
5	789	9.42
6	889	10.08
7	874	9.45
8	510	6.73
9	529	7.24
10	420	6.12
11	679	7.63
12	872	9.43
13	924	9.46
14	607	7.64
15	452	6.92
16	729	8.95
17	794	9.33
18	844	10.23
19	1010	11.77
20	621	7.41

a) Grafique el diagrama de dispersión.

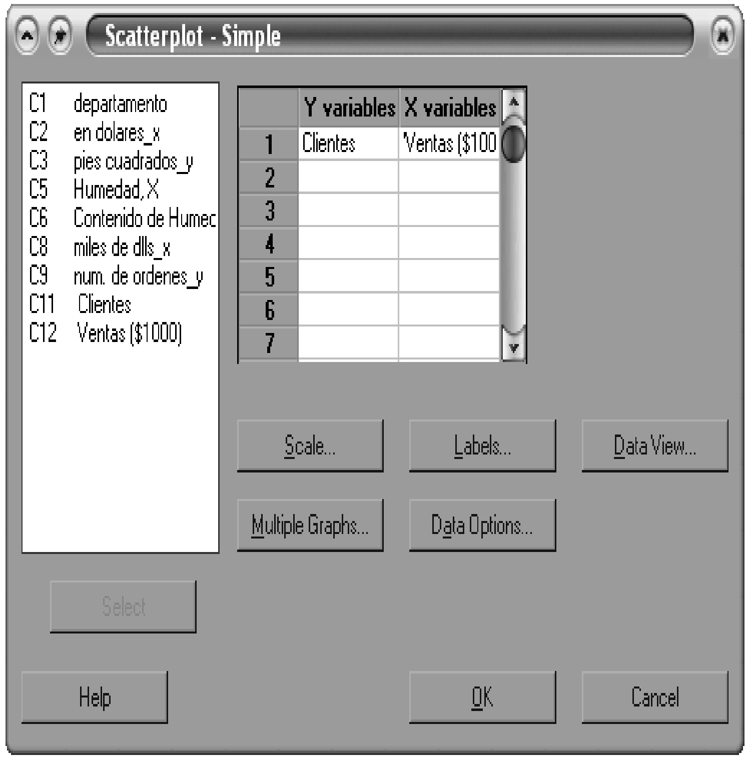
Menú graph seleccionar scatter plot.

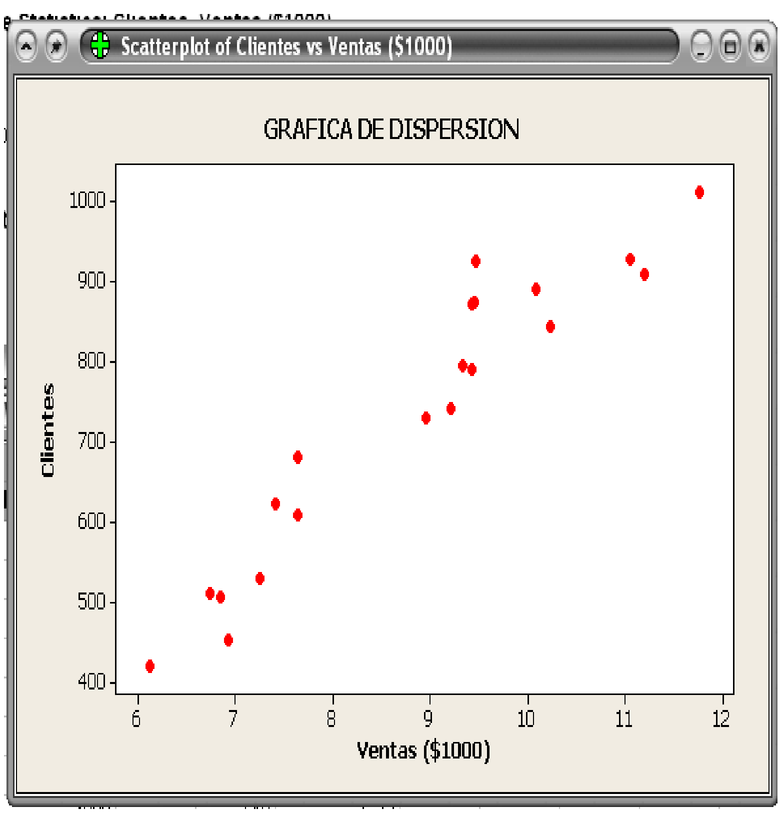


Seleccionar el tipio de grafica deseada en este caso simple.



