

Capítulo

Análisis de correlación y regresión

13

Secciones

Introducción

13.1 Correlación lineal.

13.2 Regresión lineal.

13.3 Regresión no lineal (funciones intrínsecamente lineales).

13.4 Regresión multilínea.

Antecedentes

Intervalos de confianza

Pruebas de hipótesis

Funciones lineal, exponencial, potencial, logarítmica, recíproca y polinomial.

Objetivos

Proporcionar elementos para

- Construir e interpretar diagramas de dispersión
- Calcular e interpretar, en el contexto propio, el coeficiente de correlación r de Pearson
- Hacer e interpretar inferencias sobre el coeficiente de correlación r de Pearson entre dos variables
- Calcular e interpretar la recta de regresión por mínimos cuadrados para una muestra de puntos dados
- Hacer inferencias sobre la estimación y los parámetros de la recta de regresión.
- Identificar y transformar en lineales las funciones intrínsecamente lineales más comunes.
- Calcular e interpretar la regresión multilínea.

Introducción

En los cursos de geometría, álgebra y otros que el lector haya tomado, generalmente la relación entre las variables es de tipo *determinista*; es decir, dado un valor de una de las variables, el valor de la otra variable se determina *automáticamente* y, podría decirse, *sin error*. Ejemplos típicos son las fórmulas geométricas y expresiones del tipo $C = 400 + 0.10k$, donde C es el costo de renta de un automóvil y k los kilómetros recorridos.

En estadística estamos interesados en relaciones entre variables aleatorias y, por lo tanto, una de las variables no queda determinada por completo por otra u otras variables. Así, expresiones como $P = 5E - 190$ que dan la relación entre el peso P de un hombre (en libras) y

su estatura E (en pulgadas) para una cierta población, son relaciones estadísticas en donde se espera obtener sólo *estimaciones*.

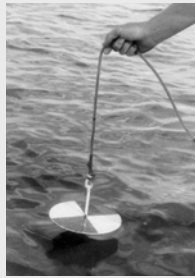
Las relaciones estadísticas se obtienen mediante una primera fase de exploración conocida como *análisis de correlación*. Consiste en analizar los datos *muestrales* para saber el *grado de asociación o correlación* entre dos o más variables de una población. El grado de correlación se expresa como un número comprendido entre -1 y +1 y se le conoce como *coeficiente de correlación*.

Como corresponde a un estudio exploratorio, el análisis de correlación no es un fin en sí mismo sino que su objetivo es establecer la pertinencia de la segunda fase o *análisis de regresión*. Este da lugar a una función $y = f(x)$ que describe *estadísticamente* la asociación o relación entre las variables en estudio y, por tanto, su fin no es calcular sin error sino obtener *predicciones* del valor de una variable, para un valor dado de la otra variable.

Debido a que los cálculos para el coeficiente de correlación y los parámetros que definen la función se basan en una muestra aleatoria, se espera que varíen de una muestra a otra (tal como la media varía de una muestra a otra). Esto plantea preguntas de *significancia* del coeficiente de correlación, de los parámetros de la función y de los valores de predicción obtenidos con ella. Tales preguntas son respondidas mediante intervalos de confianza y pruebas de hipótesis; esto es, mediante análisis inferencial.

Ventana al conocimiento 1

Un **disco Secchi** es un disco de 8 pulgadas con cuadrantes pintados de blanco y negro alternativamente. Se ata a una cuerda marcada en centímetros. Se sumerge en el agua (de lagos, ríos o mares) hasta no ser visible al observador. La lectura observada en la cuerda es conocida como profundidad Secchi y es una medida de la transparencia del agua. La transparencia del agua se ve afectada por el color, las algas y sedimentos suspendidos. Las algas son pequeñas plantas acuáticas cuya abundancia está asociada a la cantidad de nutrientes, especialmente fósforo y nitrógeno. Los lagos y los ríos se monitorean regularmente, tomando muestras en puntos elegidos aleatoriamente, para establecer la calidad del agua. En cada una de las muestras se determina la profundidad Secchi y algunos parámetros como *clorofila a*, *nitrógeno total*, *carbón orgánico*, *fósforo total*, *sólidos totales suspendidos*, *conductividad específica* y *densidad total*. Los resultados (variables) así obtenidos son de naturaleza aleatoria.



<http://dipin.kent.edu/images/Secchi%20Disk.jpg>

El análisis de regresión puede también dar lugar a una función del tipo $y = f(x, z, v)$ para describir la relación entre varias variables (ver sección 13.4)

13.1 Correlación lineal

Se empieza el estudio de correlación con el caso más sencillo, el de la *correlación lineal* entre dos variables aleatorias cuantitativas X y Y (en adelante se manejarán como x y y respectivamente, por ser la notación más común en la literatura). Los datos muestrales suelen reducirse a parejas y la forma estándar para designarlas es:

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

Se tiene entonces una muestra de n datos (parejas) donde x_1 denota el primer valor de la variable aleatoria x y y_1 denota el primer valor de la variable aleatoria y . En correlación es indistinto a qué variable se le llame x y a qué variable se le llame y ; de hecho, si hay correlación entre x y y en ese orden, la hay también entre y y x en ese orden.

No obstante lo anterior y dado que generalmente el proceso no termina en la correlación sino que se pasa al análisis de regresión, se designará como x a aquella variable que pueda verse como un predictor potencial y cuyo valor pueda algunas veces ser seleccionado por el investigador. En cierto sentido podría verse como una *variable explicativa*. La otra variable sería denotada como y y sería aquella variable que pueda verse como *respuesta*. Algunos autores, tomando los nombres de las relaciones deterministas, usan los términos variable independiente y dependiente para x y y respectivamente.

Dicho lo anterior y con el fin de contextualizar el estudio de la correlación lineal, se recurre a una situación real.

Situación de estudio: cigarrillos

La Comisión de Comercio Federal de Estados Unidos evalúa anualmente distintas marcas de cigarrillos de acuerdo con su contenido de alquitrán, nicotina y monóxido de carbono. La Asociación de Médicos de Estados Unidos juzga peligrosas cada una de estas sustancias para la salud del fumador. Estudios anteriores han demostrado que un aumento en el contenido de alquitrán y nicotina de un cigarrillo está acompañado de un incremento en el monóxido de carbono emitido en el humo del cigarrillo. La tabla siguiente muestra los valores para 25 marcas de cigarrillos comunes en Estados Unidos.

Tabla 13.1 Contenido de sustancias peligrosas para la salud en cigarrillos.

Marca	Peso (g)	Alquitrán (mg)	Nicotina (mg)	CO (mg)
Alpine	0.9853	14.1	0.86	13.6
Benson&Hedges	1.0938	16.0	1.06	16.6
BullDurham	1.1650	29.8	2.03	23.5
CamelLights	0.9280	8.0	0.67	10.2
Carlton	0.9462	4.1	0.40	5.4
Chesterfield	0.8885	15.0	1.04	15.0
GoldenLights	1.0267	8.8	0.76	9.0
Kent	0.9225	12.4	0.95	12.3
Kool	0.9372	16.6	1.12	16.3
L&M	0.8858	14.9	1.02	15.4
LarkLights	0.9643	13.7	1.01	13.0
Marlboro	0.9316	15.1	0.90	14.4
Merit	0.9705	7.8	0.57	10.0
MultiFilter	1.1240	11.4	0.78	10.2
NewportLights	0.8517	9.0	0.74	9.5
Now	0.7851	1.0	0.13	1.5
OldGold	0.9186	17.0	1.26	18.5
PallMallLight	1.0395	12.8	1.08	12.6
Raleigh	0.9573	15.8	0.96	17.5
SalemUltra	0.9106	4.5	0.42	4.9
Tareyton	1.0070	14.5	1.01	15.9
True	0.9806	7.3	0.61	8.5
ViceroyRichLight	0.9693	8.6	0.69	10.6
VirginiaSlims	0.9496	15.2	1.02	13.9
WinstonLights	1.1184	12.0	.82	14.9

Fuente: http://www.amstat.org/publications/jse/jse_data_archive.html

Un estudio de correlación empieza seleccionando las variables de interés. Así, si se desea analizar la relación entre los miligramos de alquitrán y los miligramos de CO emitidos por los cigarrillos, puede tomarse los miligramos de alquitrán como la variable predictiva, x y los miligramos de CO como la variable respuesta, y . Ordenando los datos respecto a x y separando las columnas relevantes al estudio del resto de la información, se obtiene la tabla 13.2.

Tabla 13.2 Datos ordenados respecto al Alquitrán

Marca	Alquitrán: x (mg)	CO: y (mg)
Now	1	1.5
Carlton	4.1	5.4
SalemUltra	4.5	4.9
True	7.3	8.5
Merit	7.8	10
CamelLights	8	10.2
ViceroyRichLight	8.6	10.6
GoldenLights	8.8	9
NewportLights	9	9.5
MultiFilter	11.4	10.2
WinstonLights	12	14.9
Kent	12.4	12.3
PallMallLight	12.8	12.6
LarkLights	13.7	13
Alpine	14.1	13.6
Tareyton	14.5	15.9
L&M	14.9	15.4
Chesterfield	15	15
Marlboro	15.1	14.4
VirginiaSlims	15.2	13.9
Raleigh	15.8	17.5
Benson&Hedges	16	16.6
Kool	16.6	16.3
OldGold	17	18.5
BullDurham	29.8	23.5

Ordenados los datos, el recorrido simultáneo de las columnas x y y de arriba abajo puede en algunas ocasiones proporcionar información preliminar. En la tabla 13.2, por ejemplo, se aprecia una relación entre ambas variables que se expresa así:

Relación observada entre x y y :

Al aumentar x “aumenta” y

Debe precisarse, sin embargo, que a diferencia de x , el aumento de y no es estricto; en algunos casos, al pasar de una marca a otra, el CO disminuye para después aumentar. La expresión *al aumentar x “aumenta” y* describe más bien un *patrón* de comportamiento *global* de las parejas en estudio.

El siguiente paso consiste en graficar las parejas de datos $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ en un sistema cartesiano, resultando un diagrama de puntos bivariable conocido como *diagrama de dispersión*. El diagrama de dispersión correspondiente a las parejas de la tabla 13.2 se da en la figura 13.1.

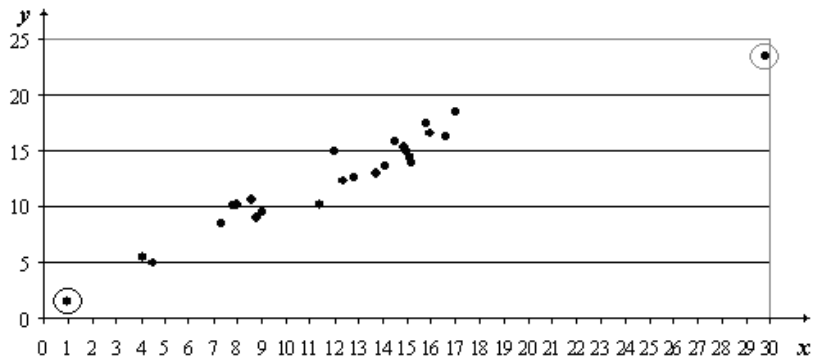


Figura 13.1 Diagrama de dispersión del Alquitrán-CO.

El diagrama de dispersión permite *visualizar* las parejas y establecer algún patrón de comportamiento gráfico. En la figura 13.1 se confirma la relación *al aumentar x “aumenta” y*; además, se resaltan algunos aspectos de interés, como el que los valores extremos (encerrados en círculos) pudieran ser atípicos, dada la dimensión de los huecos entre éstos y los racimos más cercanos (ver capítulo 3 del libro para la definición y empleo de los términos racimos, huecos y valores atípicos). No obstante el valor de la información anterior, hay, sin embargo, un aspecto visual importante:

La disposición de los puntos sigue un patrón gráfico “lineal”

Los diagramas de dispersión de la figura 13.2 pueden ser descritos también con la expresión *al aumentar x “aumenta” y*; no obstante, su patrón gráfico no corresponde al tipo lineal sino al de otras curvas.

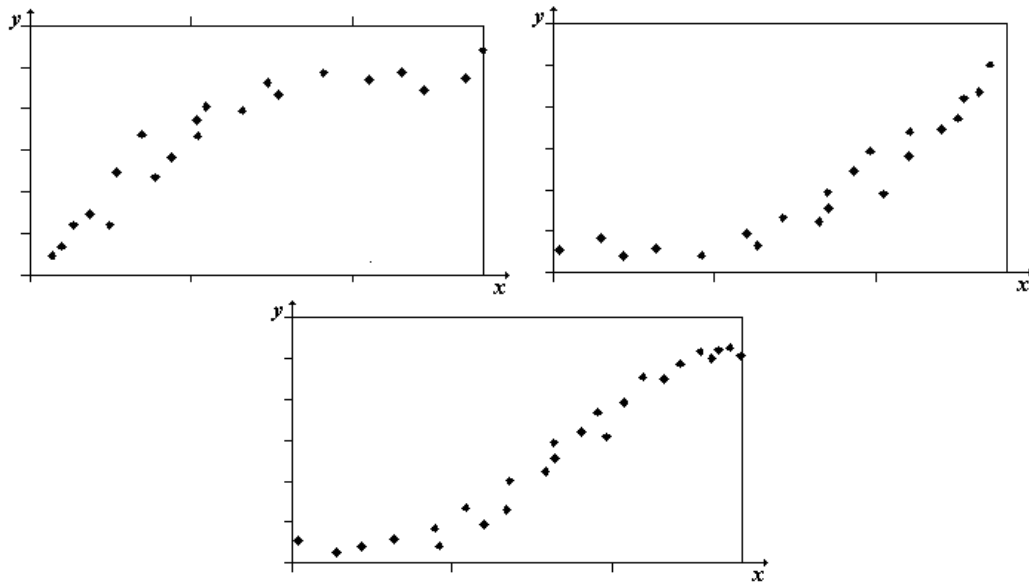


Figura 13.2 Gráficas de dispersión no lineales en las que al aumentar x “aumenta” y

Un comportamiento global descrito por la expresión *al aumentar x “aumenta”* y (en lo que resta de esta sección nos referiremos al tipo lineal) suele describirse como una *correlación o asociación positiva* de y respecto a x (vea figura 13.3 *a*). En caso contrario, esto es, si al aumentar x disminuye globalmente y siguiendo un patrón gráfico lineal (vea figura 13.3 *b*), se dice que hay una *correlación o asociación negativa* de y respecto a x .

Por otro lado, si el diagrama de dispersión es del tipo mostrado en el inciso *c*) de la figura 13.3, el recorrido de izquierda a derecha en el eje x no muestra asociación o relación de ningún tipo entre los valores de x y y ya que, al aumentar x igualmente aumenta y disminuye y . Un diagrama de estas características, es indicativo de que **no** hay relación (correlación) entre las variables en estudio.

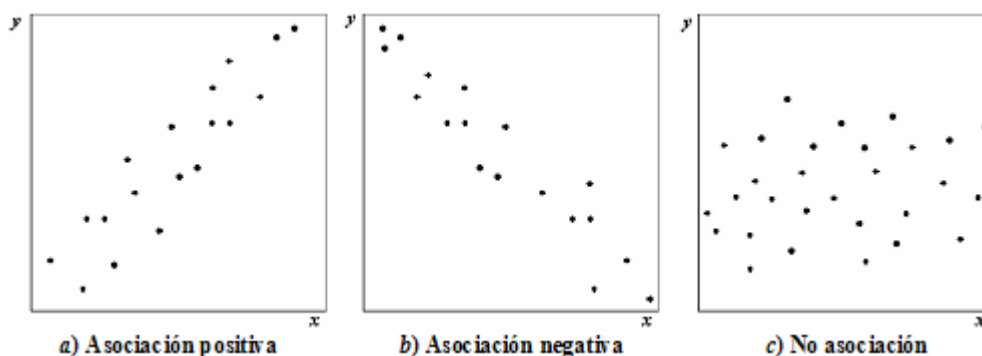


Figura 13.3 Distintos tipos de correlación o asociación de datos

En el caso de variables aleatorias es poco probable tener una correlación lineal perfecta; sin embargo, para fines de análisis, resulta útil e importante considerarla. En los incisos *a*) y *b*) de la figura 13.4 se muestra una correlación lineal positiva y una negativa perfectas respectivamente. Como se observa, los puntos están distribuidos a lo largo de líneas rectas.

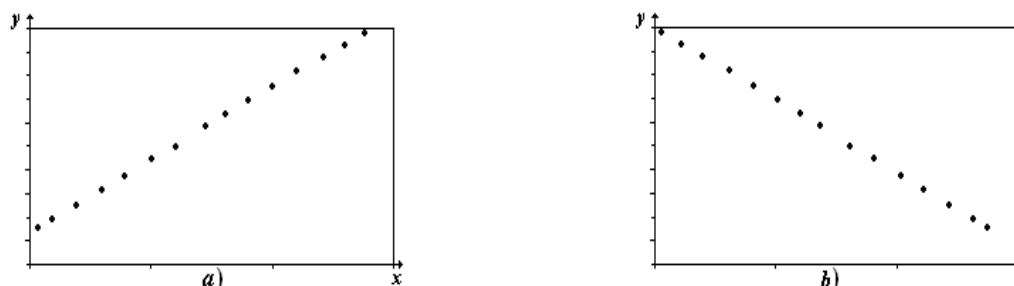


Figura 13.4 Correlación lineal positiva y negativa perfectas.

La “no relación” puede también manejarse mediante una serie de puntos a lo largo de una línea recta horizontal (ver figura 13.5). El significado algebraico de esto es que y es independiente de x o, en términos estadísticos, que no hay correlación entre x y y .

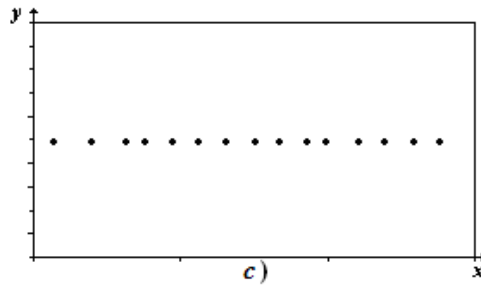


Figura 13.5 No asociación

Una consideración importante que se desprende de esto es que:

Las relaciones deterministas vistas en otros cursos, pueden verse como correlaciones perfectas y por tanto como un caso particular de las relaciones estadísticas.

Las descripciones de correlación lineal vistas hasta ahora son de tipo cualitativo. Para avanzar a una descripción cuantitativa se procede a dividir el diagrama de dispersión en cuatro regiones, dibujando líneas paralelas a los ejes por un *punto central*. El punto central puede ser el de las medianas o el de las medias; en este capítulo se considerará el punto central correspondiente a las medias (\bar{x}, \bar{y}) , llamado también *centroide* (en el capítulo 3 del libro podrá encontrarse un análisis detallado empleando como punto central las medianas (\tilde{x}, \tilde{y})).

Calculando las medias de las columnas x y y de la tabla 13.2 se obtiene $\bar{x} = 12.216$ y $\bar{y} = 12.528$. Colocando el punto central $(12.2, 12.5)$ en el diagrama de la figura 13.1 y trazando paralelas a los ejes por ese punto se llega a la figura 13.6.

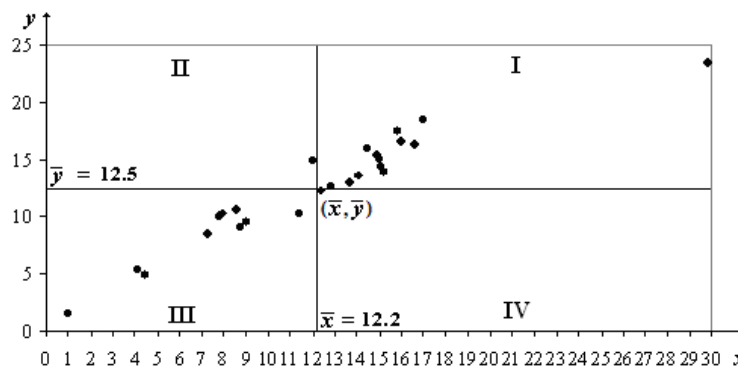


Figura 13.6 División del diagrama de dispersión en cuatro regiones.

Cualquier punto ubicado en la región I o III apoya una correlación positiva; cualquier punto en la región II o IV apoya en cambio una correlación negativa. Tomando en cuenta que se trabaja con muestras de n puntos o datos, puede llamarse $n(I)$ al número de puntos en la región

I y de igual forma $n(\text{II})$, $n(\text{III})$ y $n(\text{IV})$ el número de puntos de las regiones II, III y IV respectivamente.

Con estos elementos, se puede definir un número c que permita establecer tipo y grado de correlación o asociación entre las variables en estudio, de la siguiente manera (Peter Holmes, Correlation: From Picture to Formula, *Teaching Statistics* volume 23, Num. 3, Autumn 2001 p p. 67-70):

$$c = \frac{n(\text{I}) + n(\text{III}) - n(\text{II}) - n(\text{IV})}{n} \quad (13.1)$$

A la clase de números a que pertenece c se les conoce genéricamente como *coeficientes de correlación*. Analizando la definición 13.1 pueden verse algunas de las ideas generales con que se construyen tales coeficientes.

Propiedades del coeficiente de correlación c .

- a) Si todos los puntos están en I y III, entonces $c = 1$.
- b) Si todos los puntos están en II y IV, entonces $c = -1$.
- c) Si los puntos están repartidos equitativamente en las cuatro regiones, entonces $c = 0$.
- d) Si todos los puntos están en tres o cuatro regiones, entonces c estará entre -1 y $+1$: si los puntos están predominantemente en I y III, entonces c será positivo, pero si los puntos están predominantemente en II y IV, entonces c será negativo.

Se continúa el análisis de la situación de los cigarrillos, calculando el coeficiente de correlación c empleando la figura 13.6 o, en caso de duda, la tabla 13.2:

$$n = 25; n(\text{I}) = 13; n(\text{II}) = 1; n(\text{III}) = 10; n(\text{IV}) = 1$$

$$c = \frac{13 + 10 - 1 - 1}{25} = 0.84$$

El signo positivo de c (implícito en 0.84), indica que los puntos están ubicados predominantemente en el primer y tercer cuadrantes y por tanto que se tiene una *correlación o asociación positiva* entre x y y .

Considerando que los valores extremos de c son -1 y $+1$, la magnitud (valor absoluto) de c puede usarse como un indicador del grado o fuerza de la correlación entre las variables: el grado es *fuerte* entre más cercana se encuentre la magnitud de c a 1 y *débil* entre más cercana se encuentre a cero. Podría decirse entonces que la magnitud de c para los cigarrillos indica un grado de correlación *fuerte* entre las variables. Resumiendo:

De acuerdo al valor numérico del coeficiente de correlación c , hay una correlación *positiva fuerte* entre los miligramos de alquitrán y los miligramos de CO en los cigarrillos.

Actividad 13.1 Empleando la expresión 13.1 demostrar que $c = 0$ para el caso de una serie de puntos a lo largo de una línea recta horizontal

Con el fin de avanzar en el estudio de los coeficientes de correlación se recurre a una situación distinta a la vista pero también en un contexto real.

Situación de estudio: maratón

Un **maratón** es una prueba atlética de resistencia con categoría olímpica que consiste en correr a pie la distancia de 42195 metros. Forma parte del programa olímpico en la categoría masculina desde 1896, y en 1984 se incorporó la categoría femenina.

Muchas ciudades importantes del mundo organizan anualmente maratones. Uno de los más prestigiados es el de Nueva York. Se listan a continuación los tiempos de los y las ganadoras del Maratón de Nueva York y las temperaturas medias registradas durante el periodo 1978-1998.

Tabla 13.3 Tiempos de los ganadores del maratón de Nueva York

Año	T (°F)	t Hombres (min)	t Mujeres (min)
1978	75	132.200	152.500
1979	80	131.700	147.550
1980	50	129.683	145.700
1981	54	128.217	145.483
1982	52	129.483	147.233
1983	59	128.983	147.000
1984	79	134.883	149.500
1985	72	131.567	148.567
1986	65	131.100	148.100
1987	64	131.017	150.283
1988	67	128.333	148.117
1989	56	128.017	145.500
1990	73	132.650	150.750
1991	57	129.467	147.533
1992	51	129.483	144.667
1993	73	130.067	146.400
1994	70	131.350	147.617
1995	62	131.000	148.100
1996	49	129.900	148.300
1997	61	128.200	148.717
1998	55	128.750	145.283

Fuente: The Effects of Temperature on Marathon Runner's Performance de David Martin y John Buoncristiani (Chance, vol. 12, num 4).

El origen de la palabra maratón se encuentra en la gesta del soldado griego Filípides, quien en el año 490 a. C. murió de fatiga tras haber corrido unos 40 km desde Maratón hasta Atenas para anunciar la victoria sobre el ejército Persa. En honor a la hazaña de Filípides se creó una competencia con el nombre de "maratón", que fue incluida en los juegos de 1896 de Atenas.

Resulta plausible considerar que pudiera haber una relación entre las temperaturas (variable predictiva) en que se realiza la prueba y los tiempos de los ganadores (variable respuesta). Para analizar esta hipótesis puede empezarse ordenando los datos de acuerdo a las

temperaturas. Se omiten los tiempos de los hombres dejando solamente la información relevante al estudio (ver tabla 13.4).

Tabla 13.4 Tabla ordenada de menor a mayor considerando la temperatura

Año	T (°F)	t Mujeres (min)
1996	49	148.300
1980	50	145.700
1992	51	144.667
1982	52	147.233
1981	54	145.483
1998	55	145.283
1989	56	145.500
1991	57	147.533
1983	59	147.000
1997	61	148.717
1995	62	148.100
1987	64	150.283
1986	65	148.100
1988	67	148.117
1994	70	147.617
1985	72	148.567
1990	73	150.750
1993	73	146.400
1978	75	152.500
1984	79	149.500
1979	80	147.550

Observe que la temperatura media pueda repetirse en algunas ocasiones (*i.e.* 73°F) y que, sin embargo, le correspondan tiempos distintos. Esto es común en parejas de datos estadísticos.

Al recorrer simultáneamente las columnas de T y t de arriba abajo, no logra apreciarse una asociación entre las variables. Construyendo el diagrama de dispersión con las temperaturas en el eje horizontal y los tiempos de las ganadoras en el eje vertical, se llega a la figura 13.7.

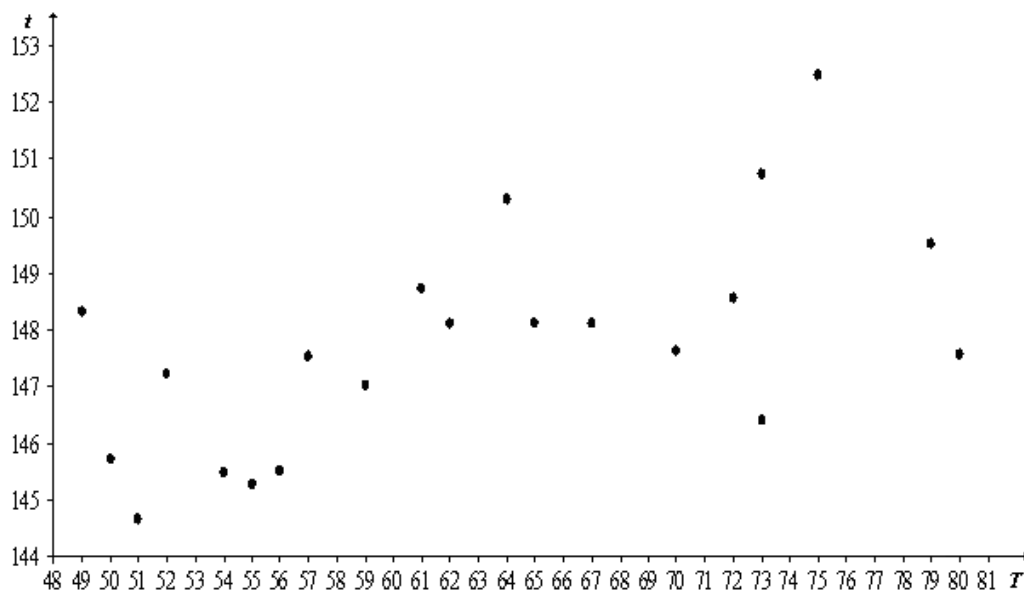


Figura 13.7 Diagrama de dispersión temperatura vs. tiempo

El diagrama tampoco es muy revelador del tipo de asociación, por lo que se obtiene el punto central y se trazan por éste las líneas de división.

$$\text{Punto central: } \bar{T} = 63.048; \bar{t} = 147.757 .$$

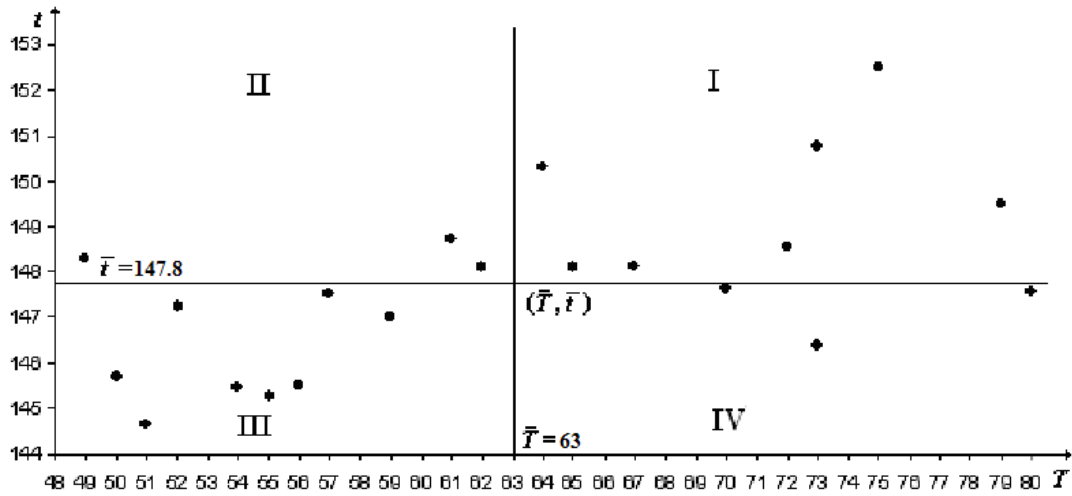


Figura 13.8 Diagrama de dispersión con líneas de división

La división permite distinguir que la distribución de los puntos se da predominantemente en las regiones I y III y por tanto considerar una *correlación positiva* entre las variables. Los puntos, sin embargo, se encuentran muy dispersos respecto a lo que pudiera ser un patrón gráfico lineal, por lo que se esperaría un grado de asociación *débil*. Con el fin de tener medidas numéricas se calcula el coeficiente de correlación c :

$$n = 21; n(\text{I}) = 7; n(\text{II}) = 3; n(\text{III}) = 8; n(\text{IV}) = 3$$

$$c = \frac{7+8-3-3}{21} = 0.42857$$

El signo positivo de c indica la preponderancia de los puntos en las regiones I y III confirmando la asociación positiva; la magnitud de c (0.42857), sin embargo, refleja un grado de correlación débil ya que se encuentra más bien cercana a cero.

Pudiera pensarse que la magnitud de c indica la dispersión de los puntos de un diagrama, sin embargo, puede no resultar así en todos los casos, ya que por ejemplo en las dos gráficas de la figura 13.9 se obtiene $c = 1$, el grado máximo de correlación. Esta falla de la magnitud del coeficiente c a diferenciar el grado de dispersión en ambos diagramas, sugiere construir un coeficiente de correlación que, por ejemplo, deje el grado máximo de asociación exclusivamente a los casos en que se tienen las correlaciones lineales positiva y negativa perfectas. Asimismo, que refleje que el diagrama de dispersión del inciso b) corresponde a una

correlación de mayor grado que la correlación que guardan los puntos del diagrama del inciso a).

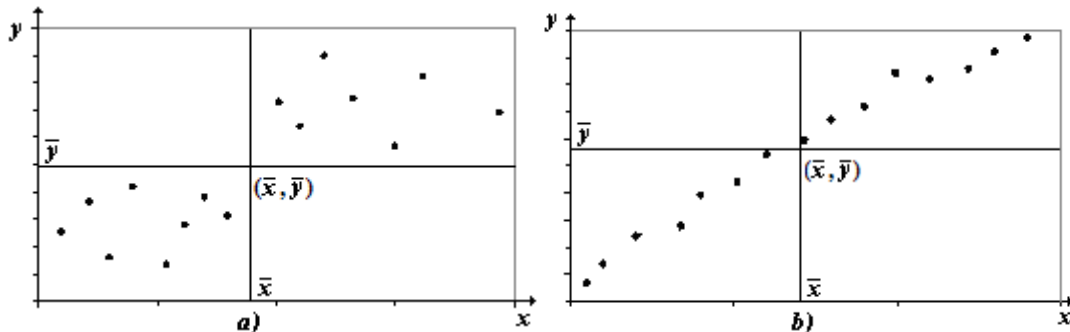


Figura 13.9 Correlación positiva débil y fuerte

El científico Inglés Karl Pearson desarrolló un coeficiente de correlación que cumple con los requisitos mencionados y es uno de los más ampliamente usados en ingeniería y ciencias.

Coefficiente de correlación de Pearson

El desarrollo del nuevo coeficiente de correlación puede plantearse asignándole peso a los puntos (x_i, y_i) en función de su ubicación respecto a las líneas de división que se trazan por el centroide. A medida que el punto (x_i, y_i) se aleja de las líneas, su peso sería mayor (ver figura 13.10). Después de todo, los puntos cerca de las líneas podrían cambiar de signo fácilmente (recuerde que son valores aleatorios), mientras que los puntos más alejados de las líneas pueden establecer con mayor fuerza la correlación.



Karl Pearson
(Londres 27 de marzo de 1857-
Londres, 27 de abril de 1936) fue un prominente científico, matemático, historiador y pensador británico, que estableció la disciplina de la *estadística matemática*. Desarrolló una intensa investigación sobre la aplicación de los métodos estadísticos en la biología y fue el fundador de la bioestadística.

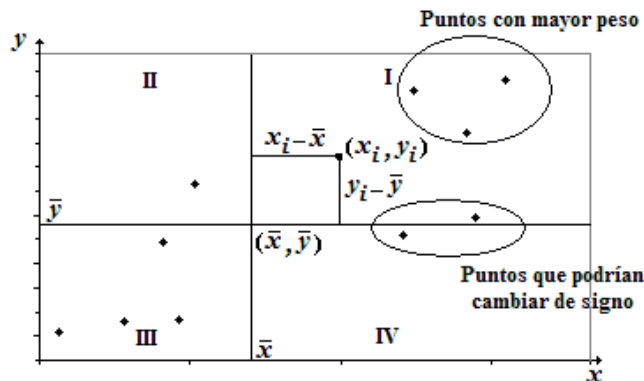


Figura 13.10 Pesos de los puntos del diagrama de dispersión.

Para la asignación del peso a un punto (x_i, y_i) , se empieza calculando las diferencias $x_i - \bar{x}$ y $y_i - \bar{y}$ (vea figura 13.10). La magnitud del producto $(x_i - \bar{x})(y_i - \bar{y})$ da una medida de la cercanía o lejanía de (x_i, y_i) a las líneas de división.

Los signos de $x_i - \bar{x}$ y $y_i - \bar{y}$ dependen de la región en que se encuentre (x_i, y_i) (vea Tabla 13.5). El signo del producto $(x_i - \bar{x})(y_i - \bar{y})$ es positivo para puntos (x_i, y_i) de las regiones I y III, reforzando la idea de asociación positiva. El signo negativo del producto para puntos de las regiones II y IV haría lo propio con la asociación negativa.

Tabla 13.5 Signos de las diferencias y del producto

Diferencias y producto	I	II	III	IV
$x_i - \bar{x}$	+	+	-	-
$y_i - \bar{y}$	+	-	-	+
$(x_i - \bar{x})(y_i - \bar{y})$	+	-	+	-

Por tanto, se obtiene un primer acercamiento al coeficiente de correlación buscado, r en adelante, sumando los productos $(x_i - \bar{x})(y_i - \bar{y})$ correspondiente a los n puntos de la muestra:

$$r = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

El resultado es un número real cuyo signo indicaría la preponderancia de los puntos de las regiones I y III o la preponderancia de los puntos de las regiones II y IV y, como se desea, una magnitud indicativa de la fuerza de correlación entre las variables x y y .

La suma de los productos, sin embargo, no daría un valor entre -1 y +1, ya que dependería de:

- a) La magnitud y unidades de las variables x y y .
- b) El número n de puntos de la muestra.

Para ver mejor a qué se refiere el inciso a), se calcula r para los cigarrillos, resultando $617.0988mg^2$ (se sugiere verificar). La magnitud resultante no sólo *no* está entre -1 y +1 sino que además pudo resultar mayor si se hubiesen usado gramos o más pequeña si se hubiesen usado miligramos. Una condición razonable a imponer es que r no dependa de las unidades utilizadas para medir las variables.

Lo anterior puede resolverse expresando cada diferencia en términos de desviaciones estándar: $\frac{x_i - \bar{x}}{s_x}$ y $\frac{y_i - \bar{y}}{s_y}$, donde s_x y s_y son las desviaciones estándar de los valores de x y de y , respectivamente. Como s_x y s_y tienen las mismas unidades que sus variables asociadas,

se elimina el aspecto unidades y, adicionalmente, se estandariza cada diferencia. Con esto, r toma la forma siguiente:

$$r = \sum_{i=1}^n \frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

No obstante la estandarización, r sigue dependiendo del número n de puntos. Así, en una asociación positiva, si los puntos fueran duplicados sin cambio en la naturaleza de la asociación, el valor de r aproximadamente se duplicaría. En los cigarrillos, por ejemplo, los primeros 12 puntos de la tabla 13.2, calculando las medias y desviaciones estándar correspondientes, dan un valor de r igual a 10.38, mientras que utilizando todos los datos se obtiene r igual a 22.98. Para solucionar esto se divide entre $n-1$ (las razones para dividir entre $n-1$ y no entre n son las mismas que en el cálculo de la desviación estándar). Con esto se obtiene un tipo de promedio que, como se verá en el ejemplo 13.1, toma los valores extremos -1 y 1 y cumple las propiedades requeridas. La expresión para r toma la forma:

$$r = \frac{1}{n-1} \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} \quad (13.2)$$

Con el fin de llegar a una expresión equivalente a la 13.2 que resulte práctica para los cálculos directos o su programación, se desarrolla algebraicamente el numerador para llegar a:

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

Sustituyendo en el denominador las desviaciones estándar por sus expresiones prácticas

correspondientes $s_x = \sqrt{\frac{\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2}{n-1}}$ y $s_y = \sqrt{\frac{\sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2}{n-1}}$ y simplificando se

tiene finalmente:

$$r = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \sqrt{n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i \right)^2}} \quad (13.3)$$

La expresión 13.3 (o equivalentes) se conoce como el *coeficiente de correlación lineal producto momentos de Pearson*.

Actividad 13.2
Justifica la equivalencia de $\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$ y $\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$

Por último y con el fin de emplear en forma práctica la magnitud de r como un indicador del grado de correlación o asociación entre las variables, se da la tabla 13.6.

Tabla 13.6 Correlación lineal entre dos variables

Valores de r	Tipo y grado de correlación
-1	Negativa perfecta
$-1 < r \leq -0.8$	Negativa fuerte
$-0.8 < r < -0.5$	Negativa moderada
$-0.5 \leq r < 0$	Negativa débil
0	No existe
$0 < r \leq 0.5$	Positiva débil
$0.5 < r < 0.8$	Positiva moderada
$0.8 \leq r < 1$	Positiva fuerte
1	Positiva perfecta

Ejemplo 13.1 Demostrar que el coeficiente de correlación de Pearson toma los valores extremos de $+1$ y -1 en los casos de correlación lineal positiva y negativa perfectas, respectivamente.

Solución. Para llevar a cabo la demostración primero se sustituye en la expresión 13.2 a s_x por

$$\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \text{ y a } s_y \text{ por } \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} :$$

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

Simplificando:

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

Luego se considera el hecho de que en la asociación perfecta positiva y negativa todos los puntos quedan en una línea recta $y = mx + b$. De la misma forma el punto central (\bar{x}, \bar{y}) queda sobre esa línea recta (ver problema 13.10), por lo que $\bar{y} = m\bar{x} + b$. Sustituyendo en la expresión simplificada de r a y_i por $mx_i + b$ y a \bar{y} por $m\bar{x} + b$ y reduciendo se obtiene:

$$r = \frac{m \sum_{i=1}^n (x_i - \bar{x})^2}{\sqrt{m^2} \sum_{i=1}^n (x_i - \bar{x})^2} = \frac{m}{\sqrt{m^2}}$$

Como $\sqrt{m^2} = |m|$, $r = +1$ si m es negativa y $r = -1$ si m es positiva, quedando con esto la demostración terminada.

Se dan a continuación las propiedades de coeficiente de correlación r .

Propiedades del coeficiente de correlación r de Pearson.

- El valor de r es independiente de las unidades en que se midan x y y .
- $r = 1$ si y sólo si todos los pares de puntos de la muestra están en una recta con pendiente positiva y $r = -1$ si y sólo si todos los pares de puntos de la muestra están en una recta con pendiente negativa.
- El rango de valores de r está dado por el intervalo $-1 \leq r \leq 1$.
- Simetría: El valor de r no depende de cuál de las dos variables bajo estudio se designe como x y cuál como y .
- r mide la fuerza de una relación lineal. No está diseñado para medir la fuerza de una relación que no sea lineal

La demostración de la propiedad del inciso c) queda fuera de los objetivos del libro. La propiedad de simetría del inciso d) se discute en la actividad 13.3.

Ejemplo 13.2 Utilizando la expresión 13.3

a). Calcular el coeficiente de correlación r de Pearson para el caso de los cigarrillos.

b). Calcular el coeficiente de correlación r de Pearson para el caso de los tiempos de las ganadoras del Maratón de Nueva York .

Solución

a). Para un empleo eficiente de la expresión 13.3, conviene organizar los cálculos en una tabla como se muestra enseguida. Tomando como base la tabla 13.2 se tiene:

	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
	1	1.5	1	2.25	1.5
	4.1	5.4	16.81	29.16	22.14
	4.5	4.9	20.25	24.01	22.05
	\vdots	\vdots	\vdots	\vdots	\vdots
	16.6	16.3	275.56	265.69	270.58
	17	18.5	289	342.25	314.5
	29.8	23.5	888.04	552.25	700.3
Sumatorias	305.4	313.2	4501.2	4462.92	4443.15

Al sustituir las sumatorias obtenidas y el valor de n en la ecuación 13.3 se tiene:

$$r = \frac{25 \times 4443.15 - 305.4 \times 313.2}{\sqrt{25 \times 4501.2 - 305.4^2} \sqrt{25 \times 4462.92 - 313.2^2}} = 0.95748533$$

b). Tomando ahora como base la tabla 13.4 y realizando los cálculos correspondientes, se llega a:

$$\sum_{i=1}^n T_i = 1324; \quad \sum_{i=1}^n t_i = 3102.9; \quad \sum_{i=1}^n T_i^2 = 85396; \quad \sum_{i=1}^n t_i^2 = 458551.716; \quad \sum_{i=1}^n T_i t_i = 195859.23$$

Al sustituir estos valores y el de n en la ecuación 13.3 se tiene:

$$r = \frac{21 \times 195859.23 - 1324 \times 3102.9}{\sqrt{21 \times 85396 - 1324^2} \sqrt{21 \times 458551.716 - 3102.9^2}} = 0.59839$$

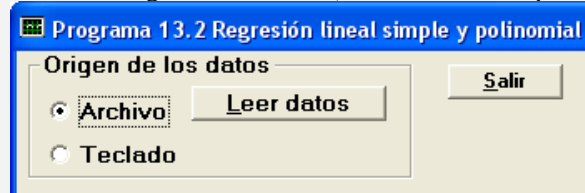
Discusión En ambos casos se tiene asociación positiva entre las variables en estudio. De acuerdo con la tabla

La realización de los cálculos para el coeficiente de correlación, tal como se vio en el ejemplo 13.2, resulta práctica ya que permite organizarlos, revisarlos y, en caso necesario, programarlos. Se recomienda, sin embargo, el empleo de programas desarrollados como los de las calculadoras científicas, los de Excel y, desde luego, los del software del libro. En cualquiera de estos casos, el cálculo del coeficiente de correlación es una primera etapa del análisis de correlación y regresión. Se muestra a continuación el uso del software del libro y más adelante el de Excel.

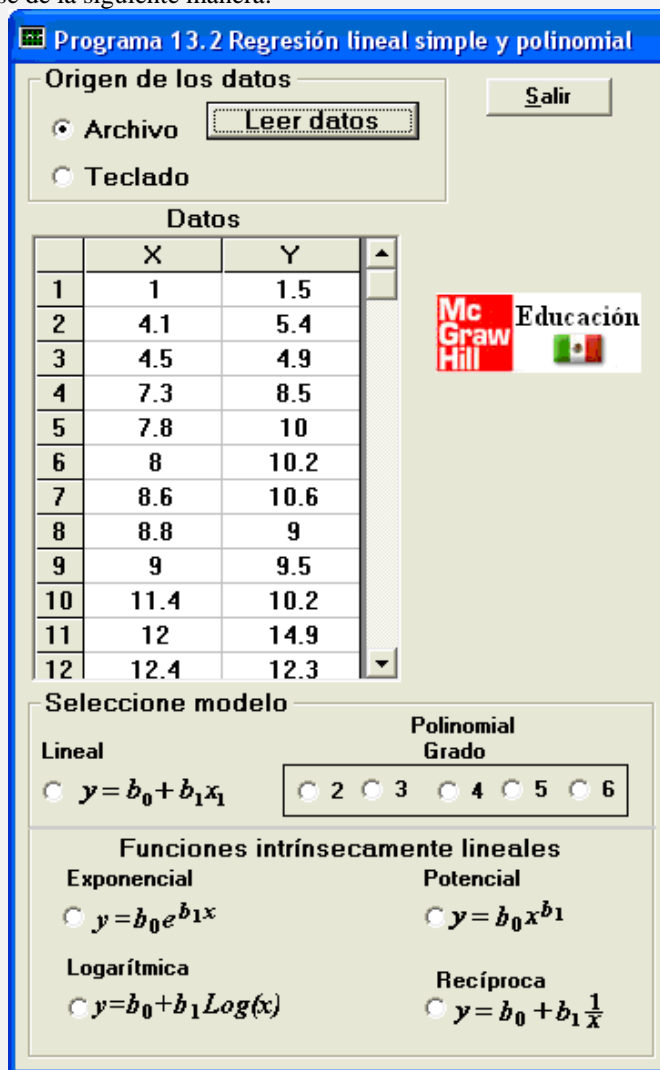
Ejemplo 13.3 Resolver el ejemplo 13.2 empleando el programa 13.2 del libro.

Solución

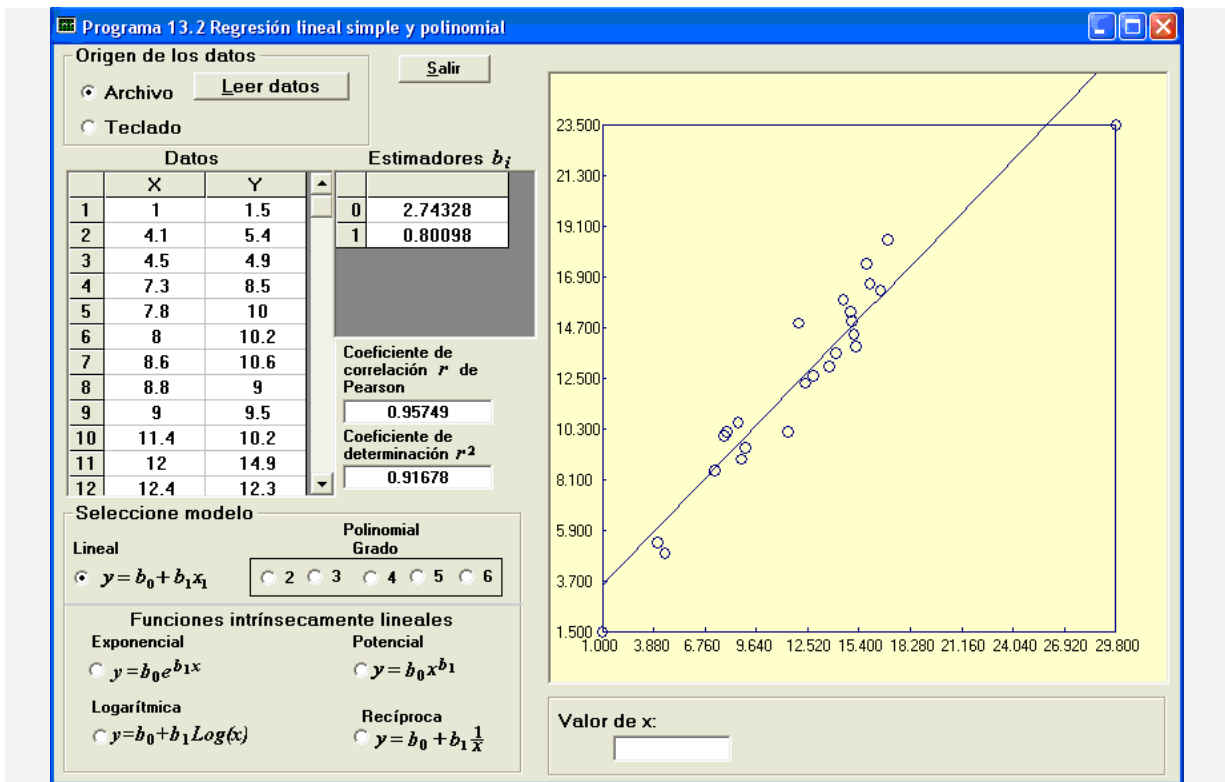
a) Al iniciar el programa 13.2 se verá la siguiente interfase (se muestra sólo la parte relevante):



Al hacer clic en el botón **Leer Datos** con la opción **Archivo** activada, se abrirá una ventana que le permitirá navegar en su computadora para seleccionar el archivo de interés. Seleccione el archivo **Cigarrillos.dat**. Si el archivo no está disponible, cree usted los datos con la opción **Teclado** y guárdelo para usos posteriores (para el lector interesado, al final del capítulo se dan indicaciones de cómo crear un archivo). Una vez que se ha leído el archivo se verá la interfase de la siguiente manera:



Al hacer clic en el botón **Lineal**, el programa elabora el diagrama de dispersión y calcula el coeficiente de correlación entre otros aspectos del análisis de correlación y regresión.



La información adicional así como las múltiples opciones restantes que el programa 13.2 muestra serán utilizadas más adelante.

Se deja al lector el inciso b.

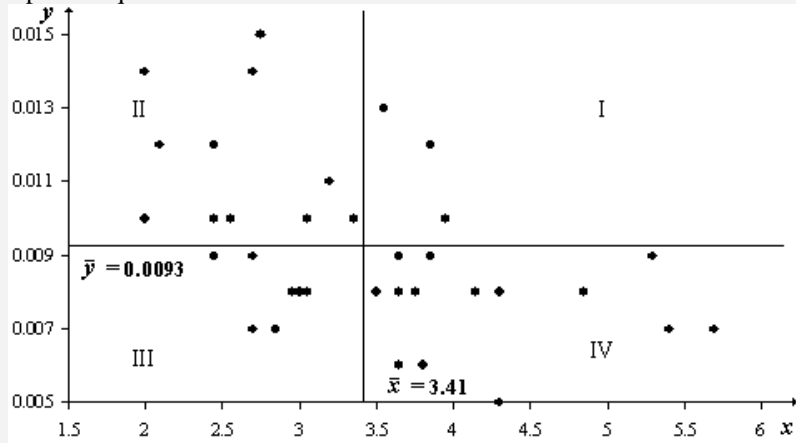
Ejemplo 13.4 La disminución de la transparencia del agua en Grand Lake, Colorado, motivó un estudio para establecer estadísticamente cuáles de varios parámetros muestran una correlación fuerte con la profundidad Secchi (ver ventana al conocimiento 1). El objetivo del estudio era establecer cuáles de los parámetros contribuían mayormente a la reducción de la transparencia. Se da a continuación los valores muestrales de parejas correspondientes a la profundidad Secchi (variable predictiva x) en metros y la cantidad de fósforo total (variable respuesta y) correspondiente (http://www.cdph.state.co.us/op/wqcc/WQClassandStandards/Regs33-37/33_37RMH2008/ProponentsPHS/33_37phsNWCCOGGrandCoEx3.pdf).

x_i	y_i	x_i	y_i	x_i	y_i
2	0.014	2.95	0.008	3.8	0.006
2	0.01	3	0.008	3.85	0.012
2.1	0.012	3.05	0.008	3.85	0.009
2.45	0.012	3.05	0.01	3.95	0.01
2.45	0.01	3.2	0.011	4.15	0.008
2.45	0.009	3.35	0.01	4.3	0.008
2.55	0.01	3.5	0.008	4.3	0.005
2.7	0.014	3.55	0.013	4.85	0.008
2.7	0.009	3.65	0.009	5.3	0.009
2.7	0.007	3.65	0.008	5.4	0.007
2.75	0.015	3.65	0.006	5.7	0.007
2.85	0.007	3.75	0.008		

- Construir un diagrama de dispersión, calcular el centroide y trazar las líneas de división de modo que quede dividido en cuatro regiones.
- Calcular el coeficiente de correlación r de Pearson y de acuerdo a la tabla 13.6, establecer el tipo y grado de correlación entre las variables

Solución

a) El diagrama de dispersión queda así:



b) Se organizan los cálculos para el cálculo de r de la siguiente manera:

	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
	2	0.014	4	0.000196	0.028
	2	0.01	4	0.0001	0.02
	2.1	0.012	4.41	0.000144	0.0252
	⋮	⋮	⋮	⋮	⋮
	5.3	0.009	28.09	0.000081	0.0477
	5.4	0.007	29.16	0.000049	0.0378
	5.7	0.007	32.49	0.000049	0.0399
Sumatorias:	119.5	0.325	438.28	0.003213	1.0726

Sustituyendo las sumatorias y el valor de n en la ecuación 13.3:

$$r = \frac{35(1.0726) - 119.5(0.325)}{\sqrt{35(438.28) - (119.5)^2} \sqrt{35(0.003213) - (0.325)^2}} = -0.4819494$$

De acuerdo con la tabla 13.6, $-0.5 \leq r < 0$. Por lo tanto se trata de una correlación negativa débil.