Seleccionar las variables X y X





b) Obtenga la ecuación que mejor ajuste a los datos.

Clientes	Ventas	(x-x)	(x-x)2	(Y-Y)	(Y-Y)2	(x-x)(Y-Y)
907	11.2	175.8	30905.64	2.394	5.731236	420.8652

926	11.05	194.8	37947.04	2.244	5.035536	437.1312
506	6.84	-225.2	50715.04	-1.966	3.865156	442.7432
741	9.21	9.8	96.04	0.404	0.163216	3.9592
789	9.42	57.8	3340.84	0.614	0.376996	35.4892
889	10.08	157.8	24900.84	1.274	1.623076	201.0372
874	9.45	142.8	20391.84	0.644	0.414736	91.9632
510	6.73	-221.2	48929.44	-2.076	4.309776	459.2112
529	7.24	-202.2	40884.84	-1.566	2.452356	316.6452
420	6.12	-311.2	96845.44	-2.686	7.214596	835.8832
679	7.63	-52.2	2724.84	-1.176	1.382976	61.3872
872	9.43	140.8	19824.64	0.624	0.389376	87.8592
924	9.46	192.8	37171.84	0.654	0.427716	126.0912
607	7.64	-124.2	15425.64	-1.166	1.359556	144.8172
452	6.92	-279.2	77952.64	-1.886	3.556996	526.5712
729	8.95	-2.2	4.84	0.144	0.020736	-0.3168
794	9.33	62.8	3943.84	0.524	0.274576	32.9072
844	10.23	112.8	12723.84	1.424	2.027776	160.6272
1010	11.77	278.8	77729.44	2.964	8.785296	826.3632
621	7.41	-110.2	12144.04	-1.396	1.948816	153.8392
14623	176.11	-1	614602.6	-0.01	51.3605	5365.074

Sacamos las medias

$$X = \frac{\sum x}{n} = \frac{14623}{20} = 731.15$$

$$Y = \frac{\sum y}{n} = \frac{176.11}{20} = 8.80$$

Sacamos las desviaciones estandares.

$$Sx^2 = \frac{\sum(x-x)^2}{n-1} = \frac{614602.6}{19} = 32347.50$$

$Sx = (32347.50)^{½} = 179.85$

$$Sy^2 = \frac{\sum(y-y)^2}{n-1} = \frac{51.36}{19} = 2.70$$

$Sy = (83322.81)^{½} = 1.64$

Sacamos la covarianza.

$$Sxy = \frac{\sum(x\text{-}x)\,(y\text{-}y)}{n-1} = \frac{5365.07}{19} = 282.37$$

Sacamos la correlación.

$$r = \frac{Sxy}{SxSy} = \frac{282.37}{179.85\,(1.64)} = 0.955$$

$$b = r\,\frac{(Sy)}{(Sx)} = 0.955\,\frac{(1.64)}{(179.85)} = 0.008729$$

a = y- bx = 8.80-0.008729(731.15) = 2.41

Ecuación de regresión:

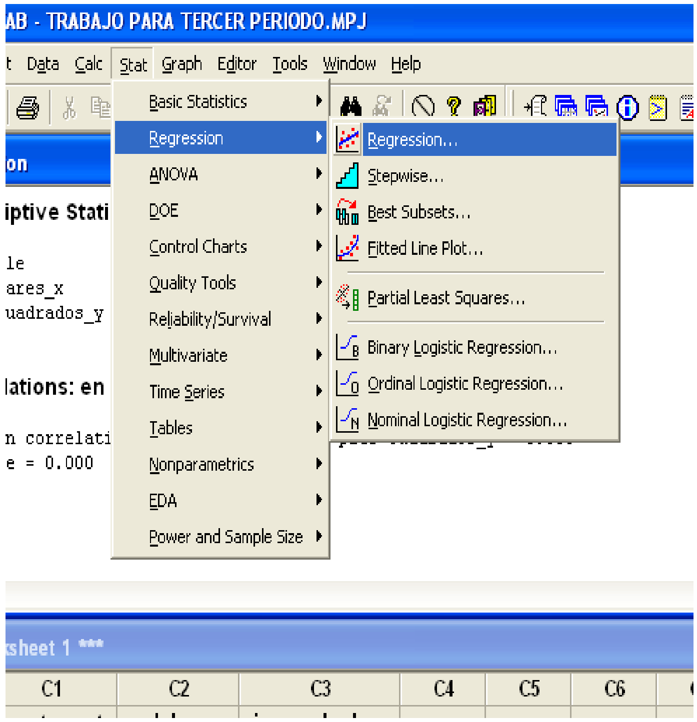
y = a + bx

c) Pronostique las ventas semanales (en miles de dólares) para las tiendas que tienen 600 clientes.

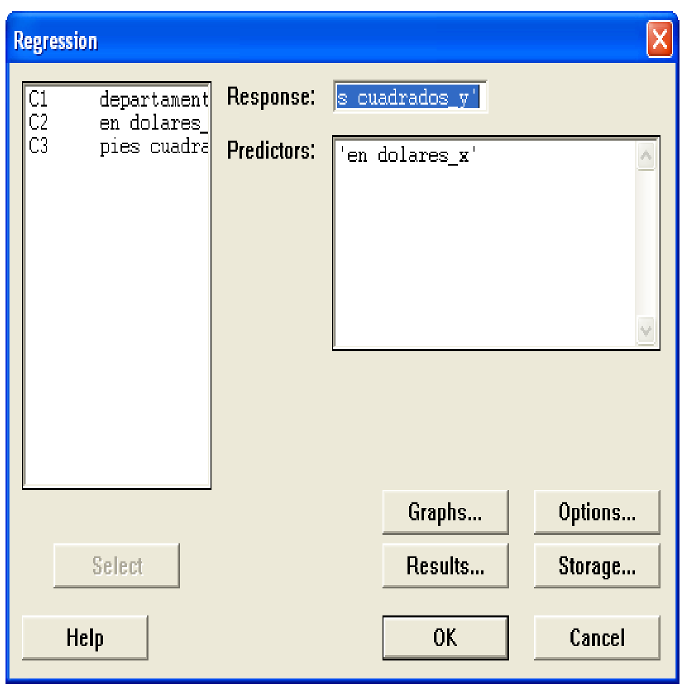
y = 2.14+0.008729(600)
y = 7.37

3. Solución en Minitab (instrucciones para resolver los ejercicios con graficas, en base a uno de los ejercicios teóricos)

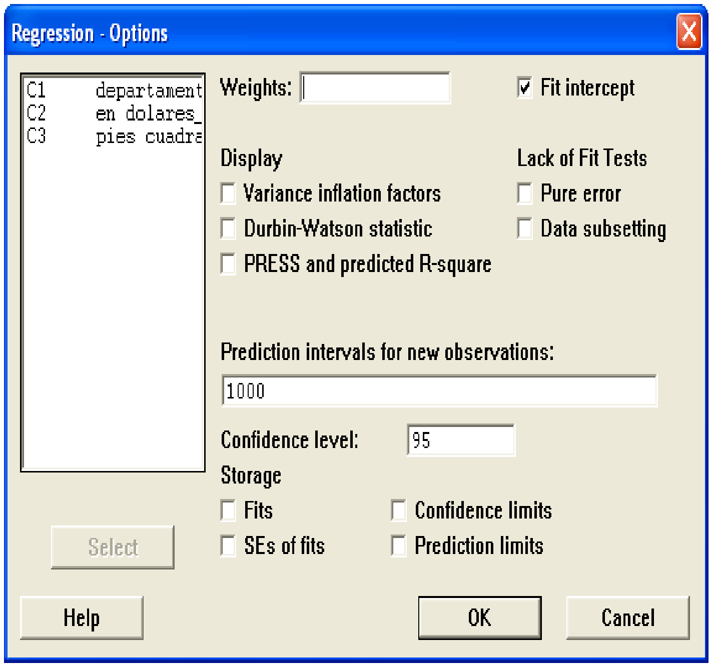
Menú stat, elegir la opción regression, después vuelve a elegir del siguiente menú la opción regresión.