El lector interesado puede utilizar el programa 13.2 como se vio en el ejemplo 13.3.

Actividad 13.3

a) Calcular el coeficiente de correlación r de Pearson para los datos de la tabla 13.2 pero designando la columna del alquitrán como y y la columna del CO como x .

Sugerencia. Utilice el programa 13.2 del libro

b) A partir del resultado obtenido en el inciso anterior, discuta la conjetura de que el valor de r es el mismo, independientemente de qué variable se designe como x y qué variable se designe como y . En otros términos, que el valor de r es el mismo para las parejas (x_i, y_i) y las parejas (y_i, x_i) . Expresé verbalmente y en forma escrita los resultados de la discusión.

Sugerencia: intercambie x y y en la expresión 13.2 o en la 13.3.

Análisis inferencial para el coeficiente de correlación r de Pearson

Prueba de hipótesis. El coeficiente de correlación r puede verse como una medida numérica de qué tan bien un modelo lineal (línea recta) representa los puntos de un diagrama de dispersión. El diagrama, sin embargo, no contiene todos los puntos posibles o población de puntos. Debido a que r se calcula con base en una muestra aleatoria de puntos, se espera que

los valores de r varíen de una muestra a otra (tal como la media o la proporción varían de una muestra a otra). Esto plantea la pregunta de la *significancia* de r . Puesto de otra manera, ¿cuál es la probabilidad que la muestra aleatoria de puntos den una correlación fuerte cuando, en realidad, los puntos de la población no están correlacionados fuertemente?

Como en los casos de la media, la desviación estándar y la proporción, se empleará una letra griega ρ (se lee ro) para representar el parámetro poblacional correspondiente a r . Con esto, la significancia de r será tratada enseguida mediante una prueba de hipótesis del coeficiente de correlación ρ . Los pasos para la prueba de hipótesis son:

Paso 1. Se elige el modelo estadístico apropiado de acuerdo a la tabla 13.7.

Cola izquierda	Cola derecha	Dos colas
$H_0 : \rho \geq 0$	$H_0 : \rho \leq 0$	$H_0 : \rho = 0$
$H_1 : \rho < 0$	$H_1 : \rho > 0$	$H_1 : \rho \neq 0$

Paso 2. Elección del estadístico de prueba. En este caso lo proporciona una conversión de la distribución de los valores muestrales r a una distribución t de Student mediante su estandarización:

$$t = \frac{r}{s_r},$$

donde s_r es la desviación estándar muestral de los valores de r , calculada de la siguiente manera:

$$s_r = \sqrt{\frac{1-r^2}{n-2}}$$

Sustituyendo se tiene

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}, \quad \text{con } gl = n-2 \quad (13.4)$$

En este paso 2, en resumen, se calcula el valor de t con los datos muestrales usando la expresión 13.4, obteniéndose el valor observado t_0 .

Paso 3. Se fija un valor de α y se calculan los valores críticos de la distribución t de Student con $n-2$ grados de libertad. Con esto quedan establecidos los intervalos de rechazo y aceptación de acuerdo al modelo estadístico del paso 1.

Actividad 13.4
Algunos autores escriben la ecuación 13.4 como

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

A partir de 13.4 obtenga la ecuación equivalente anterior.

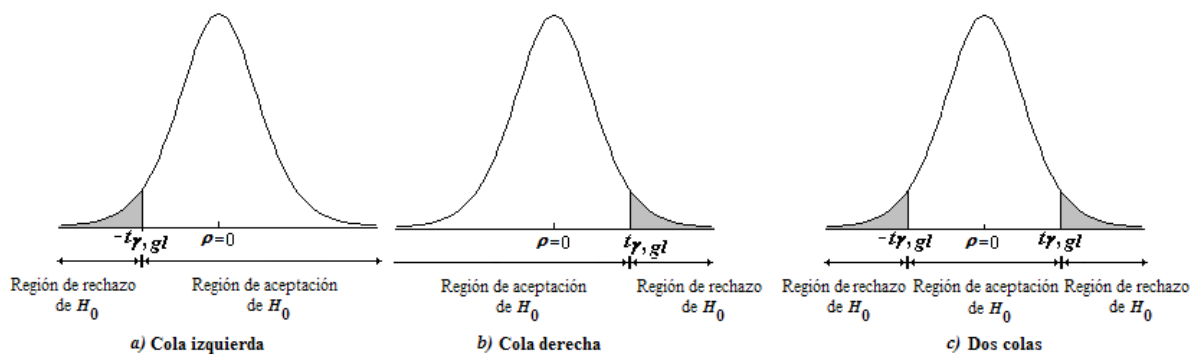


Figura 13.11 Representación gráfica de los modelos estadísticos

Paso 4.

Enfoque tradicional. Si el valor observado t_0 cae en la región de rechazo, se rechaza H_0 teniéndose una correlación lineal. Si por el contrario, el valor observado t_0 cae en la región de aceptación, se acepta H_0 y no hay una correlación lineal.

Enfoque del valor p . Al igual que en el capítulo 12 del libro, se emplea el *valor p* en forma práctica para ayudar a tomar una decisión; esto es, comparándolo con α de acuerdo a:

Si valor $p \leq \alpha$, rechazar H_0

Si valor $p > \alpha$, aceptar H_0

Se ilustra a continuación el método con el caso de los cigarrillos.

Ejemplo 13.5 Aplicar la prueba de hipótesis de dos colas al valor del coeficiente de correlación r obtenido en el ejemplo 13.2 para el caso de los cigarrillos.

Solución

Enfoque tradicional. El modelo estadístico es:

$$H_0 : \rho = 0 \text{ (no existe correlación lineal)}$$

$$H_1 : \rho \neq 0 \text{ (existe correlación lineal)}$$

Con $r = 0.95749$ y $n = 25$, el valor observado de acuerdo a la expresión 13.4 es:

$$t_0 = \frac{0.95749}{\sqrt{\frac{1 - 0.95749^2}{25 - 2}}} = 15.9185$$

Tomando el nivel de confianza $\alpha = 0.05$ y $gl = 25 - 2 = 23$, se obtienen como valores críticos a $t_{(0.025, 23)} = \pm 2.069$

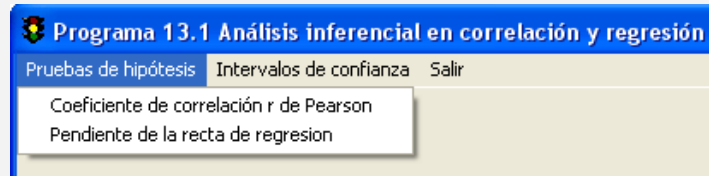
Como el valor observado queda fuera de la región de aceptación dada por $[-2.069, 2.069]$, se rechaza H_0 teniéndose una correlación lineal.

El cálculo del *valor p* mediante una tabla de distribución t de Student resulta impráctico, por lo que el enfoque del *valor p* generalmente requiere de un programa. Se muestra en el ejemplo 13.6 el uso del programa 13.1 del libro para tal fin.

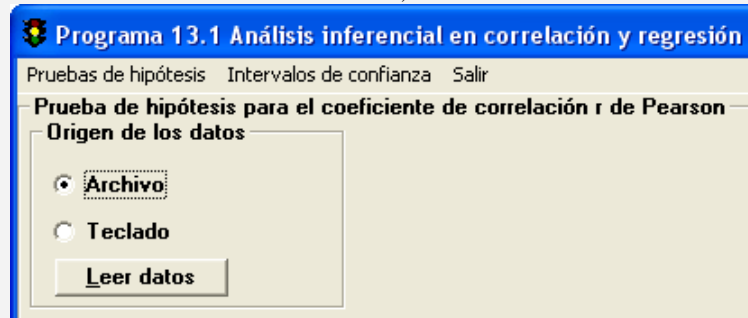
Ejemplo 13.6 Aplicar la prueba de hipótesis de dos colas al valor del coeficiente de correlación obtenido en el ejemplo 13.2 para el caso de las corredoras del maratón. Utilice el programa 13.1 del libro

Solución

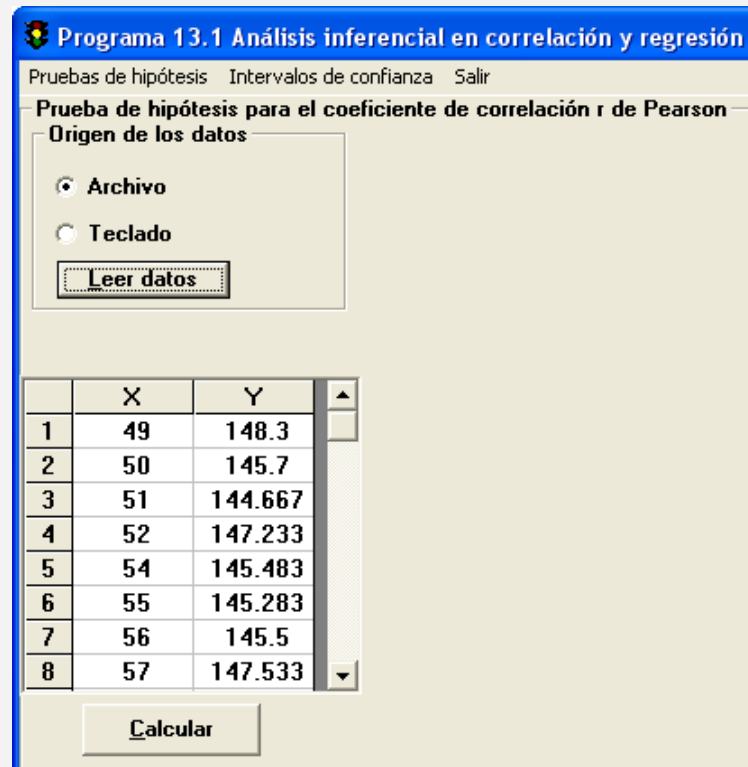
Al iniciar el programa 13.1 se verá la siguiente interfase (después de hacer clic en la opción **Pruebas de hipótesis** del menú principal):



Al hacer clic en **Coeficiente de correlación r de Pearson**, se obtiene:



Al hacer clic en el botón **Leer Datos** con la opción **Archivo** activada, se abrirá una ventana que le permitirá navegar en su computadora para seleccionar el archivo de interés. Seleccione el archivo **Maraton.dat**. Si el archivo no está disponible, cree usted los datos con la opción **Teclado** (el lector interesado podrá encontrar al final del capítulo instrucciones para crear un archivo). Una vez que se ha leído el archivo se verá la interfase de la siguiente manera:



Se hace clic en el botón **Calcular** para obtener:

Programa 13.1 Análisis inferencial en correlación y regresión

Pruebas de hipótesis Intervalos de confianza Salir

Prueba de hipótesis para el coeficiente de correlación r de Pearson

Origen de los datos

Archivo

Teclado

	X	Y
1	49	148.3
2	50	145.7
3	51	144.667
4	52	147.233
5	54	145.483
6	55	145.283
7	56	145.5
8	57	147.533

Número de parejas n

Valor de r

Hipótesis nula

$\rho =$ }

$\rho \geq$

$\rho \leq$

Hipótesis alterna

$\rho \neq 0$

Nivel de significancia

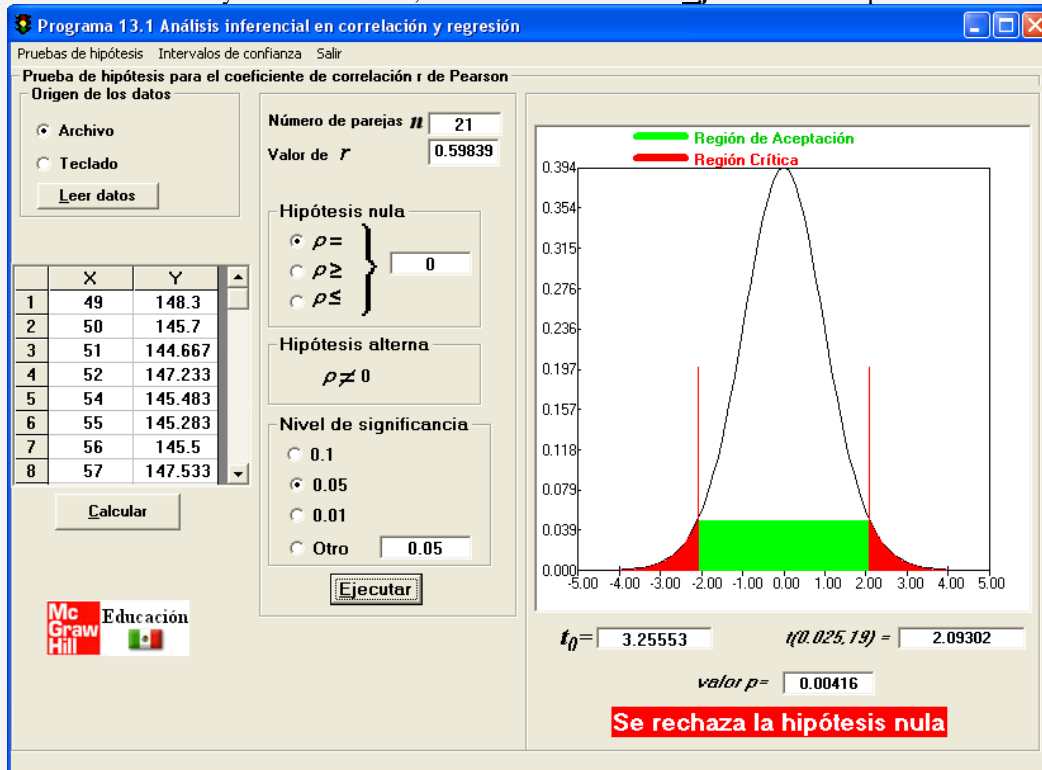
0.1

0.05

0.01

Otro

Como se ve en esta última interfase, se obtiene el valor del coeficiente de correlación y, de manera predeterminada la **Hipótesis nula** correspondiente a una prueba de dos colas y **Nivel de significancia** en **0.05**. Dado que este es el modelo y el nivel deseado, se hace clic en el botón **Ejecutar** con lo que se obtiene:



Enfoque clásico. El valor $t_o = 3.25553$ cae fuera de la región de aceptación (franja verde acotada por ± 2.09302) por lo que se rechaza la hipótesis nula favoreciéndose la hipótesis alterna: existe correlación entre las variables.

Enfoque del valor p . Se compara el *valor p* con α . Como $0.00416 < 0.05$ se rechaza la hipótesis nula. El *valor p* , sin embargo, permite continuar el análisis. Por ejemplo, de acuerdo a la definición de *valor p* como *el nivel de significancia menor que llevaría al rechazo de la hipótesis nula* (ver capítulo 12 del libro), ninguno de los valores preestablecidos de α llevaría a aceptar la hipótesis nula. Compruébelo haciendo clic en 0.01 para nivel de significancia y luego en **Ejecutar**. Puede también corroborar la definición de *valor p* activando la opción **Otro de Nivel de significancia** y escribiendo un valor menor que 0.00416. La ejecución dará como resultado la aceptación de la hipótesis nula. Esto significa que para aceptar la hipótesis nula se requeriría un valor de α extremadamente pequeño (en comparación con el menor valor de los recomendados: 0.01), dando mayor confianza (no certidumbre) en rechazar la hipótesis nula.

En conclusión, el programa 13.1 permite abordar ambos enfoques pero no sólo como un instrumento de cálculo y de visualización, sino también de exploración y análisis.

Comentario: En algunos textos modernos, en las revistas científicas y en artículos de investigación se reporta simplemente el *valor p* para que el analista o lector pueda concluir con *cualquier* nivel de significancia especificado.

Actividad 13.5 Aplique una prueba de hipótesis al coeficiente de correlación r obtenido en el caso de los cigarrillos. Use α igual a 0.05 y 0.01. Sugerencia. Utilice el programa 13.1 del libro.

Pruebas de una cola

En los ejemplos anteriores se aplicó una prueba de dos colas. En general, los ejemplos y problemas de este capítulo implicarán únicamente pruebas de dos colas, pero puede presentarse una prueba de una cola para una declaración de correlación lineal positiva o una declaración de correlación lineal negativa. El programa 13.1 permite cualquiera de estas pruebas.

Intervalos de confianza para ρ

Una vez que se encuentra el valor de r para una muestra de n pares de puntos, por ejemplo 0.60, y la prueba de hipótesis resulta significativa, se puede pensar en estimar ρ . La opción más recomendable es mediante un intervalo de confianza. El procedimiento es similar a los vistos en el capítulo 11 para la media y la proporción. Por ello se plantea como el proyecto 13.1 al final del capítulo.

Precauciones sobre correlación

El coeficiente de correlación r es una herramienta matemática para medir la fuerza de una relación lineal entre dos variables. Como tal, no tiene implicaciones de causa o efecto. El hecho de que dos variables tiendan a aumentar o disminuir juntas **no significa que el cambio**

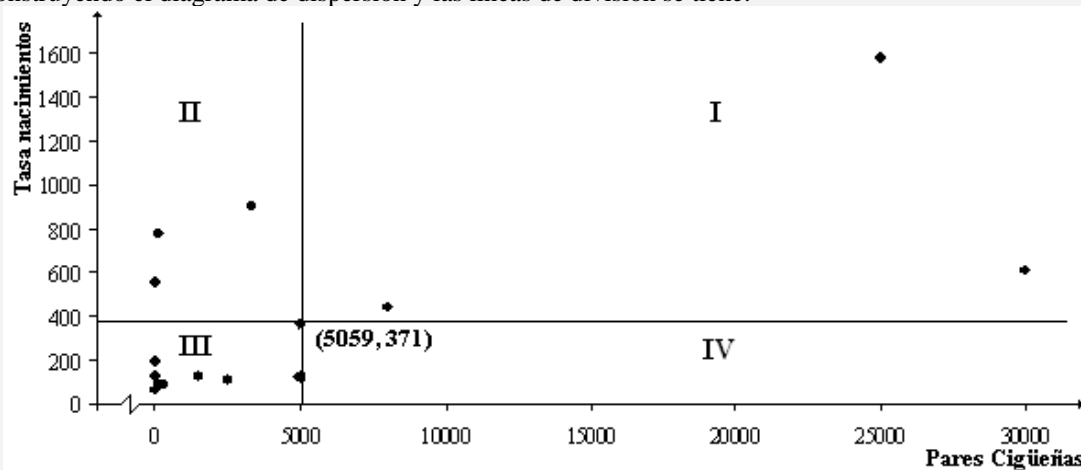
en una *cause* un cambio en la otra. En estadística, cuando r indica una correlación lineal significativa entre x y y , se considera que cambios en los valores de y tienden a responder a cambios en los valores de x de acuerdo a un modelo lineal.

Una correlación significativa entre x y y se debe algunas veces a otras variables, llamadas *variables o factores de confusión*. Una variable o factor de confusión es una variable que no es predictiva ni variable respuesta; no obstante, puede ser responsable de cambios en x y y . El siguiente ejemplo ilustra de manera amena esto.

Ejemplo 13.7 La cigüeña blanca es un pájaro sorprendentemente común en muchas partes de Europa. La tabla siguiente muestra datos geográficos y demográficos de 17 países Europeos.

País	Área (km ²)	Cigüeñas (parejas con crías)	Población (10 ⁶)	Tasa nacimientos (10 ³ /año)
Albania	28750	100	3.2	83
Austria	83860	300	7.6	87
Bélgica	30520	1	9.9	118
Bulgaria	111000	5000	9.0	117
Dinamarca	43100	9	5.1	59
Francia	544000	140	56	774
Alemania	357000	3300	78	901
Grecia	132000	2500	10	106
Holanda	41900	4	15	188
Hungría	93000	5000	11	124
Italia	301280	5	57	551
Polonia	312680	30000	38	610
Portugal	92390	1500	10	120
Rumania	237500	5000	23	367
España	504750	8000	39	439
Suiza	41290	150	6.7	82
Turquía	779450	25000	56	1576

Construyendo el diagrama de dispersión y las líneas de división se tiene:



La gráfica del número de parejas de cigüeñas con crías versus el número de nacimientos en cada país sugiere una posible correlación! entre ambas variables.

Calculando el coeficiente de correlación de Pearson: $r = 0.62$, cuya significancia estadística puede medirse con una prueba de hipótesis de dos colas:

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

El valor observado es: $t_0 = 3.062$.

Tomando el nivel de significancia $\alpha = 0.05$ y $gl = 17 - 2 = 15$, se obtiene como valores críticos a $t_{(0.025,15)} = \pm 2.131$

El valor observado queda fuera de la región de aceptación y se rechaza H_0 , teniéndose una correlación lineal.

El lector desprevenido podría pensar que se trata de una demostración estadística de que las cigüeñas traen a los bebés. Una explicación plausible a esta “correlación” es la existencia de un factor común a ambas variables que, no obstante no tener nada en común entre ellas, produce una aparente correlación. El factor podría ser el área. A mayor área, mayor tasa de nacimientos y mayor número de cigüeñas.

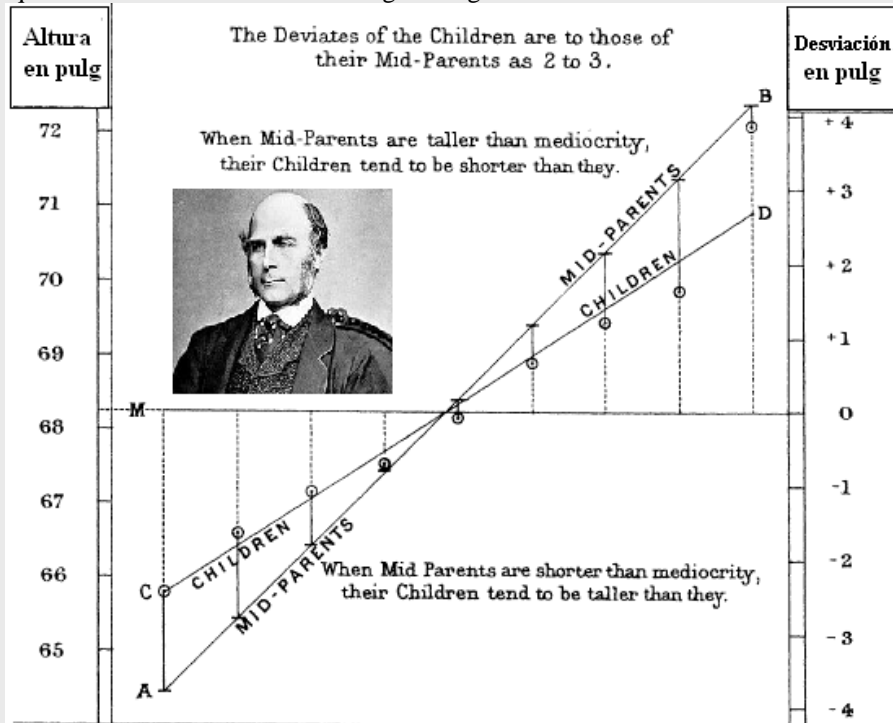
<http://score.kings.k12.ca.us/lessons/wwwstats/lurking.variables.html>

Ventana al conocimiento. 2

En 1880 el científico Inglés Sir Francis Galton introdujo el concepto de correlación, así como el uso de la línea de regresión en sus estudios de investigaciones genéticas. He aquí algunas de sus conclusiones en el memorable artículo *Regression towards Mediocrity in Hereditary Stature*. By Francis Galton, FRS, &c. *Journal of the Anthropological Institute of Great Britain* 1886, 246 *Anthropological Miscellanea*.

“Han pasado ya algunos años desde que realicé una serie extensa de experimentos sobre el producto de semillas de diferentes tamaños pero de las mismas especies... El resultado de estos experimentos parece indicar que las semillas producidas no tendían a parecerse en tamaño a las semillas de origen, sino a ser más mediocres que ellas –a ser más pequeñas que los padres, si los padres eran muy grandes; a ser más grandes que los padres, si los padres eran pequeños.

Más tarde busqué evidencia antropológica, considerando el caso de las semillas como un medio que arrojará luz sobre la herencia en el hombre... Un análisis de los datos confirmó completamente y, fue más allá de las conclusiones que obtuve con las semillas.” La siguiente gráfica ilustra lo anterior.



El método de regresión lineal por mínimos cuadrados también puede emplearse con una variable aleatoria y la otra determinista o ambas deterministas (caso común en ingeniería y ciencias).

13.2 Regresión lineal

El paso siguiente en el análisis de dos variables aleatorias x y y consiste en encontrar la función lineal $y = b_0 + b_1x$ que sirva para modelar la relación entre ellas. Este proceso es llamado *regresión lineal* y a la línea resultante *recta de regresión*. Para ilustrarlo se considera de nuevo el caso de los cigarrillos.

Tomando el diagrama de dispersión de la figura 13.1, pero con una línea recta trazada “arbitrariamente” por *entre los puntos* se obtiene la figura 13.12.

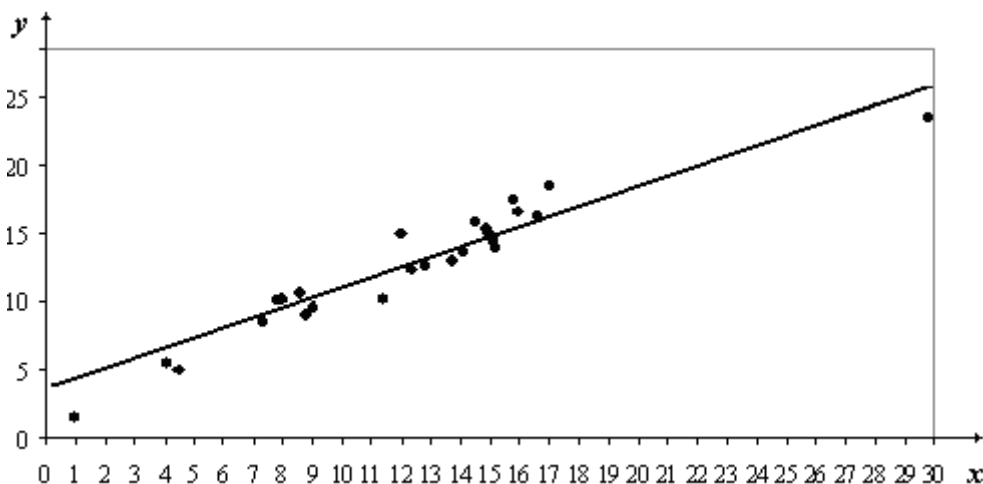


Figura 13.12 Trazo de una línea recta cualquiera por entre los puntos del diagrama de dispersión.

La finalidad de la recta es representar algebraicamente a los datos, es decir, con una ecuación del tipo $y = b_0 + b_1x$. Se desearía entonces que la recta trazada representara los datos muestrales de la “mejor manera posible”. La “mejor” representación puede interpretarse de diferentes formas: la recta que toque más puntos; aquella recta que permita tener igual número de puntos arriba y debajo de ella; la recta que pase por el punto central (\bar{x}, \bar{y}) ; etc. Tales criterios, sin embargo, son subjetivos y generalmente no conducen a una recta única. Uno de los criterios formales más ampliamente usado es el del ajuste por *mínimos cuadrados* (en el capítulo 3 se estudió el criterio que da lugar a la *recta de ajuste mediana*). Se presenta a continuación el criterio de ajuste por mínimos cuadrados, mediante una serie de pasos gráficos, de modo que se capte intuitivamente la idea que lo sustenta.

Considere para el primer paso el diagrama de dispersión de la figura 13.12. Enseguida se trazan líneas verticales desde cada uno de los puntos a la recta trazada arbitrariamente (ver figura 13.13). Se dan las distancias verticales de algunos de estos puntos a la recta (el cálculo de tales distancias en este momento es irrelevante ya que sólo tienen fines ilustrativos del método).

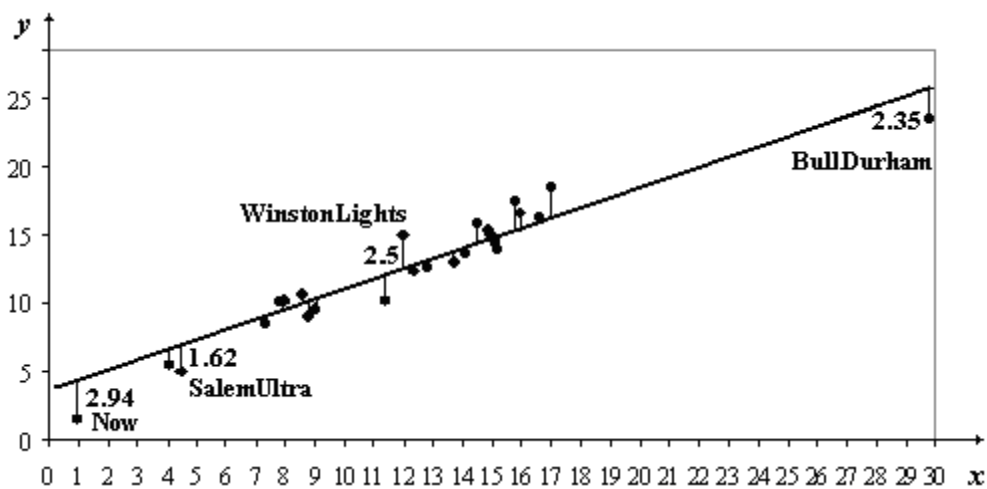


Figura 13.13 Trazo de líneas verticales de los puntos a la recta

Luego se toma cada una de las verticales trazadas como el *lado* de un cuadrado. A cada cuadrado le corresponde un área igual a $lado \times lado$; por ejemplo, la distancia del punto correspondiente a Now (1, 1.5) a la recta es 2.94 y su área es 8.64, mientras que la distancia de Salem Ultra (4.5, 4.9) es 1.62 y su área es 2.62. En el caso de la marca Bull Durham el área es 5.52 como se ve en la figura 13.14.

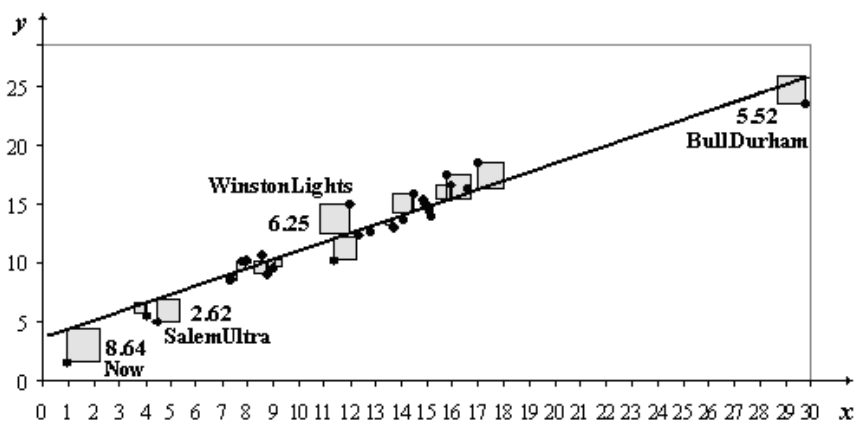


Figura 13.14 Construcción de los cuadrados y cálculo de sus áreas.

El siguiente paso consiste en *sumar las áreas de los cuadrados generados por cada uno de los 25 puntos*.

Si se traza arbitrariamente otra recta por entre los puntos se generaría otro juego de 25 cuadrados, cuya suma daría un área total seguramente distinta a la del caso inicial. Con estas consideraciones el escenario queda listo para enunciar el criterio para seleccionar una recta de ajuste:

La recta de ajuste por mínimos cuadrados es aquella que pasa por entre los puntos de la muestra, de tal modo que produce el área total mínima.

El criterio así establecido da lugar a una recta única. Su deducción o, dicho de otra forma, la deducción del cálculo de la ordenada al origen b_0 y la pendiente b_1 es un proceso técnico que se puede consultar en el Apéndice G del libro. Las expresiones resultantes para dichos parámetros son:

$$b_0 = \frac{\left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n (x_i)^2\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n x_i y_i\right)}{n \sum_{i=1}^n (x_i)^2 - \left(\sum_{i=1}^n x_i\right)^2} \quad (13.5)$$

$$b_1 = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n x_i\right)}{n \sum_{i=1}^n (x_i)^2 - \left(\sum_{i=1}^n x_i\right)^2} \quad (13.6)$$

La recta de ajuste por mínimos cuadrados o recta de regresión queda entonces

$$\hat{y} = \frac{\left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n (x_i)^2\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i x_i\right)}{n \sum_{i=1}^n (x_i)^2 - \left(\sum_{i=1}^n x_i\right)^2} + \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n x_i\right)}{n \sum_{i=1}^n (x_i)^2 - \left(\sum_{i=1}^n x_i\right)^2} x \quad (13.7)$$

Los valores que da la recta de regresión (13.7) correspondientes a x_1, x_2, \dots, x_n son, en general, diferentes a los valores observados y_1, y_2, \dots, y_n , por lo que suelen denotarse como $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$ y llamarse valores ajustados o estimados. De la misma manera, cualquier otro valor calculado con la recta para un valor arbitrario de la variable x se denota como \hat{y} .

Por último, dado que b_0 y b_1 se obtuvieron a partir de una muestra de n puntos y no de la población, son estimaciones de los parámetros poblacionales correspondientes β_0 y β_1 . La letra griega empleada se pronuncia beta.

Actividad 13.6 Examina en equipo las expresiones 13.3, 13.5 y 13.6. ¿Encuentra algunos elementos comunes? Descríbalos verbalmente y en forma escrita.

Cálculos para encontrar la recta de regresión

Directos. En los cálculos para los parámetros b_0 y b_1 de la recta de regresión se emplean los valores de las sumatorias, a excepción de $\sum_{i=1}^n y_i^2$, que se emplean para el cálculo del coeficiente de correlación r de Pearson.

Así, en el caso de los cigarrillos (ver ejemplo 13.2) se tiene:

$$\sum_{i=1}^n x_i = 305.4; \sum_{i=1}^n y_i = 313.2; \sum_{i=1}^n x_i^2 = 4501.2; \sum_{i=1}^n x_i y_i = 4443.15$$

Al sustituir los valores de las sumatorias y de n (25) en las ecuaciones 13.5 y 13.6:

$$b_0 = \frac{(313.2)(4501.2) - (305.4)(4443.15)}{25(4501.2) - (305.4)^2} = 2.74327755$$

$$b_1 = \frac{25(4443.15) - (313.2)(305.4)}{25(4501.2) - (305.4)^2} = 0.80097597$$

La ecuación de la recta de ajuste por mínimos cuadrados para la muestra de marcas de cigarrillos queda entonces:

$$\hat{y} = 2.74327755 + 0.800975997x$$

Al graficar la recta de ajuste en el diagrama de dispersión de la figura 13.12 pero conservando la recta trazada arbitrariamente (en gris) para comparación, se tiene:

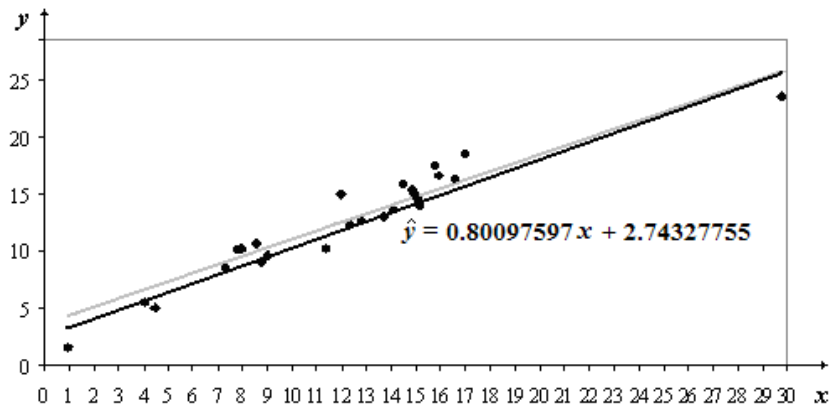


Figura 13.15 Recta de regresión y recta arbitraria.

Programa 13.2. El cálculo de los parámetros puede hacerse mediante el programa 13.2. En la última interfase mostrada en el ejemplo 13.3, se tienen los valores de los parámetros como:

Estimadores b_i	
0	2.74328
1	0.80098

Se dispone también del diagrama de dispersión y de la representación gráfica de la recta de regresión.

Ejemplo 13.8 Empleando la ecuación de la recta de regresión para los cigarrillos, calcular los valores \hat{y}_i correspondientes a las x_i y comparar con los valores observados y_i .

Solución

Valores de CO			Valores de CO		
Alquitrán	Observados	Ajustados	Alquitrán	Observados	Ajustados
x	y	\hat{y}	x	y	\hat{y}
1	1.5	3.54	13.7	13	13.72
4.1	5.4	6.03	14.1	13.6	14.04
4.5	4.9	6.35	14.5	15.9	14.36
7.3	8.5	8.59	14.9	15.4	14.68
7.8	10	8.99	15	15	14.76
8	10.2	9.15	15.1	14.4	14.84
8.6	10.6	9.63	15.2	13.9	14.92
8.8	9	9.79	15.8	17.5	15.40
9	9.5	9.95	16	16.6	15.56
11.4	10.2	11.87	16.6	16.3	16.04
12	14.9	12.35	17	18.5	16.36
12.4	12.3	12.68	29.8	23.5	26.61
12.8	12.6	13.00			

Comentarios. Los valores \hat{y}_i rara vez coinciden con sus correspondientes valores observados y_i . Esto es el resultado de tener una línea que pasa “por entre” los puntos no “por los puntos”. La técnica garantiza, sin embargo, que la suma $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ es mínima respecto a cualquier otra línea recta que pase por entre los puntos. La diferencia $(y_i - \hat{y}_i)$ es conocida como error, desviación o residual. La suma de estos valores y de sus cuadrados es utilizada ampliamente en análisis subsecuentes.

**Material
opcional**

Exploración del método de ajuste por mínimos cuadrados con el programa 4.1

El programa 4.1 permite, una vez que se han leído los datos (puntos), tener el diagrama de dispersión y una recta que une los puntos extremos. Puede luego manipularse la recta (en negro) y observar el área total (ver ventana *suma de cuadrados*) de cada recta que se vaya formando. Se sugiere explorar visualmente con varias rectas hasta encontrar aquella que minimice el área y luego comparar con la recta de ajuste por mínimos cuadrados o de regresión que da también el programa. Se muestra a continuación una etapa en la exploración para encontrar la recta en el caso de los cigarrillos.

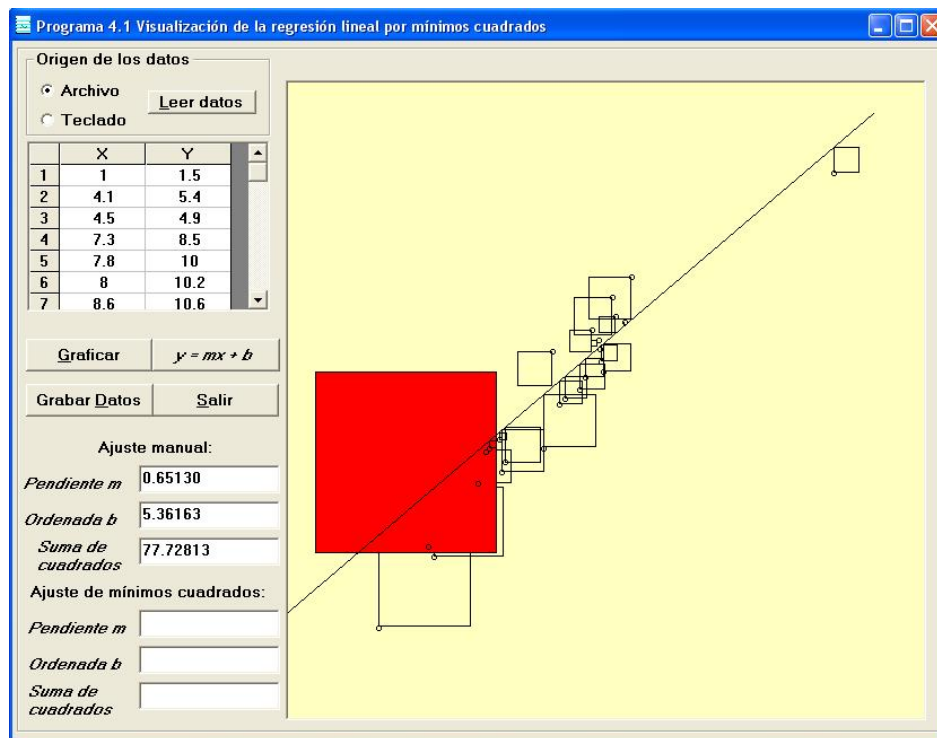


Figura 13.16. Interfase que muestra una etapa en la exploración de la mejor recta


**Material
opcional**

Obtención de la recta de regresión y del coeficiente de correlación con Excel

Para ilustrar el procedimiento se usará la tabla 13.2. Primero se capturan los datos en una hoja de cálculo de Excel. Por ejemplo en las celdas **A1:C27**, como se ve en la figura siguiente.

	A	B	C
1	Marca	Alquitran: x	CO: y
2		(mg)	(mg)
3	Now	1	1.5
4	Carlton	4.1	5.4
5	SalemUltra	4.5	4.9
6	True	7.3	8.5
7	Ment	7.8	10
8	CamelLights	8	10.2
9	ViceroyRichLight	8.6	10.6
10	GoldenLights	8.8	9
11	NewportLights	9	9.5
12	MultiFilter	11.4	10.2
13	WinstonLights	12	14.9
14	Kent	12.4	12.3
15	PallMallLight	12.8	12.6
16	LarkLights	13.7	13
17	Alpine	14.1	13.6
18	Tareyton	14.5	15.9
19	L&M	14.9	15.4
20	Chesterfield	15	15
21	Marlboro	15.1	14.4
22	VirginiaSlims	15.2	13.9
23	Raleigh	15.8	17.5
24	Benson&Hedges	16	16.6
25	Kool	16.6	16.3
26	OldGold	17	18.5
27	BullDurham	29.8	23.5

Figura 13.17 Interfase con la tabla de datos en Excel

Después se señalan las celdas **B3:C27** y se invoca al asistente para gráficos que se encuentra en la barra de herramientas **Estándar** . Seleccionar **XY (Dispersión)** en **Tipo de gráfico** y **Dispersión** en **Subtipo de gráfico** y hacer clic en el botón **Finalizar** (ver figura 13.13)

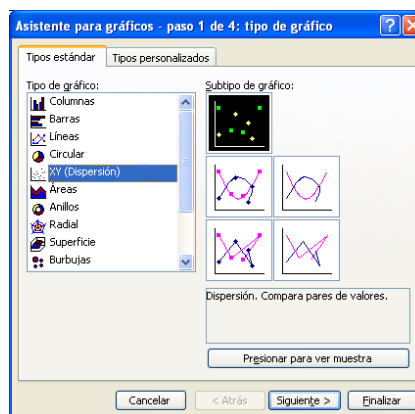


Figura 13.18 Interfase del asistente para gráficos de Excel

El resultado después de estas operaciones es:

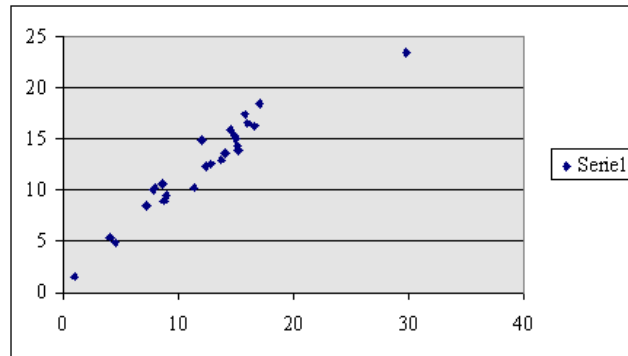


Figura 13.19 Interfase con la gráfica resultante.

Hay diversas opciones para presentación de la gráfica: Se puede señalar la Leyenda que dice **Serie 1** y usar la tecla **Supr** para eliminarla. Hacer clic en alguna de las líneas de división, con lo que se señalarán todas, y usar nuevamente la tecla **Supr** para eliminarlas. Hacer doble clic en el sombreado del área de graficación, con lo que aparecerá el cuadro de diálogo **Formato del área de trazado**. Del lado derecho (Área) hacer clic en **Ninguna** y luego en **Aceptar**. Cambiar la escala en ambos ejes, haciendo doble clic sobre cada uno de ellos. Por ejemplo, si se hace doble clic sobre el eje *x* aparecerá el cuadro de diálogo **Formato de ejes**. Después de hacer clic en la ficha **Escala**, se verá así:

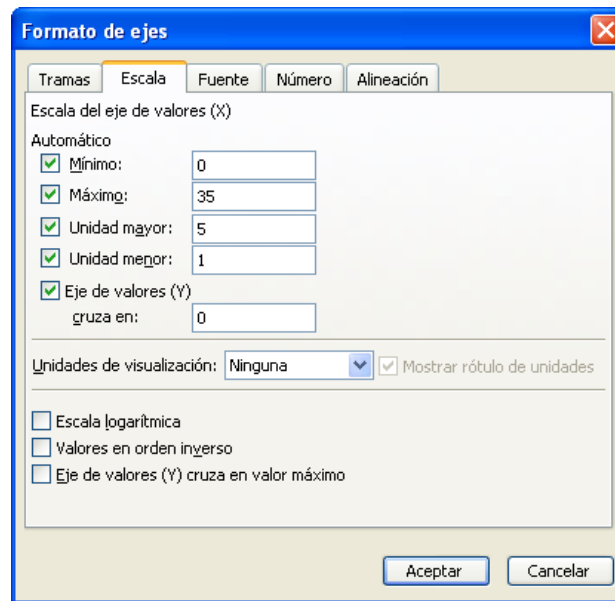


Figura 13.20 Interfase del formato de ejes.

En el **eje x** dejar **Mínimo 0** y **Máximo 30**, como se muestra en la figura anterior. Repetir lo mismo para el **eje y** dejando **Mínimo 0** y **Máximo 25**. El resultado debe ser:

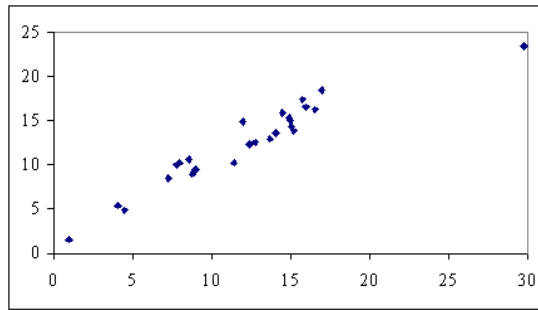


Figura 13.21 Diagrama de dispersión elaborado por Excel.

Ahora proceder a solicitar la línea de regresión. Después dar clic en el gráfico, hacer clic en el menú **Gráfico** y luego en **Agregar línea de tendencia**.

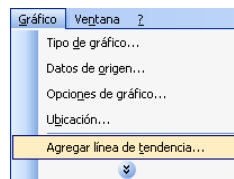


Figura 13.22 Interfase del menú Gráfico de Excel

Aparecerá el cuadro de diálogo de esta opción. Seleccionar **Lineal** en **Tipo de tendencia o regresión**. Cambiar a la ficha **Opciones** y seleccionar el cuadro **Presentar ecuación en el gráfico** y el cuadro **Presentar el valor de R cuadrado en el gráfico**. Por último, hacer clic en **Aceptar**. El resultado es:

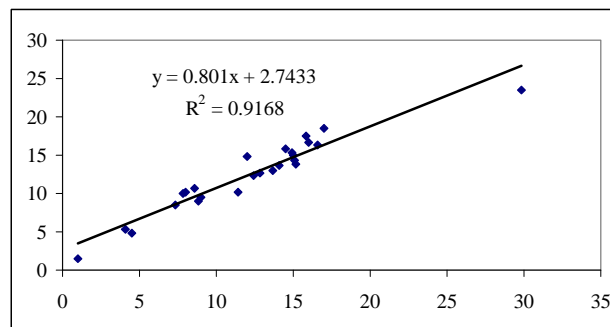


Figura 13.23 Diagrama de dispersión y recta de regresión con el coeficiente de correlación.

Excel emplea R para el coeficiente de correlación y lo da al cuadrado. Más adelante se explica el nombre, sentido y uso del coeficiente al cuadrado.

Se recomienda utilizar la tecnología para obtener la recta de regresión (programa 4.1, programa 13.2 (ver ejemplo 13.9), Excel o una calculadora que disponga del programa de regresión lineal). En caso que se quieran realizar paso a paso, se sugiere organizarlos como se mostró anteriormente en el ejemplo 13.2.

Predicción de valores utilizando la recta de regresión

Una vez que se tiene la representación analítica de los puntos, se pueden llevar a cabo distintas actividades, siendo una de las más importantes la *predicción*. Para ver como se realiza y se emplea, se recurre a un ejemplo.

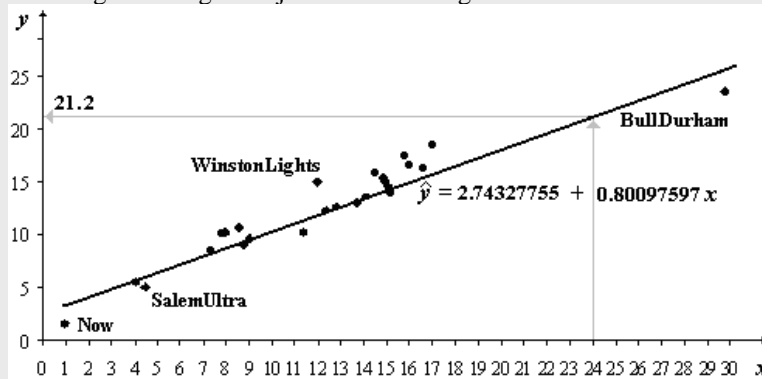
Ejemplo 13.9 Considere que se elabora un cigarrillo con un contenido de alquitrán igual a 24 mg y se desea estimar (sin realizar la medición química) la cantidad de CO que se desprende en su consumo.

Solución

La estimación de la cantidad de CO se obtiene sustituyendo 24 en la ecuación de la recta de regresión encontrada y calculando el valor de y:

$$y = 0.80097597 \times 24 + 2.74327755 = 21.9667008$$

La cantidad de CO es 22.0 mg. En la figura adjunta se muestra gráficamente la estimación.



La estimación es un valor que naturalmente indica un referente de lo que puede esperarse e incluso para calcular un intervalo de confianza.

La predicción puede verificarse empleando el programa 13.2 del libro. En el ejemplo 13.3 se dieron los primeros pasos. Lo siguiente es escribir 24 en la ventana correspondiente a **Valor de x** y hacer clic en **Calcular**. El resultado es:

Programa 13.2 Regresión lineal simple y polinomial

Origen de los datos: Archivo Teclado

Datos	X	Y	Estimadores b_i
1	1	1.5	0 2.74328
2	4.1	5.4	1 0.80098
3	4.5	4.9	
4	7.3	8.5	
5	7.8	10	
6	8	10.2	
7	8.6	10.6	
8	8.8	9	
9	9	9.5	
10	11.4	10.2	
11	12	14.9	
12	12.4	12.3	

Coefficiente de correlación r de Pearson: 0.95749
 Coeficiente de determinación r^2 : 0.91678

Selección de modelo: Lineal $y = b_0 + b_1x_1$

Funciones intrínsecamente lineales:
 Exponencial: $y = b_0e^{b_1x}$
 Logarítmica: $y = b_0 + b_1 \log(x)$
 Potencial: $y = b_0x^{b_1}$
 Recíproca: $y = b_0 + b_1 \frac{1}{x}$

Valor de x: 24 Calcular Estimación: \hat{y}
 $y(24.00) = 21.97$

Efecto de los valores extremos o atípicos sobre la recta de regresión

En algunos casos, se tienen valores que se desvían considerablemente del patrón que siguen los demás puntos. Puede tratarse de valores extremos, influyentes y/o atípicos (ver capítulo 3

del libro). Este tipo de puntos influyen los resultados de la recta de regresión, por lo que se recomienda realizar los cálculos con y sin dichos puntos. Así, si en el caso de los cigarrillos se elimina el punto correspondiente a BullDurham y se encuentra la ecuación de la nueva recta, se tiene $\hat{y} = 1.4129 + 0.9281x$. Las gráficas de las rectas de regresión con todos los puntos y sin BullDurham se muestra en la figura 13.24

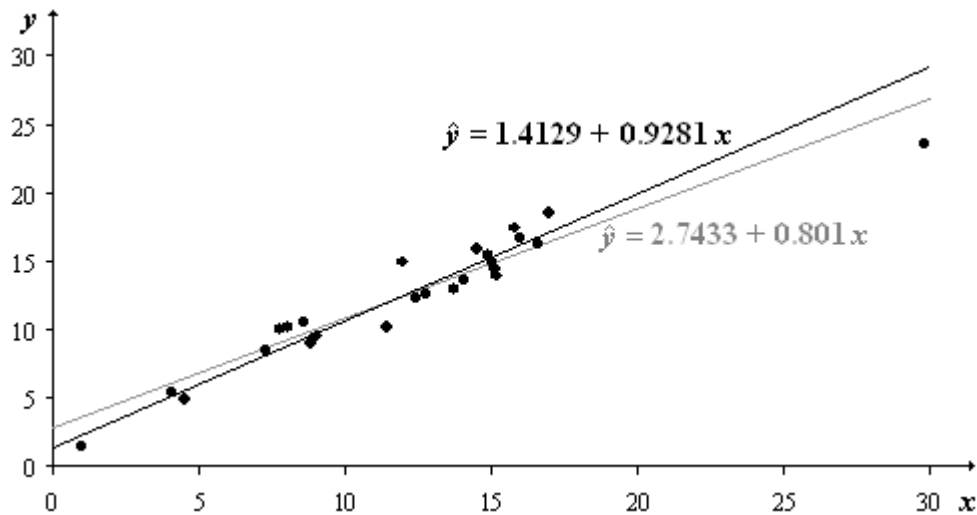


Figura 13.24 Efecto de BullDurham sobre la recta de regresión.