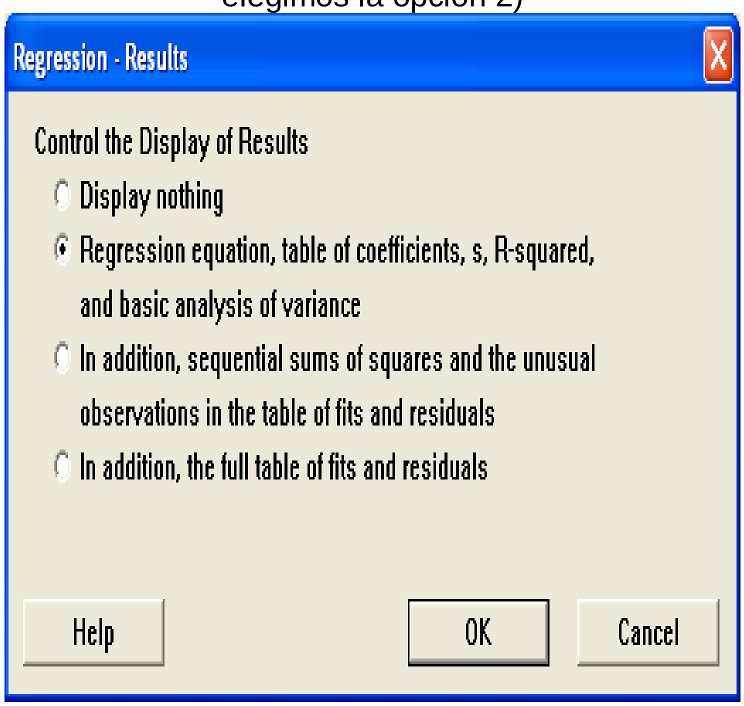
Selecciona las variables a analizar X y Y. Elige option.



Ingresa la predicción en este caso 1000



Elige la opción que quieres que te muestre el software. (En nuestro caso elegimos la opción 2)



Estos son los resultados obtenidos:

Regression Analysis: pies cuadrados_y versus en dolares_x

```
The regression equation is
pies cuadrados_y = 195 + 0.678 en dolares_x


Predictor        Coef  SE Coef     T      P
Constant        194.8    125.4  1.55  0.134
en dolares_x  0.67841  0.08765  7.74  0.000


S = 155.301   R-Sq = 72.3%   R-Sq(adj) = 71.1%


Analysis of Variance

Source          DF       SS       MS      F      P
Regression       1  1445024  1445024  59.91  0.000
Residual Error  23   554723    24118
Total           24  1999747


Predicted Values for New Observations

New
Obs    Fit  SE Fit      95% CI           95% PI
  1  873.2    46.0  (778.1, 968.2)  (538.1, 1208.2)


Values of Predictors for New Observations

New
Obs  en dolares_x
  1          1000
```

4. Conclusiones

Los diagramas de dispersión nos permite establecer una correlación entre dos características de un objeto o muestra, debe existir una relación entre ambos datos de lo contrario la grafica estará errónea.
Gracias a los diagramas de dispersión se puede observar gráficamente los resultados de análisis de proceso.
El empleo adecuado de estas herramientas estadísticas nos llevara a elevar la competitividad de nuestra empresa pues nos permite hacer predicciones a largo plazo aun sin que estas hallan ocurrido.

5. Experiencia de aprendizaje

La estadística es una rama de las matemáticas que se emplea para el estudio de los procesos, así como para su control y es fundamental.

6. **Bibliografía**

http://es.wikipedia.org/wiki/Variable
http://es.wikipedia.org/wiki/Medici%C3%B3n#Medici.C3.B3n
http://www.monografias.com/trabajos11/contrest/contrest.shtml
http://es.wikipedia.org/wiki/Coeficiente_de_correlaci%C3%B3n_de_Pearson