Resulta interesante observar cómo un solo punto modifica la ordenada al origen y la pendiente de la recta de regresión. Por ello es conveniente analizar este tipo de puntos, primero para ver si no se trata de algún error de medición y segundo para establecer si se tiene un valor atípico (ver problemas 13.26 y 13.27).

Coefficiente de determinación y error estándar de estimación

Como se vio en el ejemplo 13.8, los valores estimados \hat{y}_i no coinciden con los valores observados y_i correspondientes. Con el fin de analizar estas desviaciones, considérese un diagrama de dispersión y la correspondiente línea de regresión (ver figura 13.25). Se ha adicionado al diagrama una línea horizontal $y = \bar{y}$, a la que se llamará *línea base*; su finalidad es servir de referente para el análisis de las desviaciones. A fin de cuentas \bar{y} es el valor representativo de los valores de esa variable.

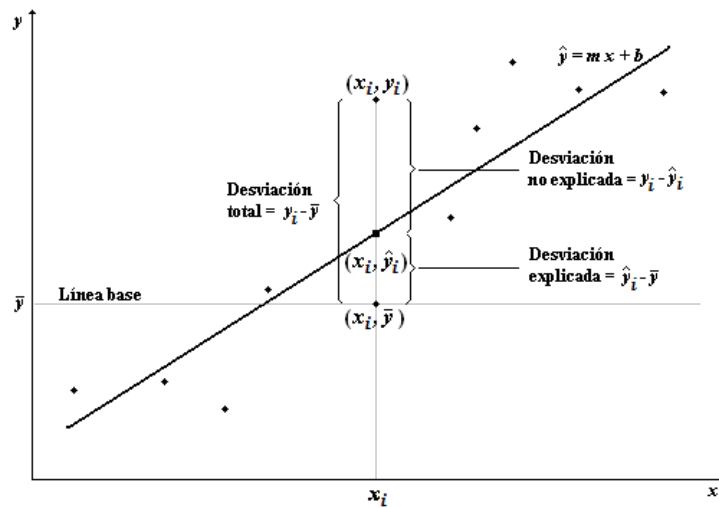


Figura 13.25 Análisis de la desviación total.

Considere un punto cualquiera (x_i, y_i) . La desviación (o diferencia) del valor y_i respecto a la línea base se representa por $y_i - \bar{y}$ y se conoce como *desviación total*.

La desviación total puede dividirse en dos partes:

1. La *desviación explicada* $\hat{y}_i - \bar{y}$ que expresa la desviación del valor \hat{y}_i a la línea base.

Podría decirse que la línea de regresión “explica” esa parte de la desviación: Imagine un punto que puede desplazarse **sobre** la línea base (ver figura 13.26); al mover el punto a la derecha, la desviación, representada por las líneas en gris, aumenta (tome en cuenta que son valores negativos); llega a cero en la intersección con la recta de regresión y sigue aumentando al avanzar a la derecha.

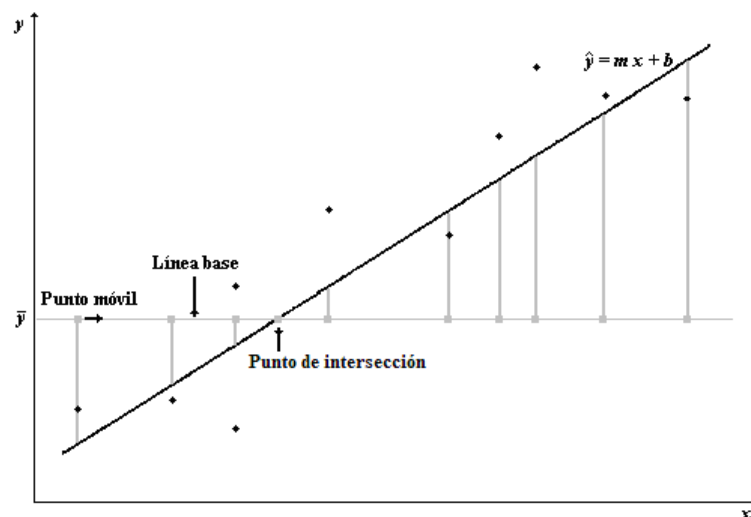


Figura 13.26 Desviación explicada $\hat{y}_i - \bar{y}$.

2. La *desviación no explicada* $y_i - \hat{y}_i$ que indica la desviación del valor y_i de la línea de regresión. Suponga ahora un punto móvil que se desplaza **sobre** la recta de regresión (ver figura 13.27). Al desplazarse sobre ésta, la desviación de los puntos de la muestra a la recta de regresión no siguen un patrón ya que su distribución es aleatoria: su posición (arriba o debajo de la recta) así como su magnitud son aleatorios. En resumen, hay factores aleatorios y de otro tipo que la recta no explica en forma alguna.

La desviación no explicada $y_i - \hat{y}_i$ recibe también los nombres de desviación aleatoria o residual.

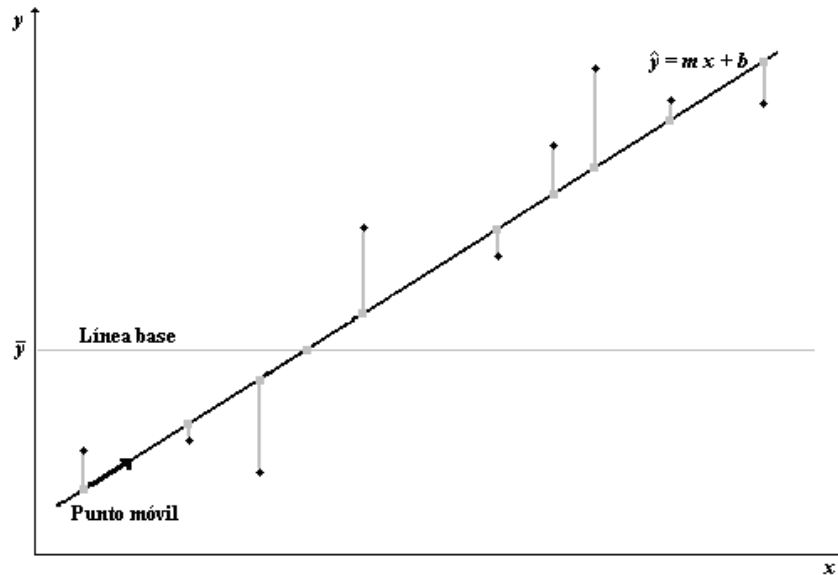


Figura 13.27 Aleatoriedad de los puntos respecto a la recta de regresión.

Para analizar algebraicamente las desviaciones considérese la siguiente relación:

$$y_i - \bar{y} = \hat{y}_i - \bar{y} + y_i - \hat{y}_i$$

(Desviación total) = (Desviación explicada) + (Desviación no explicada)

Elevando al cuadrado ambos miembros y sumando sobre todos los puntos (para ver la justificación de elevar al cuadrado se sugiere realizar la actividad 13.7):

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

Desarrollando algebraicamente el lado derecho:

Actividad 13.7
Demostrar que la suma de las desviaciones totales sobre todos los puntos muestrales, da cero:
 $\sum_{i=1}^n (y_i - \bar{y})^2 = 0$

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

La simplificación en el lado derecho se debe a que el término $2\sum_{i=1}^n (\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$ es cero (ver problema 13.22).

Como se incluye a todos los puntos de la muestra, el término variación resulta más apropiado que el de desviación.

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ \left(\begin{array}{c} \text{Variación} \\ \text{total} \end{array} \right) &= \left(\begin{array}{c} \text{Variación} \\ \text{explicada} \end{array} \right) + \left(\begin{array}{c} \text{Variación no} \\ \text{explicada} \end{array} \right) \end{aligned} \quad (13.8)$$

Dividiendo entre la variación total ambos lados de la ecuación 13.8.

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

Simplificando:

$$\begin{aligned} 1 &= \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} + \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ 1 &= \frac{\text{Variación explicada}}{\text{Variación total}} + \frac{\text{Variación no explicada}}{\text{Variación total}} \end{aligned} \quad (13.9)$$

El primer término del lado derecho es denotado como r^2 ya que la raíz cuadrada es equivalente al coeficiente de correlación de Pearson r . Se conoce como el *coeficiente de determinación* y suele manejarse así:

$$r^2 = \frac{\text{Variación explicada}}{\text{Variación total}} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (13.10)$$

La expresión 13.10 da pie a continuar el análisis iniciado con el coeficiente de correlación de Pearson pero tomando ahora en cuenta la recta de regresión. Para ello se da primero un resumen y después su aplicación:

1. El valor de r^2 es la razón de la variación explicada sobre la variación total. Es decir, r^2 es la fracción de la variación total en y que puede explicarse usando el modelo lineal $\hat{y} = b_0 + b_1x$.
2. $1 - r^2$ es la fracción de la variación total en y debida al azar o a la posibilidad de variables ocultas (desconocidas) que influyen en y .

En el caso de los cigarrillos se tiene $r = 0.96$ con lo que el coeficiente de determinación es $r^2 = 0.92$. Puede decirse entonces, de acuerdo al punto 1, que alrededor de 92% del comportamiento (variación) de la variable y , puede explicarse por medio del correspondiente comportamiento (variación) de la variable x mediante la ecuación de regresión.

Como $r^2 = 0.92$, $1 - r^2 = 0.08$. De acuerdo al punto 2, el comportamiento (variación) de alrededor de 8% de la variable y se debe al azar o a posibles variables, desconocidas para el investigador, que influyen en y .

El programa 13.2 proporciona el coeficiente de determinación, como puede verse en los ejemplos 13.3 y 13.9.

Actividad 13.8 Realizar un análisis similar al dado arriba empleando los resultados del maratón de Nueva York.

Actividad 13.9 Los datos de la tabla siguiente corresponden a las profundidades Secchi de Grand Lake, Colorado y la cantidad de fósforo total correspondiente (ver ejemplo 13.4).

x	Y	x	y	x	y
2	0.014	2.95	0.008	3.8	0.006
2	0.01	3	0.008	3.85	0.012
2.1	0.012	3.05	0.008	3.85	0.009
2.45	0.012	3.05	0.01	3.95	0.01
2.45	0.01	3.2	0.011	4.15	0.008
2.45	0.009	3.35	0.01	4.3	0.008
2.55	0.01	3.5	0.008	4.3	0.005
2.7	0.014	3.55	0.013	4.85	0.008
2.7	0.009	3.65	0.009	5.3	0.009
2.7	0.007	3.65	0.008	5.4	0.007
2.75	0.015	3.65	0.006	5.7	0.007
2.85	0.007	3.75	0.008		

- En el ejemplo 13.4 se calculó el coeficiente de correlación r con la expresión 13.3. Eleva al cuadrado el valor encontrado.
- Encontrar el coeficiente de determinación con la expresión 13.10 y comparar con el valor obtenido en el inciso anterior.

Los coeficientes de correlación permiten medir la dispersión de los puntos alrededor de la línea de regresión. Otra forma de medir la dispersión es empleando la variación no explicada en la forma siguiente:

$$\text{Error estándar de estimación} = s_e = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}} \quad (13.11)$$

Corresponde a la desviación estándar para una variable y y de ahí el nombre de *error estándar de estimación*. Se emplea para construir intervalos de confianza, por ejemplo de las estimaciones dadas por la recta de ajuste por mínimos cuadrados, como se ve a continuación.

Intervalos de confianza para las estimaciones \hat{y} correspondientes a un valor dado de x

Los parámetros b_0 y b_1 de la recta de regresión $\hat{y} = b_0 + b_1x$ se calculan con las n parejas de puntos (x, y) que constituyen una muestra de la población. Si se empleara la población de puntos posibles (x, y) , los parámetros correspondientes serían denotados como β_0 y β_1 respectivamente, y la recta de ajuste quedaría representada como $\hat{Y} = \beta_0 + \beta_1 x$. Esta ecuación, sin embargo, no daría los valores verdaderos de la variable respuesta, representados en adelante por Y , ya que faltaría considerar el error aleatorio en Y . Se denota como ε y su

presencia se debe a que siempre hay una variación en Y debida estrictamente al fenómeno aleatorio, inherente a cualquier situación. Adicionando a la recta de regresión de la población dicho componente, se obtiene la ecuación de la recta que da el *valor verdadero*.

Recta de regresión (muestra)	$\hat{y} = b_0 + b_1x$	Estimación
Recta de regresión (población)	$\hat{Y} = \beta_1 x + \beta_0$	Estimación
Recta de regresión (población con ε)	$Y = \beta_1 x + \beta_0 + \varepsilon$	Valor verdadero

Debido al término aleatorio ε , para cada valor de x hay una distribución de valores de Y . El método de regresión lineal visto se desarrolló basándose en el supuesto de que la distribución de valores de Y correspondientes a un valor dado de x está centrada en la recta de ajuste de la población. Además, que las distribuciones de Y correspondientes a cada valor de x tienen todas la misma desviación estándar σ_ε .

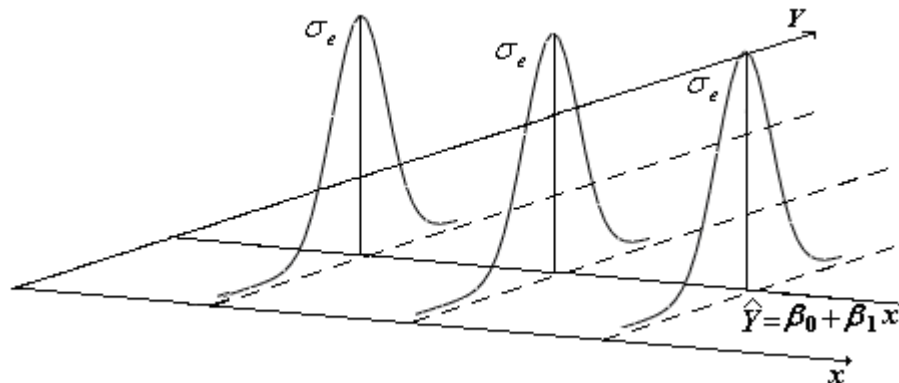


Figura 13.28 Distribuciones con centro en la recta de regresión \hat{Y} con la misma σ_ε

Usando las consideraciones teóricas anteriores, puede plantearse para un valor dado de x , un *intervalo de confianza* para el valor verdadero Y a partir de la estimación \hat{y} obtenida mediante la recta de regresión:

La expresión general para construir un intervalo de confianza de Y a partir de la estimación \hat{y} correspondiente a un valor dado de x , viene dado por:

$$\left(\hat{y} - t_{\gamma, gl} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}}, \hat{y} + t_{\gamma, gl} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}} \right) \quad (13.12)$$

Donde: \hat{y} es la estimación para un valor cualquiera de x ,

γ = nivel de significancia,

n = número de pares de datos ($n \geq 3$),

$t_{\gamma, gl}$ = valores críticos de la distribución t de Student con $gl = n - 2$,

y s_e = error estándar de estimación.

Expresando los límites del intervalo por separado:

$$LCI = \hat{y} - t_{\gamma, gl} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}}$$

$$LCS = \hat{y} + t_{\gamma, gl} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}}$$

Finalmente, el intervalo de confianza puede quedar de manera simplificada así:

$$LCI \leq Y \leq LCS$$

Ejemplo 13.10 Para las temperaturas medias de $63^{\circ}F$ y $78^{\circ}F$ en el maratón de Nueva York, estime los tiempos en minutos de la ganadora y los respectivos intervalos de confianza con $\alpha = 0.95$.

Solución. Utilizando las sumatorias del ejemplo 13.2 y sustituyendo en la ecuación 13.7 se tiene la recta de regresión $\hat{t} = 140.249142 + 0.11908461 T$. Los datos siguientes son comunes para los dos valores de temperatura: $n = 21$, $\bar{x} = 63.048$. Como $\gamma = 1 - \alpha = 1 - 0.95 = 0.05$, $t_{0.95,19} = 2.093$; $s_e = 1.60322$;

Para el caso de $T = 63^{\circ}F$

$$\hat{t} = 0.11908461(63) + 140.249142 = 147.751472$$

$$\sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}} = \sqrt{1 + \frac{1}{21} + \frac{21(63 - 63.048)^2}{21 \times 85396 - 1324^2}} = 1.02352042$$

Los límites de confianza son:

$$LCI = 147.751472 - 2.093 \times 1.60322 \times 1.02352042 = 144.317009$$

$$LCI = 147.751472 + 2.093 \times 1.60322 \times 1.02352042 = 151.185935$$

Redondeando:

$$144 \leq Y \leq 151$$

Para el caso de $T = 78^{\circ}F$

$$\hat{t} = 0.11908461(78) + 140.249142 = 149.537742$$

$$\sqrt{1 + \frac{1}{21} + \frac{21(78 - 63.048)^2}{21 \times 85396 - 1324^2}} = 1.0788910$$

Los límites de confianza son:

$$LCI = 149.537742 - 2.093 \times 1.60322 \times 1.0788910 = 145.9174858$$

$$LCS = 149.537742 + 2.093 \times 1.60322 \times 1.0788910 = 153.1579974$$

Redondeando:

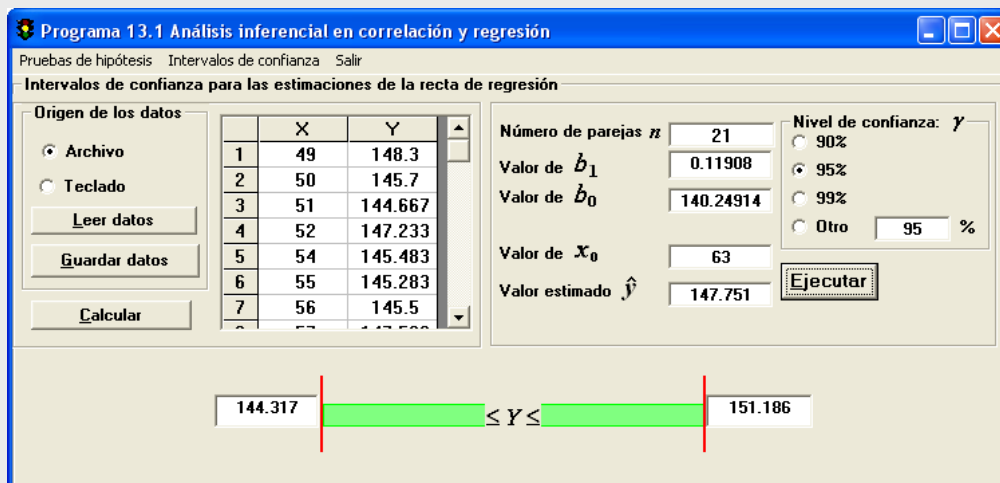
$$146 \leq Y \leq 153$$

Lo anterior significa que se puede asegurar con una confianza de 95% que los intervalos dados contienen el tiempo de la ganadora del maratón cuando este se corra a las temperaturas promedio especificadas.

Ejemplo 13.11 Resolver el ejemplo 13.10 utilizando el programa 13.1.

Solución

Se inicia el programa 13.1. Se selecciona el menú **Intervalos de confianza** y luego la opción **Estimación de la recta de regresión**. Una vez leídos los datos se usa el botón **Calcular**, se proporciona el valor de x_0 (63), se selecciona el nivel de confianza y se hace clic en el botón **Ejecutar**. El resultado es:



Se deja al lector el cálculo para $T = 78^\circ F$.

Al comparar los resultados del ejemplo 13.10, se observa que el intervalo de confianza de 95 % para $63^\circ F$ está $3.43^\circ F$ arriba y debajo de la línea de ajuste, mientras que para $78^\circ F$ está $3.45^\circ F$ arriba y debajo de la línea de ajuste. Esta comparación refleja la propiedad general que los intervalos de confianza son más angostos entre más cerca del valor medio se encuentre el valor que se quiere estimar. Al moverse a los extremos de la distribución de los puntos de la muestra, los intervalos son más anchos. Esta es una razón por la cual no debería usarse la línea de ajuste para predecir valores más allá de los datos extremos.

Si se calcula un intervalo de confianza de 95 % para todos los valores de T en el rango de los datos, se tendría geoméricamente una banda que se amplía hacia los extremos como se aprecia en la figura 13.29.

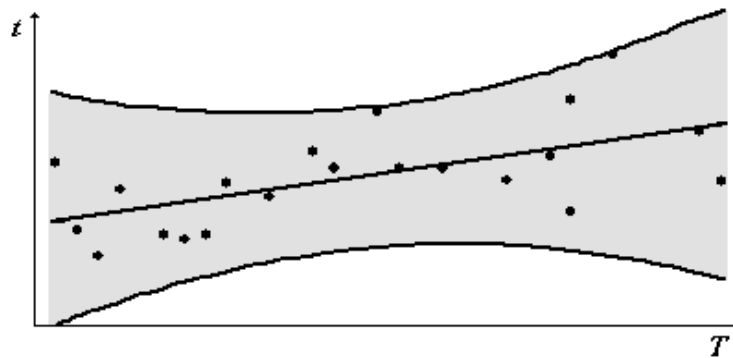


Figura 13.29 Banda de 95 % de confianza para los valores de predicción

Intervalos de confianza y pruebas de hipótesis para la pendiente β_1 de la recta de regresión

La línea de regresión $\hat{y} = b_0 + b_1x$ se calcula con la muestra de n parejas de puntos (x, y) , mientras que la línea de regresión $\hat{Y} = \beta_0 + \beta_1x$ se calcula, teóricamente, con la población de puntos (x, y) . El parámetro β_1 (pendiente de la línea de regresión de la población), resulta particularmente de interés ya que en muchas aplicaciones se requiere medir el cambio de y por unidad de cambio de x ; es decir, la velocidad de cambio de y con respecto a x . Asimismo, es importante ya que si su valor es cercano a cero, indica que posiblemente *no* hay relación entre las variables en estudio. Debido a que la línea de regresión muestral proporciona solamente una estimación b_1 de β_1 , conviene construir intervalos de confianza y pruebas de hipótesis para β_1 .

El estadístico de prueba requerido en ambos casos viene dado por la expresión

$$t = \frac{b_1 - \beta_1}{s_e / \sqrt{\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2}} \quad (13.13)$$

La expresión 13.13 sigue una distribución t de Student con $gl. = n - 2$.

La expresión $s_e / \sqrt{\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2}$ es conocida como el *error estándar* de b_1 . Se

emplea para los intervalos de confianza y para las pruebas de hipótesis de β_1 como se muestra en el ejemplo siguiente.

Ejemplo 13.12 Construya un intervalo de confianza de 95% para la pendiente de la recta de ajuste de los datos del Maratón (ver ejemplo13.10)

Solución Despejando en la ecuación 13.13 a β_1 y construyendo el intervalo:

$$b_1 - \frac{t_{\gamma, gl} s_e}{\sqrt{\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2}} \leq \beta_1 \leq b_1 + \frac{t_{\gamma, gl} s_e}{\sqrt{\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2}}$$

$$0.1191 - \frac{2.093(1.6032)}{\sqrt{85396 - \frac{1}{21}(1324)^2}} \leq \beta_1 \leq 0.1191 + \frac{2.093(1.6032)}{\sqrt{85396 - \frac{1}{21}(1324)^2}}$$

$$0.0425 \leq \beta_1 \leq 0.1957$$

Se puede emplear el programa 13.1 (ver ejemplo 13.11) seleccionando en **Intervalos de confianza** la opción **Pendiente de la recta de regresión**. Una vez proporcionados los datos y seleccionado el nivel de confianza se obtiene:

Programa 13.1 Análisis inferencial en correlación y regresión

Pruebas de hipótesis Intervalos de confianza Salir

Intervalos de confianza para la pendiente de la recta de regresión

Origen de los datos

- Archivo
- Teclado

Leer datos

Guardar datos

Calcular

	X	Y
1	49	148.3
2	50	145.7
3	51	144.667
4	52	147.233
5	54	145.483
6	55	145.283
7	56	145.5

Número de parejas n: 21

Valor de b_1 : 0.11908

Valor de b_0 : 140.24914

Nivel de confianza: γ

- 90%
- 95%
- 99%
- Otro: 95 %

Ejecutar

0.043 ≤ β_1 ≤ 0.196

Ejemplo 13.13 Realizar una prueba de hipótesis para la pendiente de la recta de ajuste de los datos del maratón.

Solución

El modelo estadístico es:

$$H_0 : \beta_1 \leq 0 \text{ (la pendiente es negativa)}$$

$$H_1 : \beta_1 > 0 \text{ (la pendiente es positiva)}$$

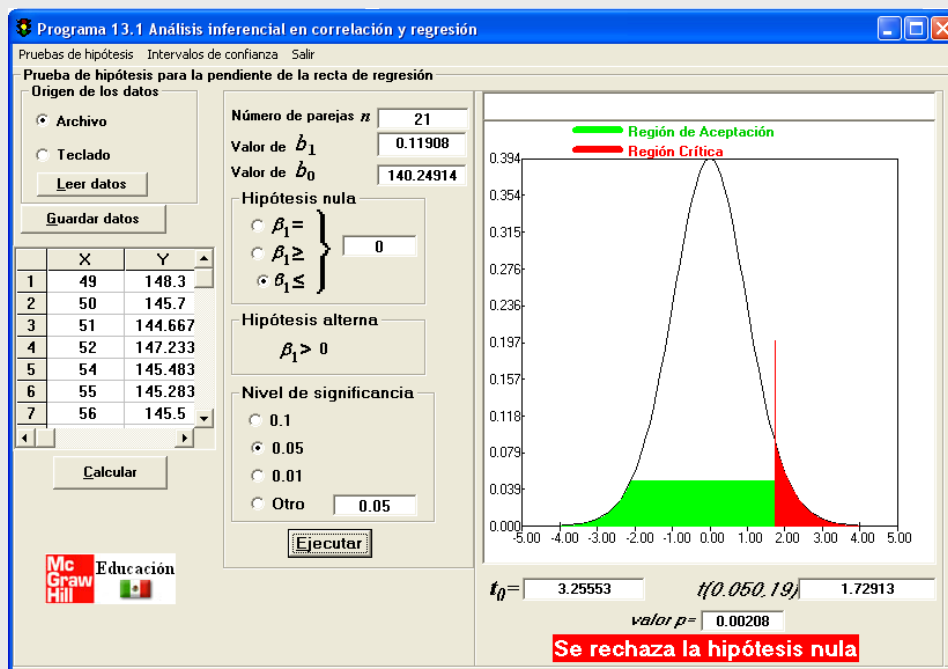
Con $b_1 = 0.1191$ y $n = 21$, el valor observado es:

$$t_o = \frac{0.1191}{1.6032 / \sqrt{85396 - \frac{1}{21}(1324)^2}} = 3.2560$$

Tomando el nivel de confianza $\alpha = 0.05$ y $gl = 21 - 2 = 19$, se obtienen como valor crítico a $t_{(0.95,19)} = 1.729$

Como el valor observado queda fuera de la región de aceptación, se rechaza H_0

Se puede emplear el programa 13.1 seleccionando en **Pruebas de hipótesis** la opción **Pendiente de la recta de regresión**. Una vez proporcionados los datos y seleccionado el modelo estadístico y el nivel de significancia se obtiene:



13.3 Regresión no lineal (funciones intrínsecamente lineales)

No siempre es conveniente ajustar una línea recta a un diagrama de dispersión. En algunos casos el diagrama perfila una línea *curva*. Si bien el diagrama es importante, no lo es menos la teoría o experiencia de la situación en estudio. Conjuntando estos elementos se puede advertir que la relación entre dos variables de interés sea curvilínea; algunos ejemplos típicos son las

reacciones químicas, el crecimiento poblacional, la relaciones entre gasto en publicidad y ventas, etcétera.

En tales casos, es importante analizar la posibilidad de usar un modelo matemático cuyos parámetros se puedan estimar con facilidad. Una clase importante de estos modelos está formada por las funciones “intrínsecamente lineales”. Un ejemplo típico de ellas es el de la función exponencial

$$y = b_0 e^{b_1 x} \quad (13.14)$$

Para ver el significado de la expresión “intrínsecamente lineal”, se toman logaritmos base e en ambos lados de la ecuación 13.14, quedando:

$$\ln(y) = \ln(b_0 e^{b_1 x})$$

Aplicando las propiedades de los logaritmos se llega a

$$\ln(y) = \ln(b_0) + b_1 x \quad (13.15)$$

Como y es una variable, también lo es $\ln(y)$, de modo que puede llamarse y' a esta “nueva” variable. Por otro lado, dado que b_0 es una constante, también lo es $\ln(b_0)$ y puede denotarse como b'_0 a la “nueva” constante. Sustituyendo en la ecuación anterior:

$$y' = b'_0 + b_1 x \quad (13.16)$$

La función exponencial 13.14 se ha transformado en una “nueva” función 13.16 cuya relevancia consiste en que es *lineal* y por tanto el que sus parámetros b'_0 y b_1 se puedan calcular en la forma vista en la sección anterior.

Una función $y = f(x)$ que relaciona a y con x es *intrínsecamente lineal*, si por medio de una transformación en x o en y o en ambas, la función se puede expresar en general como una función lineal $y' = b_0' + b_1'x'$, con x' = variable predictiva transformada, y' = variable respuesta transformada y parámetros b_0' y b_1' .

Actividad 13.10 Demostrar que la función exponencial general $y = b_0 a^{b_1 x}$, donde a es una constante conocida, es intrínsecamente lineal.

Cuatro de las funciones intrínsecamente lineales más empleadas se dan en la tabla 13.8. En los incisos a) y b) la transformación apropiada es logarítmica y en los incisos c) y d) es simplemente un cambio de variable.

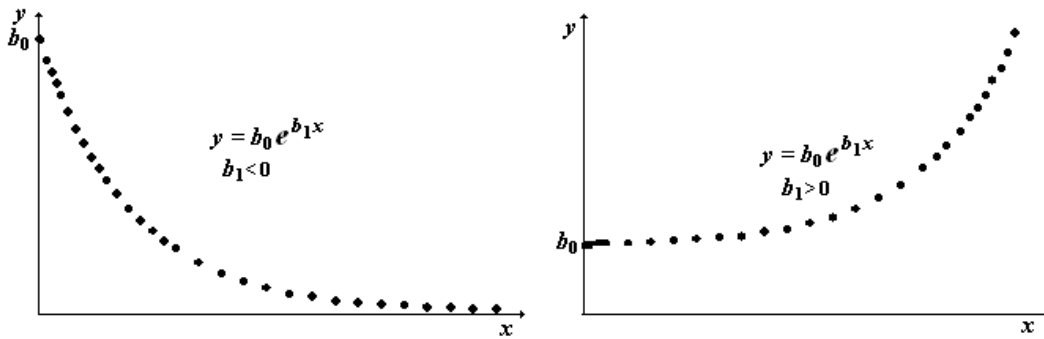
Tabla 13.8 Funciones intrínsecamente lineales más comunes*.

Función	Variable(s) y parámetro(s) transformado(s)	Función transformada
a) Exponencial: $y = b_0 e^{b_1 x}$	$y' = \ln(y)$, $b_0' = \ln(b_0)$	$y' = b_0' + b_1 x'$
b) Potencial: $y = b_0 x^{b_1}$	$y' = \log(y)$, $x' = \log(x)$, $b_0' = \log(b_0)$	$y' = b_0' + b_1 x'$
c) Logarítmica: $y = b_0 + b_1 \times \log(x)$	$x' = \log(x)$	$y = b_0 + b_1 x'$
d) Recíproca: $y = b_0 + b_1 \frac{1}{x}$	$x' = \frac{1}{x}$	$y = b_0 + b_1 x'$

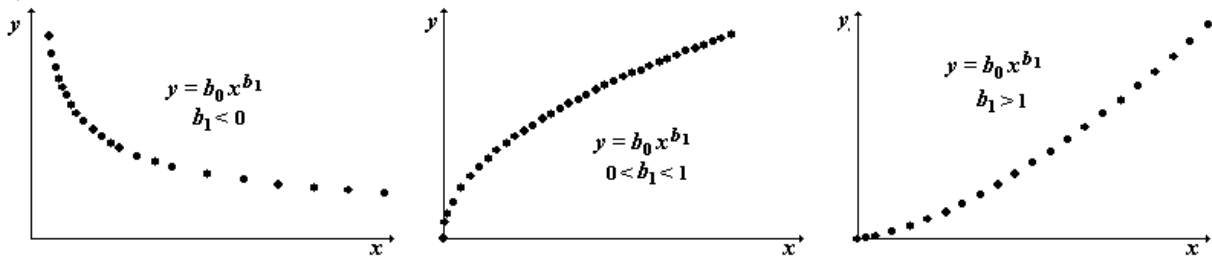
* Cuando aparece $\log(*)$, se puede usar ya sea el logaritmo base 10 o el logaritmo base e .

Las gráficas representativas de las cuatro funciones se ilustran en la figura 13.30. Tales gráficas corresponderían a correlaciones perfectas, por lo que sirven de modelos para comparar los diagramas de dispersión con que se trabaje. Se resuelven a continuación algunos ejemplos.

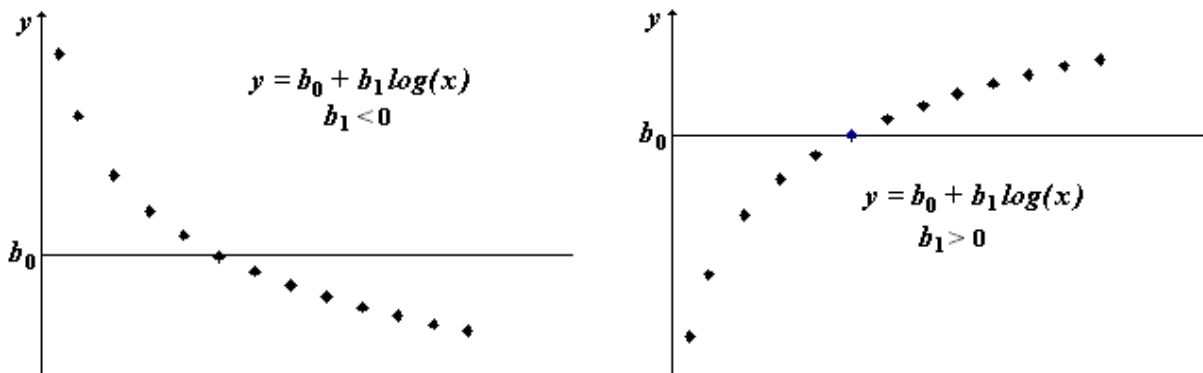
a) Exponencial



b) Potencial



c) Logarítmica



d) Recíproca

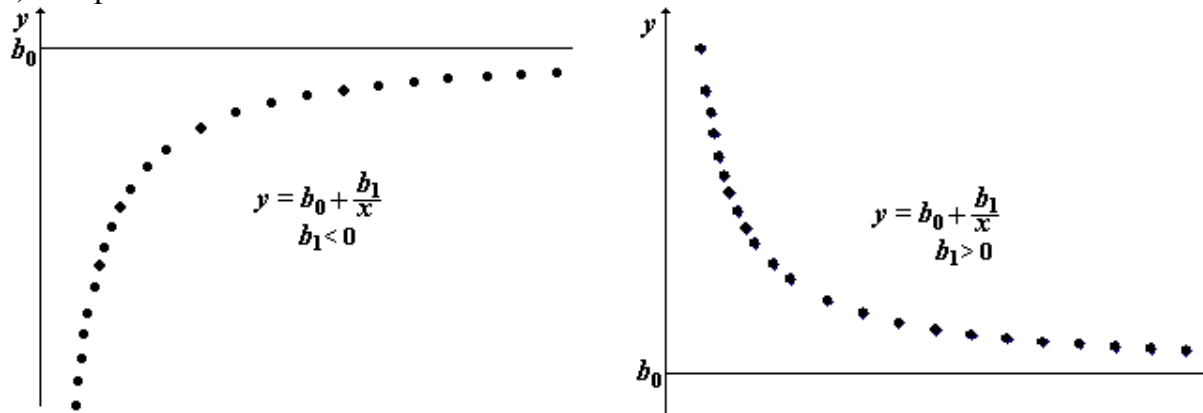


Figura 13.30 Correlaciones perfectas de funciones intrínsecamente lineales

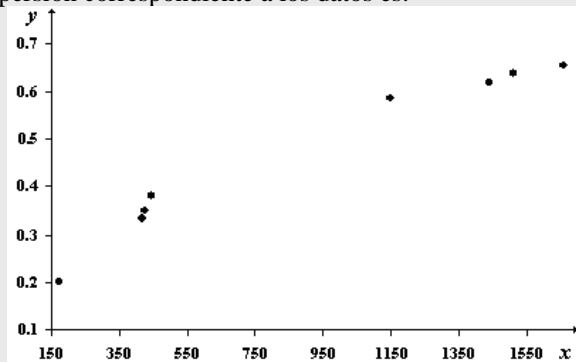
Ejemplo 13.14 La reacción química en fase líquida $A + B \rightarrow R + S$ se lleva a cabo en forma isotérmica (temperatura constante) a 25 °C en un reactor intermitente. Partiendo de las condiciones iniciales de concentración de los reactivos A y B: $C_{A0} = 0.054 \text{ mol/l}$ y $C_{B0} = 0.106 \text{ mol/l}$, los resultados experimentales son:

x (min)	174	418	426	444	1150	1440	1510	1660
y	0.203	0.335	0.35	0.383	0.588	0.618	0.638	0.655

x representa el tiempo transcurrido desde que se inicia la reacción y y el rendimiento o fracción del reactivo A que ha reaccionado al tiempo x . (Ancheyta J. J. y Valenzuela Z. M. A. Cinética Química para sistemas homogéneos. Dirección de Publicaciones, IPN (2002) p. 192-193)

Encontrar la función que mejor ajuste los datos experimentales.

Solución. La gráfica de dispersión correspondiente a los datos es:



El rendimiento y aumenta sustancialmente en los primeros minutos de la reacción, pero después los aumentos son moderados y más tarde resultan muy pequeños. Comparando con las gráficas de la figura 13.19 el modelo parece ser el logarítmico o el potencial. Se hará el desarrollo para el modelo logarítmico y en el problema 13.34 se pide el desarrollo del modelo potencial.

En el modelo logarítmico, de acuerdo a la tabla 13.8, se transforma solamente la variable x : $x' = \ln(x)$.

Para resolver el modelo lineal transformado $y = b_0 + b_1 x'$, se calculan los logaritmos de x para tener la nueva variable. Luego, se procede a realizar los cálculos tipo para la regresión lineal:

	x'	y	$(x')^2$	$x' y$
	5.1590553	0.203	26.6158516	1.04728823
	6.03548143	0.335	36.4270361	2.02188628
	6.05443935	0.35	36.6562358	2.11905377
	6.09582456	0.383	37.1590771	2.33470081
	7.04751722	0.588	49.667499	4.14394013
	7.27239839	0.618	52.8877784	4.49434221
	7.31986493	0.638	53.5804226	4.67007383
	7.41457288	0.655	54.975891	4.85654524
Sumatorias	52.3991541	3.77	347.969792	25.6878305

Sustituyendo las sumatorias en las ecuaciones 13.5 y 13.6 se llega a los parámetros:

$$b_0 = -0.8972743 \quad y \quad b_1 = 0.20893838$$

Ninguno de los parámetros se transforma, de modo que se sustituyen en la ecuación potencial, llegándose al resultado buscado:

$$y = b_0 + b_1 \times \log(x) = -0.8972743 + 0.20893838 \log(x)$$

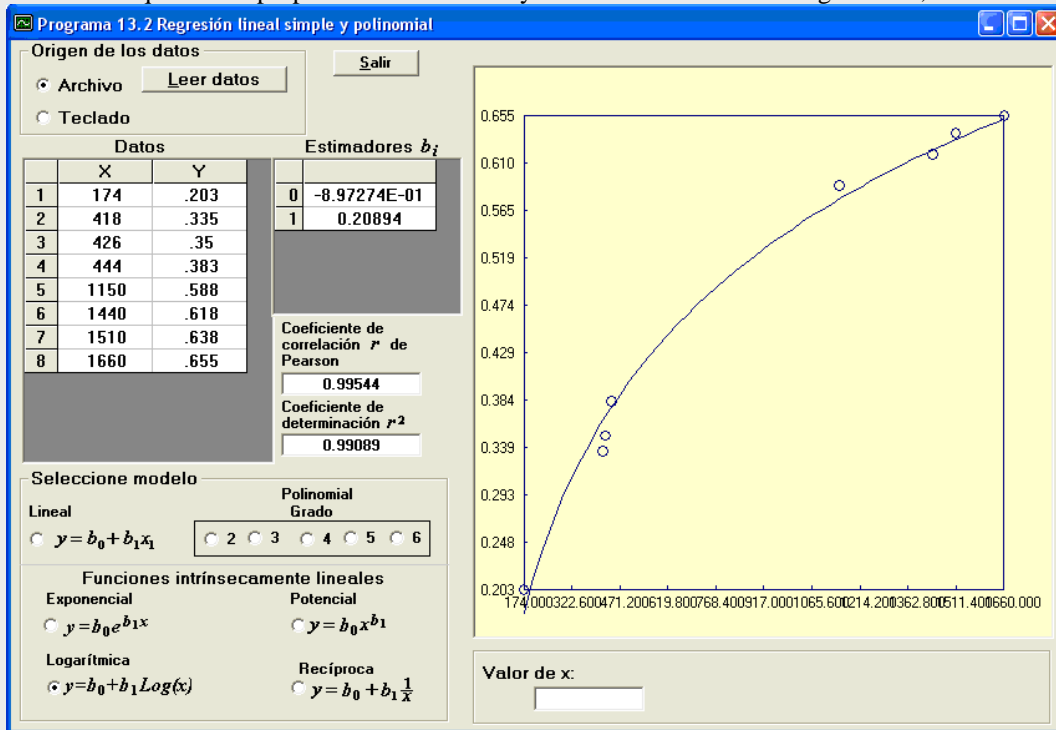
Se calcula ahora el coeficiente de determinación de acuerdo a la ecuación 13.10

$$r^2 = \frac{\text{Variación explicada}}{\text{Variación total}} = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 0.9909$$

Ejemplo 13.15 Resolver el ejemplo 13.14 utilizando el programa 13.2.

Solución

Como se ha visto en los ejemplos 13.3 y 13.9, el programa 13.2 dispone de las funciones intrínsecamente lineales. Una vez que se han proporcionado los datos y seleccionado el modelo logarítmico, se obtiene:



Como en el caso lineal se dispone de la característica de que se pueden obtener estimaciones de y para algún valor de x dado.

Análisis gráfico del proceso de transformación

El proceso de transformación de algunas funciones intrínsecamente lineales puede interpretarse gráficamente. Considere, por ejemplo, el diagrama de dispersión correspondiente a los datos originales del ejemplo 13.14 (vea el inciso *a*) de la figura 13.31). Si ahora se grafican los mismos datos pero en un sistema coordenado semilogarítmico; esto es, en el eje de las abscisas se tiene una escala logarítmica y en el eje vertical una escala decimal, se obtiene el diagrama de dispersión del inciso *b*) de la figura 13.31, donde ¡los puntos parecen tener un patrón lineal!

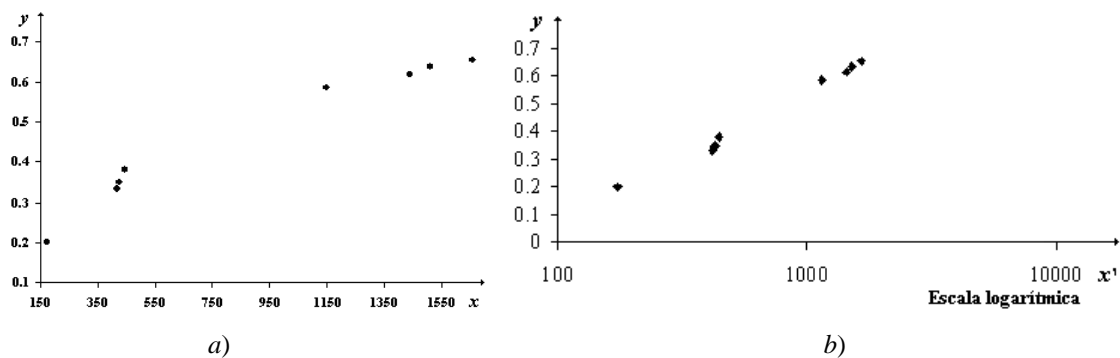


Figura 13.31 Diagramas de dispersión en escalas decimales y semilogarítmica

La escala logarítmica en el eje horizontal tiene el mismo efecto que la transformación analítica de tomar $x' = \log(x)$. En el caso de la función potencial se requeriría graficar en un sistema cartesiano en el que ambos ejes tuvieran escalas logarítmicas para conseguir la transformación deseada (ver tabla 13.8).

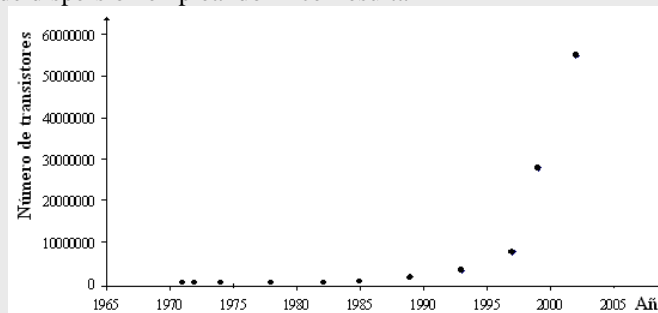
La graficación con escalas semilogarítmica y logarítmicas puede llevarse a cabo con Excel. Se realiza el diagrama de dispersión de los datos originales siguiendo las instrucciones dadas anteriormente; luego, se hace doble clic sobre el eje x del diagrama resultante con lo que aparecerá el cuadro de diálogo **Formato de ejes** (ver figura 13.20). Se hace clic en **Escala logarítmica**. En el caso de escalas logarítmicas se repite la operación anterior con el eje y .

Ejemplo 13.16 La “ley de Moore” establecida en 1965 dice: *cada 18 meses la potencia de las computadoras se duplica*. Este dato puede parecer sorprendente pero el caso es que la Ley de Moore cumplió 40 años en vigor el 19 de abril de 2005. (<http://petra.euitio.uniovi.es/~arrai/historia/trilobytes/5-Moore%20y%20la%20ley%20de%20Moore/Moore.htm>)

Procesador	Año	Nº de Transistores
4004	1971	2250
8008	1972	3500
8080	1974	6000
8086	1978	29000
286	1982	134000
386	1985	275000
486DX	1989	1200000
Pentium	1993	3100000
Pentium II	1997	7500000
Pentium III	1999	28000000
Pentium4	2002	55000000

Encontrar la función matemática que mejor ajuste la ley de Moore empleando Excel.

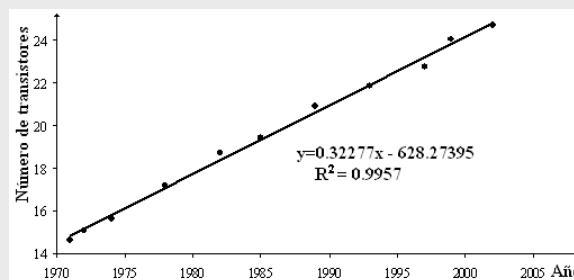
Solución. El diagrama de dispersión empleando Excel resulta



La escala vertical de la gráfica tiene un rango muy amplio, de tal modo que parecieran iguales los primeros 5 valores. Esto naturalmente no es así por lo que debemos ser precavidos en la lectura de la gráfica. Por otro

lado, el modelo se asemeja a la función exponencial: $y = b_0 e^{b_1 x}$

En Excel se señala el gráfico, se usa el menú **Gráfico** y luego **Agregar línea de tendencia**. Se elige **Exponencial** y se solicita **Presentar ecuación en el gráfico** y **Presentar el valor de R cuadrado en el gráfico**. El resultado es:



Excel hace internamente las transformaciones $y' = \ln(y)$ y $b'_0 = \ln(b_0)$, por lo que habrá que retransformar el coeficiente $b'_0 = \ln(b_0)$ para obtener $b_0 = e^{b'_0} = 1.393447E - 273$ y la ecuación resultante es:

$$y = 1.393447 \times 10^{-273} e^{0.32277x}$$

Actividad 13.11 Resolver el ejemplo 13.16 empleando el programa 13.2 del libro. Los datos son intrínsecamente lineales. Para ver esto considérese la población de México en diferentes años de los últimos tres siglos:

Tabla 13.9 Población de México en los últimos tres siglos

AÑO	POBLACIÓN EN MILLONES	AÑO	POBLACIÓN EN MILLONES
1836	7.8	1940	19.6
1846	7.5	1950	25.8
1858	8.3	1960	34.4
1862	8.4	1970	48.2
1872	9	1980	66.8
1876	9.5	1990	81.2
1900	13.6	2000	97.5
1910	15.2	2009	107.6
1929	15.6		

La gráfica de dispersión correspondiente es:

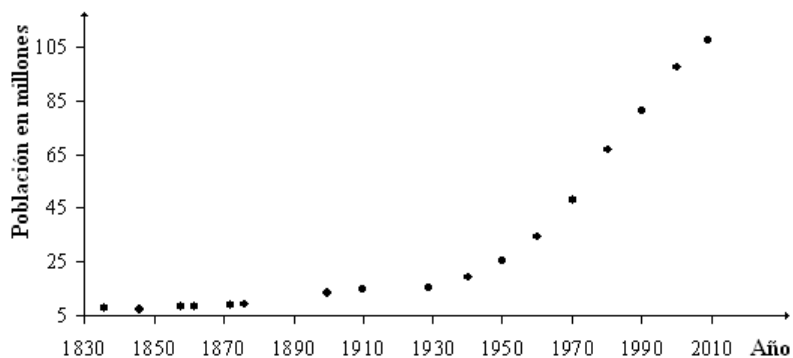


Figura 13.32 Diagrama de dispersión de la población de México

Al igual que en el ejemplo 13.14, la escala del eje vertical de la figura 13.32 puede causar interpretaciones erróneas del diagrama por lo que es recomendable interactuar con la tabla 13.9. El análisis de la gráfica es interesante; por ejemplo, el hecho de que la población de 1910 a 1929 sea prácticamente la misma se explica por la muerte de 2 millones de personas durante la revolución. En 1930, sin embargo, se aprecia el inicio de una curvatura (cambio de pendiente del “patrón lineal” que se venía dando), debiéndose a una mayor estabilidad política y económica del país. En 2000, sin embargo, se asoma otra curvatura que se puede explicar como una estabilización del aumento de población debido, entre otros factores, al control de la natalidad. Esto último obliga a pensar en una función diferente a las vistas ya que en ninguno de los casos se tienen dos curvaturas.

Ventana al conocimiento 3

México ocupa el 5º lugar dentro de los países de América con mayor prevalencia de diabetes y el 10.7% de su población la presenta*, siendo el más alto de Latinoamérica. Uno de los factores de riesgo para presentarla es el sobrepeso y la obesidad que se define como un exceso de tejido adiposo (graso).

El estudio de laboratorio para diagnosticar a una persona con diabetes o “prediabetes” (que está en riesgo a desarrollarla) consiste en tomar su glucosa (azúcar) en sangre en ayuno y durante 2 horas después de una carga de glucosa. Existen también estudios en los que se monitorea la glucosa e insulina durante 3 ó 5 horas y que permiten diagnosticar “prediabetes”. Esto resulta costoso para el sector salud e invasivo para el paciente.

En el Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán se tienen en curso investigaciones para establecer una relación matemática que, a través de medidas antropométricas (peso, estatura, grasa corporal, circunferencia de cuello, cintura, cadera, muslo, pantorrilla y brazo), puedan determinar si una persona es prediabética. Lo anterior requeriría de sólo unos minutos, resultando económico y menos molesto para el paciente.

*Fuente: Organización Panamericana de la Salud (OPS)



Instrumentos de medición antropométrica

13.4 Regresión multilínea

En estadística hay muchos ejemplos donde una variable puede predecirse con exactitud en términos de solamente otra variable. Sin embargo, las predicciones pueden mejorar si se considera *información relevante* adicional. En el caso del maratón, por ejemplo, si además de tomar la temperatura media se incluyera la velocidad del viento y la humedad relativa, podría esperarse una mejor predicción del tiempo de la ganadora o ganador.

Existen además situaciones en las que resulta indispensable considerar de inicio múltiples variables. Por ejemplo, si se desea estimar el *crecimiento económico* (y) de un estado o un país, es necesario tomar en cuenta elementos como la *inversión extranjera directa* (x_1), las *exportaciones* (x_2), el *capital humano* (x_3) –proporción de la población total del estado o país con educación media y superior –y la *captación de la banca comercial* (x_3). La relación funcional en este caso se representa como $y = f(x_1, x_2, x_3)$ y en general como $y = f(x_1, x_2, \dots, x_k)$.

El modelo más sencillo que relaciona la variable respuesta “y” con las variables predictivas x_i , $i = 1, 2, \dots, k$, es una generalización del modelo lineal por lo que es llamado *multilineal*. Su representación es:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k \quad (13.17)$$

Para encontrar los parámetros $b_0, b_1, b_2, \dots, b_k$ se utiliza la técnica de ajuste por mínimos cuadrados empleando los n datos muestrales. El método de ajuste por mínimos cuadrados para el caso multilineal es una generalización del caso lineal (ver apéndice G). Se trata por tanto de minimizar la función

$$f(b_0, b_1, b_2, \dots, b_k) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (b_0 + b_1x_{1,i} + b_2x_{2,i} + \dots + b_kx_{k,i} - y_i)^2$$

Derivando parcialmente la función con respecto a $b_0, b_1, b_2, \dots, b_k$ sucesivamente e igualando a cero las derivadas resultantes, se llega a un sistema de $k + 1$ ecuaciones lineales conocido como ecuaciones normales para el modelo multilineal.

$$\begin{array}{rcccccc} b_0n & + b_1 \sum_{i=1}^n x_{1,i} & + b_2 \sum_{i=1}^n x_{2,i} & + \dots & + b_k \sum_{i=1}^n x_{k,i} & = \sum_{i=1}^n y_i \\ b_0 \sum_{i=1}^n x_{1,i} & + b_1 \sum_{i=1}^n x_{1,i}^2 & + b_2 \sum_{i=1}^n x_{1,i}x_{2,i} & + \dots & + b_k \sum_{i=1}^n x_{1,i}x_{k,i} & = \sum_{i=1}^n x_{1,i}y_i \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ b_0 \sum_{i=1}^n x_{k,i} & + b_1 \sum_{i=1}^n x_{k,i}x_{1,i} & + b_2 \sum_{i=1}^n x_{k,i}x_{2,i} & + \dots & + b_k \sum_{i=1}^n x_{k,i}^2 & = \sum_{i=1}^n x_{k,i}y_i \end{array} \quad (13.18)$$

Resolviendo el sistema por alguno de los métodos conocidos (regla de Cramer, eliminación de Gauss, etc.) se encuentran los parámetros $b_0, b_1, b_2, \dots, b_k$ y se sustituyen en la ecuación 13.17.

Como en casos previos, se considerará una situación de estudio familiar a los lectores.

Situación de estudio: Calorías en los alimentos

La cantidad de energía (calorías) en una porción de alimento puede determinarse a partir de los gramos de grasa, proteínas y carbohidratos que contiene. La idea es *descubrir* esta relación a

partir de información recolectada de alimentos comunes en cualquier alacena o en el supermercado. Dentro de la información nutrimental en la etiqueta de cada producto pueden leerse las cantidades relevantes: Calorías (y), grasas (x_1), proteínas (x_2) y carbohidratos (x_3) por cada porción de alimento. A continuación se dan los datos encontrados en 11 alimentos comunes. Se sugiere reunir sus propios datos y, trabajando en equipo, realizar los cálculos correspondientes.

Tabla 13.10. Datos nutrimentales de alimentos comunes

Alimento	Calorías (Kcal)	Grasa (g.)	Proteínas (g.)	Carbohidratos (g.)
Leche Light (LALA)	100	2.5	7.8	11.6
Salchichas de pavo (KIR)	180	12	10	8
Agua	0	0	0	0
Pollo fresco (Bachoco)	215	15.10	18.60	0
Avena (Quaker)	152	2.7	5	27
Salsa Catsup (Del Monte)	112	0	1	27
Mantequilla (Becel)	27	3	0.0	0.0
Cereal de caja (La Lechera Flakes)	112	0.15	1.6	26.3
Filete de pescado (Sierra Madre)	200	11.0	25.5	0.0
Jamón de pierna (Oscar Mayer)	49	1.0	10.0	0.0
Arrachera marinada (Rancho San Fco.)	99	3	17	1

El modelo que se ensaya es el multilíneo, teniéndose:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

Considerando el agua como referente importante, podría tomarse $b_0 = 0$. Lo anterior resulta lógico ya que en ese caso $x_1 = x_2 = x_3 = 0$ y el valor de predicción \hat{y} debería ser también cero. Modificando el modelo:

$$\hat{y} = b_1x_1 + b_2x_2 + b_3x_3$$

Organizando los cálculos de acuerdo a las ecuaciones 13.18:

Tabla 13.11 Organización de los datos para el cálculo de b_1 , b_2 y b_3

y	x_1	x_2	x_3	$x_1 y$	$x_2 y$	$x_3 y$	$x_1 x_2$	$x_1 x_3$	$x_2 x_3$	x_1^2	x_2^2	x_3^2
100	2.5	7.8	11.6	250	780	1160	19.5	29	90.48	6.25	60.84	134.56
180	12	10	8	2160	1800	1440	120	96	80	144	100	64
0	0	0	0	0	0	0	0	0	0	0	0	0
215	15.1	18.6	0	3246.5	3999	0	280.86	0	0	228.01	345.96	0
152	2.7	5	27	410.4	760	4104	13.5	72.9	135	7.29	25	729
112	0	1	27	0	112	3024	0	0	27	0	1	729
27	3	0	0	81	0	0	0	0	0	9	0	0
112	0.15	1.6	26.3	16.8	179.2	2945.6	0.24	3.945	42.08	0.0225	2.56	691.69
200	11	25.5	0	2200	5100	0	280.5	0	0	121	650.25	0
49	1	10	0	49	490	0	10	0	0	1	100	0
99	3	17	1	297	1683	99	51	3	17	9	289	1
1246	50.45	96.5	100.9	8710.7	14903	12773	775.6	204.85	391.56	525.57	1574.6	2349.3

Sustituyendo valores en la ecuación 13.18:

$$\begin{aligned}
 525.573b_1 + 775.6b_2 + 204.845b_3 &= 8710.7 \\
 775.6b_1 + 1574.61b_2 + 391.56b_3 &= 14903.2 \\
 204.845b_1 + 391.56b_2 + 2349.25b_3 &= 12772.6
 \end{aligned}$$

Resolviendo (se sugiere utilizar Excel o una calculadora científica) se obtiene:

$$b_1 = 9.21187 ; b_2 = 3.93821 \text{ y } b_3 = 3.97725$$

Sustituyendo en el modelo multilineal:

$$\hat{y} = 9.21187x_1 + 3.93821x_2 + 3.97725x_3$$

En realidad, la relación de las calorías de una porción de alimento está relacionada con la cantidad de grasa, proteínas y carbohidratos que contiene, de la siguiente manera:

$$\text{Calorias} = 9(\text{Grasa}) + 4(\text{Proteina}) + 4(\text{Carbohidratos})$$

El acercamiento que se obtuvo (redondeando a enteros se obtienen los mismos valores) resulta bueno, considerando que se emplearon solamente 11 alimentos.

Ejemplo 13.17 Resolver el caso de las calorías en los alimentos con el programa 13.3 del libro.

Solución

Inicie el **programa 13.3** y haga clic en el botón **Leer datos** (la opción predeterminada es la lectura de un archivo). Use el navegador para ubicar el archivo **Alimentos.dat** y haga clic en el botón **Aceptar**. Se obtiene:

Programa 13.3 Regresión lineal múltiple

Origen de los datos

Archivo Teclado

Leer datos Salir

Mc Graw Hill Educación

Datos

	Y	X1	X2	X3
1	100	2.5	7.8	11.6
2	180	12	10	8
3	0	0	0	0
4	215	15.1	18.6	0
5	152	2.7	5	27
6	112	0	1	27
7	27	3	0	0
8	112	.15	1.6	26.3
9	200	11	25.5	0
10	49	1	10	0

Selección de modelo

$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$

$\hat{y} = b_1x_1 + b_2x_2 + b_3x_3$

Seleccione el modelo (en este caso se usa el modelo que no incluye b_0). El resultado es:

Programa 13.3 Regresión lineal múltiple

Origen de los datos

Archivo Teclado

Leer datos Salir

Mc Graw Hill Educación

Datos

	Y	X1	X2	X3
1	100	2.5	7.8	11.6
2	180	12	10	8
3	0	0	0	0
4	215	15.1	18.6	0
5	152	2.7	5	27
6	112	0	1	27
7	27	3	0	0
8	112	.15	1.6	26.3
9	200	11	25.5	0
10	49	1	10	0

Estimadores b_i

i	b
1	9.21187
2	3.93821
3	3.97725

Selección de modelo

$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$

$\hat{y} = b_1x_1 + b_2x_2 + b_3x_3$

Breve introducción al caso de funciones intrínsecamente multilineales

Al igual que en el caso de funciones de dos variables no lineales, es posible para las funciones no lineales de múltiples variables, transformarlas para llegar a un modelo multilinear. Se resuelve a continuación un ejemplo como ilustración.

Ejemplo 13.18 El porcentaje de impurezas que se encuentra, a varias temperaturas y tiempos de esterilización, en una reacción asociada con la fabricación de cierta bebida, está representado por los datos siguientes:

Porcentaje de impurezas	Temperatura °C	Tiempo de esterilización (min)
y	x_1	x_2
14.05	75	15
14.93	75	15
16.56	75	20
15.87	75	20
22.41	75	25
21.66	75	25
10.55	100	15
9.48	100	15
13.63	100	20
11.75	100	20
18.55	100	25
17.98	100	25
7.55	125	15
6.59	125	15
9.23	125	20
8.78	125	20
15.93	125	25
16.44	125	25

Estime los coeficientes en el modelo no lineal siguiente:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_1^2 + b_4x_2^2 + b_5x_1x_2$$

Solución

Se trata de un modelo no lineal múltiple. Al igual que en el caso no lineal de una variable, puede transformarse en multilinear mediante ciertas transformaciones. En este caso cambiando las variables de la siguiente manera:

$$x_3 = x_1^2, \quad x_4 = x_2^2 \quad \text{y} \quad x_5 = x_1x_2$$

Sustituyendo en el modelo propuesto se tiene:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5$$

Los parámetros b_0, b_1, \dots, b_5 pueden obtenerse a partir del sistema 13.18 con $k = 5$.

La organización de los cálculos en este caso queda:

y	14.05	14.93	16.56	15.87	22.41	21.66	10.55	...
x_1	75	75	75	75	75	75	100	...
x_2	15	15	20	20	25	25	15	...
x_3	5625	5625	5625	5625	5625	5625	10000	
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots

Sustituyendo las sumatorias resultantes y el valor de n en la ecuación 13.18 se llega a

$$\begin{aligned}
18b_0 + 1800b_1 + 360b_2 + 187500b_3 + 7500b_4 + 36000b_5 &= 251.94 \\
1800b_0 + 187500b_1 + 36000b_2 + 20250000b_3 + 750000b_4 + 3750000b_5 &= 24170 \\
360b_0 + 36000b_1 + 7500b_2 + 3750000b_3 + 162000b_4 + 750000b_5 &= 5287.9 \\
187500b_0 + 20250000b_1 + 3750000b_2 + 2254687500b_3 + 78125000b_4 + 405000000b_5 &= 2420850 \\
7500b_0 + 750000b_1 + 162000b_2 + 78125000b_3 + 3607500b_4 + 16200000b_5 &= 115143 \\
36000b_0 + 3750000b_1 + 750000b_2 + 405000000b_3 + 16200000b_4 + 78125000b_5 &= 508702.5
\end{aligned}$$

Resolviendo con Excel se obtiene:

$$b_0 = 56.423333, b_1 = -0.3625333, b_2 = -2.7476667, b_3 = 0.000816, b_4 = 0.081600 \text{ y } b_5 = 0.003140$$

Sustituyendo:

$$y = 56.423333 - 0.3625333x_1 - 2.7476667x_2 + 0.000816x_3 + 0.0816x_4 + 0.00314x_5$$

Una vez obtenidos los coeficientes, puede estimarse el porcentaje de impurezas correspondiente a un tiempo de esterilización y una temperatura dados; por ejemplo, a un tiempo de 19 minutos y una temperatura de 80 °C se tiene un porcentaje de impurezas de:

$$\begin{aligned}
y &= 56.423333 - 0.3625333(80) - 2.7476667(19) + 0.000816(80)^2 + 0.0816(19)^2 + 0.00314(80 \times 19) \\
y &= 14.67
\end{aligned}$$

Ejemplo 13.19 Resolver el ejemplo 13.18 con el programa 13.3.

Solución

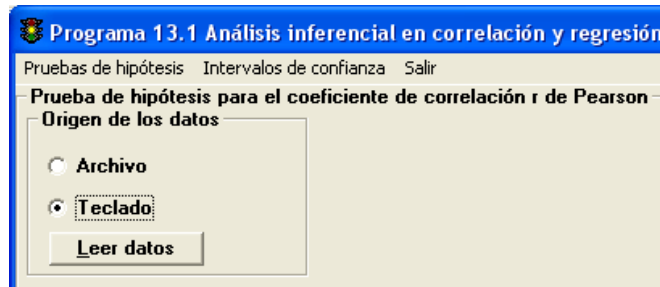
Considerando las transformaciones de las variables y organizando los datos se obtiene:

y	x ₁	x ₂	x ₃	x ₄	x ₅
14.05	75	15	5625	225	1125
14.93	75	15	5625	225	1125
16.56	75	20	5625	400	1500
⋮	⋮	⋮	⋮	⋮	⋮
15.93	125	25	15625	625	3125
16.44	125	25	15625	625	3125

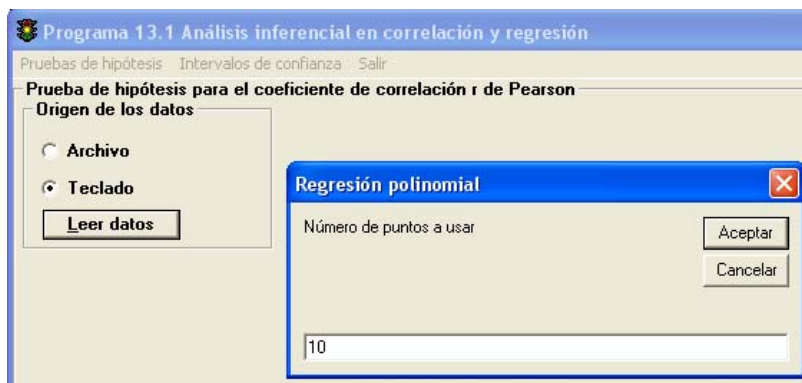
Resolviendo con el programa 13.3 (ver ejemplo 13.17) se tiene:

Creación de un archivo de datos

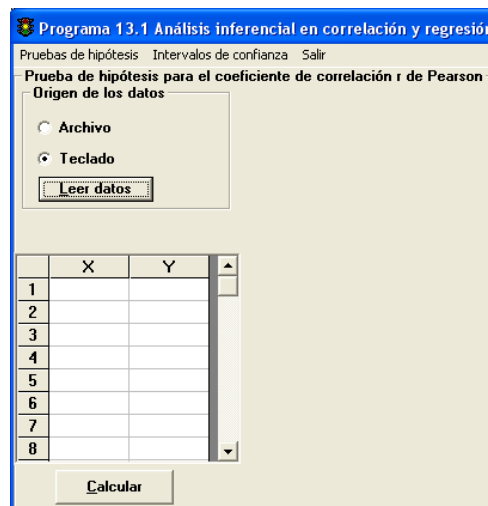
Inicie el programa 13.1 o 13.2 o 13.3 y haga clic en la opción **Teclado**



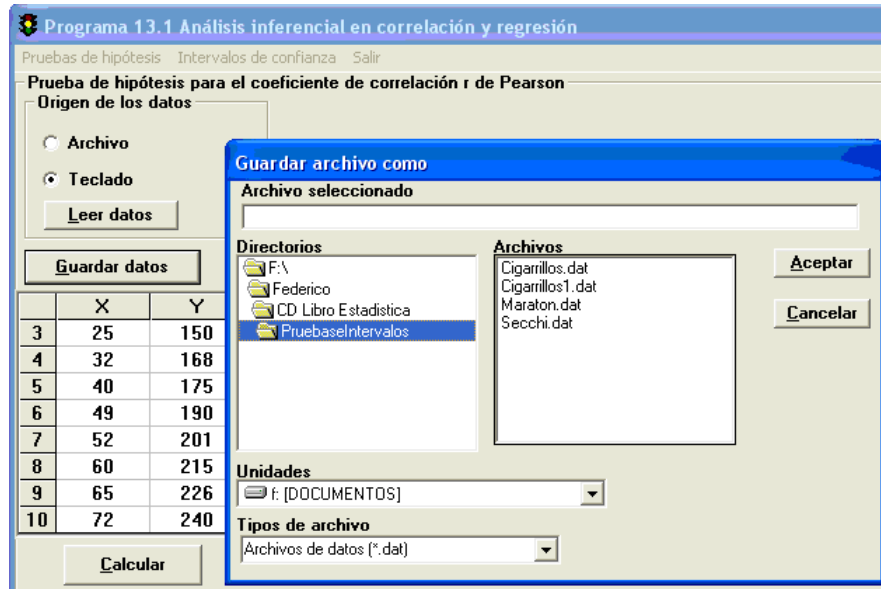
Haga clic en el botón **Leer datos**. Aparecerá una ventana solicitando el **Número de puntos a usar**:



Escriba el número de parejas de datos (x, y) de su ejemplo y oprima la tecla Enter o haga clic en el botón **Aceptar**. Aparecerá una tabla donde podrá escribir los valores numéricos de sus parejas de datos. Además aparecerá el botón **Calcular**.



Una vez que empiece a escribir sus datos aparecerá el botón **Guardar datos**. Cuando haya escrito todos sus datos haga clic en el botón **Guardar datos**. Aparecerá una ventana del navegador de Windows (la unidad, el directorio y los archivos dependerán de la computadora que se esté usando):



Seleccione la unidad y el directorio donde desee guardar sus datos y escriba el nombre del archivo que desea crear con la extensión **.dat**. Por último oprima la tecla Enter o haga clic en el botón **Aceptar**. En caso de que el archivo ya exista, el programa preguntará si desea sobrescribirlo. En adelante podrá utilizar el archivo para posteriores usos del programa.

Glosario

Análisis de correlación	Técnica estadística para establecer el grado de asociación o correlación entre dos o más variables de una población a partir de una muestra aleatoria.
Análisis de regresión	Técnica estadística para establecer la relación funcional entre dos o más variables a partir de una muestra de la población.
Análisis inferencial en correlación y regresión	Estimación de intervalos y pruebas de hipótesis del coeficiente de correlación, de los parámetros de la recta de regresión y de los valores de predicción obtenidos con ella.
Coefficiente de correlación r de Pearson	Medida numérica del tipo y grado de correlación lineal entre dos variables cuantitativas, que toma valores entre -1 y +1. Los valores cercanos a +1 indican una asociación o correlación <i>positiva fuerte</i> y los cercanos a -1 una asociación o correlación <i>negativa fuerte</i> . Los valores cercanos a cero indican <i>no</i> asociación o correlación.
Coefficiente de determinación r^2	Es la razón de la variación explicada sobre la variación total. Es decir, r^2 es la fracción de la variación total en y que puede explicarse usando el modelo lineal de regresión $\hat{y} = b_0 + b_1x$: $r^2 = \frac{\text{Variación explicada}}{\text{Variación total}}$
Desviación explicada	Diferencia entre el valor obtenido usando la línea de ajuste por mínimos cuadrados \hat{y} y el valor medio de los valores observados \bar{y} : $\hat{y}_i - \bar{y}$.
Desviación no explicada o residual	Diferencia entre el valor observado de la variable respuesta y_i y el valor correspondiente obtenido usando la línea de ajuste por mínimos cuadrados \hat{y} : $y_i - \hat{y}_i$.
Desviación total	Diferencia entre el valor observado de la variable respuesta y_i y el valor medio de los valores observados \bar{y} : $y_i - \bar{y}$.
Diagrama de dispersión	Diagrama de puntos bivariable empleado para descubrir patrones de asociación o correlación entre dos variables aleatorias cuantitativas. Una variable se representa sobre el eje horizontal y la otra sobre el eje vertical.
Error estándar de estimación	Medida de la dispersión de la muestra bivariable empleando la variación no explicada. Se representa como s_e y corresponde a la desviación estándar para una variable.
Factor variable o de confusión	Es una variable que distorsiona la medida de correlación entre otras dos variables aleatorias. El resultado de la presencia de una variable de confusión puede ser la observación de un efecto donde en

	<p>realidad no existe o la exageración de una correlación real (confusión positiva) o, por el contrario, la atenuación de una correlación real e incluso una inversión del sentido de una correlación real (confusión negativa).</p>
Función intrínsecamente lineal	<p>Una función $y = f(x)$ que relaciona a y con x es <i>intrínsecamente lineal</i>, si por medio de una transformación en x o en y o en ambas, la función se puede expresar en general como una función lineal $y' = b_0 + b_1x'$, con $x' =$ variable predictiva transformada, $y' =$ variable respuesta transformada y parámetros b_0 y b_1.</p>
Función multilínea	<p>El modelo más sencillo que relaciona la variable respuesta “y” con las variables predictivas $x_i, i = 1, 2, \dots, k$. Es una generalización del modelo lineal por lo que es llamado multilínea. Su representación es: $\hat{y} = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k$.</p>
Punto central o centroide	<p>Punto correspondiente a las medias de los datos (puntos) muestrales: (\bar{x}, \bar{y}). Se trazan por él líneas paralelas a los ejes dividiendo el diagrama de dispersión en cuatro regiones.</p>
Recta de regresión o ajuste por mínimos cuadrados	<p>Línea recta $\hat{y} = b_0 + b_1x$ trazada entre los puntos de un diagrama de dispersión basada en un proceso de minimización de $\sum_{i=1}^n (y_i - b_0 - b_1x)^2$.</p>
Variable explicativa o predictiva	<p>Variable aleatoria que puede verse como un predictor potencial y algunas veces su valor es seleccionado por el investigador. Se representa generalmente por x.</p>
Variable respuesta	<p>Es la variable aleatoria potencialmente predecible. Generalmente se representa por y.</p>
Variación explicada	<p>Suma de las desviaciones explicadas elevadas al cuadrado sobre todos los puntos de la muestra: $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$</p>
Variación no explicada o residual	<p>Suma de las desviaciones no explicadas elevadas al cuadrado sobre todos los puntos de la muestra: $\sum_{i=1}^n (y_i - \hat{y}_i)^2$</p>
Variación total	<p>Suma de las desviaciones totales elevadas al cuadrado sobre todos los puntos de la muestra: $\sum_{i=1}^n (y_i - \bar{y})^2$</p>

Tabla 13.12 Fórmulas importantes

Descripción	Ecuación
Coefficiente de correlación c	$c = \frac{n(\text{I}) + n(\text{III}) - n(\text{II}) - n(\text{IV})}{n}$
Coefficiente de correlación r de Pearson	$r = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \sqrt{n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i\right)^2}}$
Estadístico de prueba para el parámetro ρ	$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}, \quad \text{con } gl = n-2$
Ordenada al origen de la recta de regresión	$b_0 = \frac{\left(\sum_{i=1}^n y_i\right) \left(\sum_{i=1}^n (x_i)^2\right) - \left(\sum_{i=1}^n x_i\right) \left(\sum_{i=1}^n x_i y_i\right)}{n \sum_{i=1}^n (x_i)^2 - \left(\sum_{i=1}^n x_i\right)^2}$
Pendiente de la recta de regresión	$b_1 = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n y_i\right) \left(\sum_{i=1}^n x_i\right)}{n \sum_{i=1}^n (x_i)^2 - \left(\sum_{i=1}^n x_i\right)^2}$
Recta de regresión	$\hat{y} = \frac{\left(\sum_{i=1}^n y_i\right) \left(\sum_{i=1}^n (x_i)^2\right) - \left(\sum_{i=1}^n x_i\right) \left(\sum_{i=1}^n y_i x_i\right)}{n \sum_{i=1}^n (x_i)^2 - \left(\sum_{i=1}^n x_i\right)^2} + \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n y_i\right) \left(\sum_{i=1}^n x_i\right)}{n \sum_{i=1}^n (x_i)^2 - \left(\sum_{i=1}^n x_i\right)^2} \cdot x$
Relación entre las variaciones	$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$ $\left(\begin{array}{c} \text{Variación} \\ \text{total} \end{array}\right) = \left(\begin{array}{c} \text{Variación} \\ \text{explicada} \end{array}\right) + \left(\begin{array}{c} \text{Variación no} \\ \text{explicada} \end{array}\right)$
Coefficiente de determinación	$r^2 = \frac{\text{Variación explicada}}{\text{Variación total}} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$

Error estándar de estimación

$$s_e = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}}$$

Intervalo de confianza para la Y verdadera dado x

$$\left(\hat{y} - t_{\gamma, gl} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}}, \hat{y} + t_{\gamma, gl} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}} \right)$$

Estadístico de prueba para el parámetro β_1

$$t = \frac{b_1 - \beta_1}{s_e / \sqrt{\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i\right)^2}}$$

Función exponencial

$$y = b_0 e^{b_1 x}$$

Función Potencial

$$y = b_0 x^{b_1}$$

Función logarítmica

$$y = b_0 + b_1 \times \log(x)$$

Función Recíproca

$$y = b_0 + b_1 \frac{1}{x}$$

Modelo de regresión multilíneal

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$$

Ecuaciones normales para el modelo de regresión multilíneal

$$\begin{array}{rcccccc} b_0 n & + b_1 \sum_{i=1}^n x_{1,i} & + b_2 \sum_{i=1}^n x_{2,i} & + \dots & + b_k \sum_{i=1}^n x_{k,i} & = \sum_{i=1}^n y_i \\ b_0 \sum_{i=1}^n x_{1,i} & + b_1 \sum_{i=1}^n x_{1,i}^2 & + b_2 \sum_{i=1}^n x_{1,i} x_{2,i} & + \dots & + b_k \sum_{i=1}^n x_{1,i} x_{k,i} & = \sum_{i=1}^n x_{1,i} y_i \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ b_0 \sum_{i=1}^n x_{k,i} & + b_1 \sum_{i=1}^n x_{k,i} x_{1,i} & + b_2 \sum_{i=1}^n x_{k,i} x_{2,i} & + \dots & + b_k \sum_{i=1}^n x_{k,i}^2 & = \sum_{i=1}^n x_{k,i} y_i \end{array}$$

Problemas

13.1 Relacione las siguientes columnas:

- | | |
|--|-------------------------|
| a) Valores altos y bajos de x están asociados con valores altos de y ; asimismo, valores altos y bajos de x están asociados con valores bajos de y . | () Asociación positiva |
| b) Valores altos de x se asocian con valores bajos de y y valores bajos de x se asocian con valores altos de y . | () Asociación negativa |
| c) Valores altos de x se asocian con valores altos de y y valores bajos de x se asocian con valores bajos de y . | () No hay asociación |
| d) Valores altos y bajos de x están asociados con valores prácticamente iguales de y . | |

13.2 Los términos “valor alto” y “valor bajo” aplicados a las parejas de las variables aleatorias x y y en los diagramas de dispersión pueden resultar vagos. Las líneas paralelas a los ejes trazadas por el centroide permiten darle a estos términos cierta precisión. En relación a las cuatro regiones formadas, establezca lo que representa:

- Valor alto de x y valor bajo de y
- Valor alto de x y valor alto de y
- Valor bajo de x y valor bajo de y .

13.3 Describir mediante diagramas de dispersión la relación: *al aumentar x “disminuye” y* . La descripción, sin embargo, no debe corresponder a un patrón lineal sino a patrones curvos. Sugerencia: ver figura 13.2.

13.4 Discuta en equipo algunas ventajas y desventajas del coeficiente de correlación c . Exprese sus conclusiones en forma escrita.

13.5 La expresión 13.3 se obtuvo a partir de la expresión 13.2. Desarrolle en detalle:

a). El paso del numerador $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ a su equivalente $\sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i$

b). Sustituya en 13.2 la equivalencia obtenida en el inciso anterior y a s_x por

$$\sqrt{\frac{\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2}{n-1}}$$

y a s_y por $\sqrt{\frac{\sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2}{n-1}}$ para llegar a la expresión 13.3.

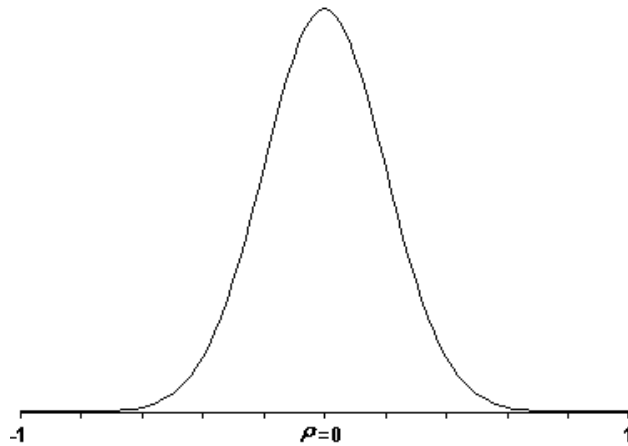
13.6 Calcule el coeficiente de correlación r para los siguientes pares de datos y establezca el tipo de asociación entre ellos.

Caballos de fuerza	74	82	86	103	123	136	161	177
Millas por galón	28	24	16	23	18	17	12	13

13.7 Los siguientes datos corresponden al rendimiento en millas por galón de gasolina y los pesos de 12 automóviles. Calcule el coeficiente de correlación r y establezca el tipo de correlación entre ellos.

Peso	2300	2100	2200	2450	2030	2700	2650	2110	3230	3210	3600	2900
MPG	28.8	29.3	34.1	27.7	33.5	26.4	23.8	30.7	18.3	19.7	14.2	20.8

- 13.8** Aplique una prueba de hipótesis con nivel de significancia igual a 0.05 para establecer si hay correlación o no en el caso del ejemplo 13.4. Concluya con el criterio tradicional y el criterio del *valor p*. Sugerencia: Utilice el programa 13.1
- 13.9** Sustituya y_i por $mx_i + b$ y \bar{y} por $m\bar{x} + b$ en la ecuación simplificada de r del ejemplo 13.1, desarrolle y obtenga $r = \frac{m}{\sqrt{m^2}}$.
- 13.10** En una asociación perfecta positiva o negativa, los puntos del diagrama de dispersión están alineados en la recta $y = mx + b$. Demuestre que el punto central (\bar{x}, \bar{y}) queda sobre dicha recta (ver ejemplo 13.1)
- 13.11** Empleando la expresión 13.3 demostrar que para el caso de una serie de puntos a lo largo de una línea recta horizontal, $r = 0/0$ (expresado comúnmente como infinito ∞ o indeterminado).
- 13.12** En la aplicación de una prueba de hipótesis al parámetro ρ (dos colas y $\alpha = 0.05$), se obtuvo *valor p* = 0.01
- Expresar la conclusión de la prueba.
 - Si se elige $\alpha = 0.01$, ¿cuál sería la conclusión de la prueba?
 - Si se elige un valor de $\alpha > 0.01$, ¿cuál sería la conclusión de la prueba?
- 13.13** Establezca la verdad o falsedad de las siguientes declaraciones.
- Si la variación de una variable no se corresponde con la variación de otra variable, entonces no existe ninguna asociación y, por tanto, ninguna correlación.
 - Una correlación fuerte entre dos variables implica necesariamente causalidad.
 - El análisis de correlación es de carácter exploratorio, de modo que si se sabe con certeza (caso frecuente en ingeniería y ciencias) que dos variables están asociadas, se procede al análisis de regresión.
 - Es altamente probable tener una correlación lineal perfecta en el caso de variables aleatorias.
 - $r = 0$ se interpreta como una falta completa de correlación.
- 13.14** La figura dada abajo muestra la distribución de la infinidad de valores de r correspondientes a todas las muestras posibles de tamaño n de una población, en la cual el coeficiente de correlación es cero.



Observe que las r se distribuyen simétricamente de -1 a +1, con la mayor parte de los valores agrupados alrededor de $\rho = 0$.

a) Suponga que $\rho = 0.90$. La distribución de los valores de r de todas las muestras posibles de tamaño n se agruparían predominantemente alrededor de _____. La distribución de las r es sesgada _____ y su gráfica es:

b) Suponga que $\rho = -0.90$. La distribución de los valores de r de todas las muestras posibles de tamaño n se agruparían predominantemente alrededor de _____. La distribución de las r es sesgada _____ y su gráfica es:

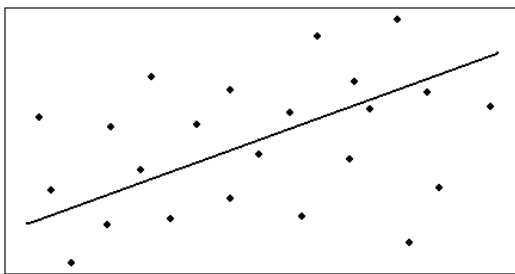
13.15 Las siguientes dos ecuaciones corresponden a líneas de regresión de dos distintos fenómenos aleatorios:

Fenómeno 1: $y_1 = 5 + 2x_1$

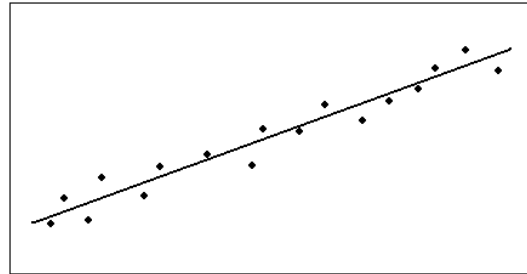
Fenómeno 2: $y_2 = -6 + 10x_2$

¿Qué fenómeno tiene el coeficiente de correlación r más grande? Argumente su respuesta.

13.16 Las dos rectas siguientes tienen las mismas pendientes y ordenadas al origen. Diga usted a quién corresponden los coeficientes $r = 0.902$ y $r = 0.1587$



a)



b)

13.17 En la ecuación general de la línea recta $y = b_0 + b_1x$, describa b_0 y b_1

13.18 Grafique los siguientes dos puntos:

Punto 1: Modelo del auto= 1999, valor en el mercado en miles de pesos = 30

Punto 2: Modelo del auto= 2004, valor en el mercado en miles de pesos = 45

a) Encuentre la ecuación de la línea recta que pasa por los puntos 1 y 2.

b) ¿Cuál es el valor de la intersección con el eje y ? ¿Tiene sentido dicho valor?

c) ¿Cuál es el valor de la pendiente?

d) Por cada año que aumente el modelo de un auto, ¿cuál es el aumento de su valor en el mercado? ¿Hay alguna relación con la pendiente?

13.19 Si todos los puntos de un diagrama de dispersión quedan exactamente sobre una línea recta, ¿sería necesaria una regresión lineal?

13.20 Discuta en equipo en qué sentido, para un conjunto de puntos, es única la recta de regresión.

13.21 La propiedad de simetría del coeficiente de correlación r de Pearson consiste en que su valor no depende de cuál de las dos variables bajo estudio se designe como x y cuál como y .

a) Es el valor de la pendiente de la recta de regresión simétrica? Justifique su respuesta.

b) ¿Es el valor de la ordenada al origen de la recta de regresión simétrica? Justifique su respuesta.

13.22 En el desarrollo del coeficiente de determinación, la elevación al cuadrado del lado derecho de

la expresión $\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n ((\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i))^2$, se simplificó a

$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$. Demuestre que el término $2 \sum_{i=1}^n (\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$ es cero.

13.23 Demostrar que el cuantificador de correlación r de Pearson no cambia su valor:

a). Si cada x_i se reemplaza por ax_i y si cada y_i se reemplaza por by_i (un cambio en la escala de medición).

b). Si cada x_i se reemplaza por $x_i - d$ y si cada y_i se reemplaza por $y_i - e$ (un cambio en la ubicación del cero de la escala de medición).

c). Los incisos anteriores resumen una de las propiedades del cuantificador r de Pearson. Discuta cuál de ellas es y emplee algún ejemplo para ilustrarlo.

13.24 En el siguiente conjunto de 7 parejas, x es una variable aleatoria que representa el promedio de bateo de un jugador de beisbol, mientras que la variable aleatoria y representa el porcentaje de ponches correspondientes del jugador.

x	0.330	0.295	0.335	0.249	0.360	0.275	0.340
y	3.1	7.5	3.8	9.0	3.4	12.0	3.2

a) Calcule r .

b) Con un nivel de significancia de 5% someta a prueba la declaración $\rho = 0$.

c) Encuentre la recta de regresión.

d) Estime el porcentaje de ponches para un jugador de beisbol cuyo promedio de bateo es 0.350

e) Encuentre un intervalo de confianza para y cuando $x = 0.350$

f) Con un nivel de significancia de 5%, someta a prueba la declaración $\beta_1 = 0$

g) Encuentre un intervalo de confianza de 90% para β_1 e interprete su significado.

13.25 En la aplicación de una prueba de hipótesis al parámetro β_1 (dos colas y $\alpha = 0.05$), se obtuvo valor $p = 0.2$

- a) Exprese la conclusión de la prueba.
- b) Analice el resultado del *valor p* en relación a la conclusión dada en el inciso anterior.

13.26 La desviación no explicada $y_i - \hat{y}_i$ es también conocida como el *residual* y se representa como e_i . Una manera de evaluar qué tan bien un modelo lineal representa los puntos muestrales es un diagrama o gráfica residual. Para construirlo, se colocan los valores de x : x_1, x_2, \dots, x_n en el eje horizontal y los correspondientes valores residuales e : e_1, e_2, \dots, e_n en el eje vertical. Los puntos resultantes: $(x_1, e_1), (x_2, e_2), \dots, (x_n, e_n)$, constituyen el diagrama residual. Debido a que la media de los residuales es siempre cero, el eje horizontal será siempre una línea central en este tipo de diagramas por lo que será conveniente resaltarla.

- a). Empleando los datos de los cigarrillos y del maratón, construya una gráfica residual para la recta de ajuste por mínimos cuadrados en ambos casos.
- b). Si el patrón de los puntos en la gráfica residual se distribuyen aleatoriamente alrededor del eje horizontal, la línea de ajuste por mínimos cuadrados proporciona un modelo razonable para los datos. ¿Es este el caso para las gráficas residuales?
- c). Si un punto en la gráfica residual parece distante del patrón que siguen los demás puntos, puede ser un punto atípico. Como se vio en la figura 13.4 tales puntos tienen una influencia importante en la recta de ajuste. ¿Hay puntos atípicos en las gráficas de arriba?

13.27 Los residuos $e_i = y_i - \hat{y}_i$ de un diagrama residual (ver problema anterior) pueden estandarizarse empleando la siguiente expresión:

$$e_i^* = \frac{e_i}{s_e \sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2}}}$$

donde s_e es el error estándar de estimación.

De esta manera los valores de e_i^* quedan en cierta forma independientes de las unidades de la variable y , pudiéndose establecer con mayor propiedad si un punto es atípico ya que, si por ejemplo, un residuo estandarizado es 2, el residuo es 2 desviaciones estándar (estimadas) más grande de lo que se esperaría de ajustar el modelo correcto.

- a) Construya un diagrama residual utilizando residuos estandarizados para el caso de los cigarrillos.
- b) Use el diagrama residual del inciso a) para discutir la validez del modelo de ajuste por mínimos cuadrados.
- c) En el diagrama residual del inciso a) trace líneas horizontales en $e^* = 2$ y $e^* = -2$. Se forma así una banda de semiancho igual a dos desviaciones estándar. Establezca si hay puntos atípicos en el diagrama residual del inciso a) utilizando como criterio el que los puntos queden fuera de la banda construida.

13.28 En una correlación lineal se manejan diferentes variables. Escriba en el paréntesis la letra que corresponda.

Variable	Descripción
a) y	() Estimación usando la línea de regresión por mínimos cuadrados
b) \hat{y}	() Variable común a dos variables que puede producir una correlación <i>aparente</i> entre ellas.
c) Y	() Variable predictiva o explicativa
d) \hat{Y}	() Error aleatorio asociado en la medición de x .
e) x	() Valores verdaderos de la variable respuesta
f) <i>Variable de confusión</i>	() Estimación de la variable respuesta usando la población
g) ε	() Variable respuesta

13.29 Empleando los resultados del ejemplo 13.13 y el programa 13.1 ensaye valores de nivel de significancia de manera tal que se acepte la hipótesis nula. Sugerencia. Ver los comentarios del ejemplo 13.6

13.30. Construir una banda de 99% de confianza para el problema de los cigarrillos. Sugerencia: Ver figura 13.18.

13.31 Se sabe que el número de pulgadas de una estructura recién construida que se hunde en el suelo está dada por

$$y = 3 - 3e^{-ax}$$

donde x es el número de meses que lleva construida la estructura. Con los valores

x	2	4	6	12	18	24
y	1.07	1.88	2.26	2.78	2.97	2.99

estime a usando una regresión de mínimos cuadrados. (Nieves H. A. y Domínguez S. F. Métodos Numéricos Aplicados a la Ingeniería. Tercera edición. Grupo Editorial Patria (2007) P. 446).

Sugerencia. Considere a $y - 3$ como una nueva variable.

13.32 Se da a continuación los valores muestrales de parejas correspondientes a la profundidad Secchi (variable predictiva x) en metros y la cantidad de *clorofila a* (variable respuesta y) correspondiente (ver ejemplo 13.4 y http://www.cdphe.state.co.us/op/wqcc/WQClassandStandards/Regs33-37/33_37RMH2008/ProponentsPHS/33_37phsNWCCO GGrandCoEx3.pdf).

x_i	y_i	x_i	y_i	x_i	y_i
1.60	20.69	2.80	7.74	3.70	4.54
1.79	18.81	3.01	6.64	3.70	2.10
1.79	17.48	3.01	4.98	3.70	1.66
1.90	18.14	3.01	3.98	3.81	5.75
1.90	12.83	3.01	3.43	3.81	1.99
2.00	11.06	3.01	2.32	3.90	5.20
2.00	10.62	3.10	4.65	3.90	2.10

2.10	14.27	3.10	3.43	4.10	4.65
2.10	7.96	3.21	4.09	4.10	1.88
2.19	4.98	3.21	2.88	4.21	1.99
2.30	9.51	3.26	2.77	4.41	1.44
2.59	4.98	3.31	4.31	4.60	1.99
2.70	6.31	3.31	3.54	4.60	1.22
2.70	4.31	3.45	1.99	4.70	3.98
2.70	3.54	3.61	3.98	4.90	1.55
2.70	2.65	3.61	2.21		

- Construir un diagrama de dispersión.
- Calcular la recta de regresión y r^2 .
- Ajuste los datos a los modelos intrínsecamente lineales siguientes: logarítmico, potencial y exponencial. Calcule además r^2 para cada uno de ellos.
- Compare los coeficientes de determinación para cada uno de los modelos empleados y determine cuál de ellos ajusta mejor los datos.

13.33 En una reacción gaseosa de expansión a volumen constante, se observa que la presión del reactor (*batch*) aumenta con el tiempo de reacción según se muestra en la tabla de abajo. ¿Qué grado de polinomio (con el criterio de ajuste exacto) aproxima mejor la función $P = f(t)$? (Nieves H. A. y Domínguez S. F. Métodos Numéricos Aplicados a la Ingeniería. Tercera edición. Grupo Editorial Patria (2007) P. 447)

P (atm)	1.0000	1.0631	1.2097	1.3875	1.7232	2.0000	2.9100
t (min)	0.0	0.1	0.3	0.5	0.8	1.0	1.3

Sugerencia. Utilice el programa 13.2.

13.34 Desarrolle el modelo potencial para los datos del ejemplo 13.14. Compare los resultados de los coeficientes de correlación y de determinación. Discuta qué modelo resultaría más adecuado para representar la información. Escriba sus conclusiones.

13.35 En la tabla

Puntos	0	1	2	3	4	5	6	7	8
v	26.43	22.40	19.08	16.32	14.04	12.12	10.51	9.15	8.00
P	14.70	17.53	20.80	24.54	28.83	33.71	39.25	45.49	52.52

v es el volumen en pie^3 de una libra de vapor y P es la presión en $psia$. Encuentre los parámetros a y b de la ecuación

$$P = av^b$$

aplicando el método de mínimos cuadrados. (Nieves H. A. y Domínguez S. F. Métodos Numéricos Aplicados a la Ingeniería. Tercera edición. Grupo Editorial Patria (2007) P. 446)

13.36 En la tabla siguiente se da la esperanza de vida(años) para las mujeres en la mayoría de los países del continente americano; así mismo se dan los valores de fecundidad (número de hijos por mujer) y el gasto público en salud (% del PIB)

País	Esperanza	Fecundidad	Gasto
Canadá	82	1.4	6.9
Estados Unidos	80	2	6.8
México	79	2.4	2.9
Antigua y Barbuda	74	2.2	3.2

Bahamas	74	2.2	3.4
Barbados	79	1.5	4.7
Belice	74	2.6	2.5
Costa Rica	80	2.1	5.8
Cuba	80	1.6	6.3
El Salvador	75	3.1	3.7
Granada	67	2.3	4
Guatemala	71	3.8	2.1
Haití	54	4.9	2.9
Honduras	70	3.4	4
Jamaica	74	2.4	2.7
Nicaragua	73	2.7	3.1
Panamá	77	2.6	5.1
Puerto Rico	82	1.7	6
República Dominicana	73	2.8	2.3
San Vicente y las Granadinas	75	1.8	3.9
Santa Lucía	77	2.1	3.4
Trinidad y Tobago	68	1.7	1.5
Argentina	80	2.16	4.3
Bolivia	68	2.8	6.4
Brasil	75	1.91	3.4
Chile	80	2	3
Colombia	76	2.4	6.4
Ecuador	79	2.6	2
Guyana	68	2	4.3
Paraguay	77	3.8	2.3
Perú	72	2.5	2.1
Surinam	71	2.4	3.6
Uruguay	79	1.8	2.7
Venezuela	77	2.2	2

Fuente: Almanaque mundial 2008 Edición 54. Editorial Televisa Internacional

- Calcule el coeficiente de correlación r y el de determinación r^2 para las columnas de esperanza de vida y fecundidad.
- Calcule la recta de regresión para la esperanza de vida (variable respuesta) y fecundidad (variable predictiva).
- Proponga un modelo multilíneal y estime los parámetros empleando los datos de la tabla.
- Para los modelos encontrados en los incisos *b*) y *c*) calcule $\sum_{i=1}^n (\hat{y}_i - y_i)^2$ y compare. ¿Son significativas las diferencias? (ver proyecto 13.3)

13.37 Empleando la tabla 13.1 de los cigarrillos.

- Ordenar los datos respecto a la nicotina. Recorrer simultáneamente las columnas de nicotina y CO y describir el comportamiento de los pares de datos.
- Construir un diagrama de dispersión. ¿Se confirma el patrón descrito en el inciso anterior? Trazar sobre el diagrama las líneas de división que se cruzan en el centroide. Describe el tipo de asociación entre las variables.
- Calcule el coeficiente de correlación de Pearson y establece el grado de asociación entre las variables.
- Ensaye una relación multilíneal entre el CO, el alquitrán y la nicotina. Calcule los parámetros

Proyectos abiertos

Proyecto 13.1 Construcción de intervalos de confianza para el coeficiente de correlación poblacional ρ .

Cuando se obtienen todas las muestras posibles de n pares de datos de una población con coeficiente de correlación ρ , la distribución de los coeficientes muestrales r no es normal. Para pasar de dicha distribución a una distribución que es aproximadamente normal, R. A. Fisher ideó la siguiente conversión $z = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right)$. La media de la distribución resultante es $\frac{1}{2} \ln\left(\frac{1+\rho}{1-\rho}\right)$ y su desviación estándar $\sqrt{\frac{1}{n-3}}$. Esta conversión es conocida como transformación de Fisher.

Construya un intervalo de confianza del 95% de confianza para el caso de los tiempos de los ganadores del Maratón de Nueva York.

Sugerencia: Use el siguiente procedimiento:

Paso 1. Empleando la tabla de distribución normal calcular $z_{\alpha/2}$.

Paso 2. Calcule los límites del intervalo de confianza de acuerdo a:

$$LCI = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right) - z_{\alpha/2} \frac{1}{\sqrt{n-3}} \qquad LCS = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right) + z_{\alpha/2} \frac{1}{\sqrt{n-3}}$$

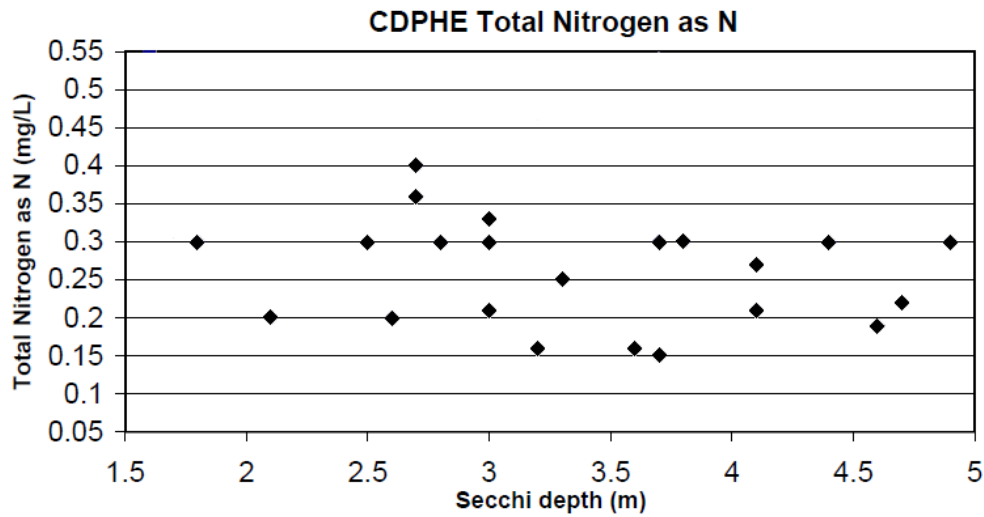
Paso 3. Los límites del paso 2 son los límites transformados. Para obtener los límites de ρ se debe reconvertir z a r . Para ello, sustituya los valores obtenidos de LCI y LCS por z en $z = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right)$ y despeje a r . Los valores resultantes son los correspondientes límites que forman el intervalo de confianza.

Finalmente, puede corroborar sus cálculos con el programa 13.1 en el menú **Intervalos de confianza**.

Proyecto 13.2 Lectura de diagramas de dispersión.

Dado el diagrama de dispersión siguiente (profundidades Secchi vs Nitrógeno total), obtener los valores numéricos aproximados correspondientes a las coordenadas de los puntos y calcular el coeficiente de correlación r de Pearson y la recta de ajuste por mínimos cuadrados. Probar estadísticamente el coeficiente de correlación.

Sugerencia. Para establecer las coordenadas con mayor exactitud, lleve el diagrama a un programa de edición de gráficos (Paint, por ejemplo).



Proyecto 13.3 Determinación de cuándo agregar o eliminar variables en un modelo multilíneal.

Como se dijo en la sección 13.4, es posible mejorar las predicciones aumentando variables en el modelo. Así, en el caso del Maratón de Nueva York, se propuso considerar además de la temperatura media la velocidad del viento y la humedad relativa. Suponga que se plantea un modelo multilíneal en el que se considere la temperatura media y la velocidad v del viento, quedando:

$$\hat{t} = b_0 + b_1T + b_2v$$

Una pregunta estadística válida es si la inclusión de la velocidad v del viento mejora significativamente la estimación \hat{t} . La respuesta puede encontrarse mediante una prueba de hipótesis utilizando la distribución F . Si, por ejemplo, se rechaza la hipótesis nula (la inclusión de la velocidad del viento no mejora la estimación \hat{t}), se tiene un nuevo modelo con el cual se puede continuar a considerar otras variables. Para ver cómo se lleva a cabo esto se emplea el caso de las calorías en los alimentos mediante los pasos siguientes.

Paso 1. Utilizando la tabla 13.10 lleve a cabo la regresión considerando solamente las grasas y las proteínas (puede usar el programa 13.3). Con los valores obtenidos para los parámetros del modelo usado, calcule la variación no explicada:

$$VNE_2 = \sum_{i=1}^n (\hat{y}_i - y_i)^2$$

Paso 2. Con el modelo de regresión que utiliza las grasas, las proteínas y los carbohidratos calcule la variación no explicada correspondiente:

$$VNE_3 = \sum_{i=1}^n (\hat{y}_i - y_i)^2$$

Pregunta de investigación: ¿Es la diferencia $VNE_2 - VNE_3$ estadísticamente significativa?

Paso 4. Calcule $\frac{VNE_2 - VNE_3}{\text{Número de variables aumentadas}}$

Calcule $\frac{VNE_3}{n-p-1}$, con n igual al tamaño de la muestra y p igual al número de variables para el cálculo de VNE_3 .

Calcule el valor observado de F mediante:

$$F_o = \frac{\frac{VNE_2 - VNE_3}{\text{Número de variables aumentadas}}}{\frac{VNE_3}{n-p-1}}$$

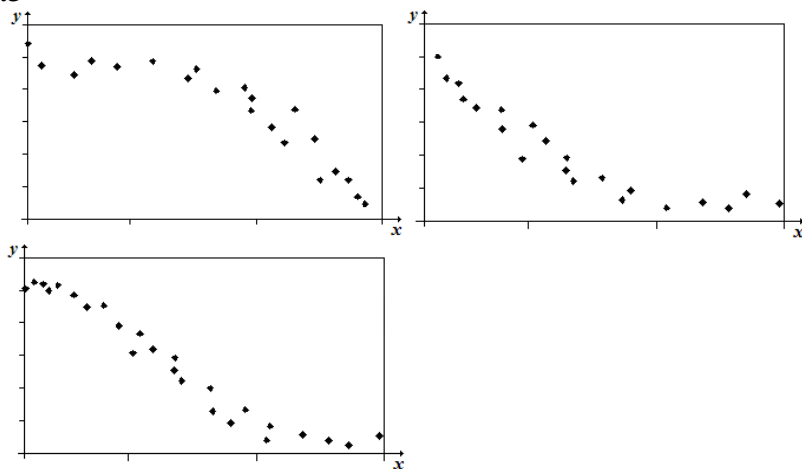
Consulte en una tabla de distribución F , con $\alpha = 0.05$ y con grados de libertad del numerador igual al número de variables aumentadas al modelo y con grados de libertad del denominador igual a $n-p-1$.

Compare los valores de F_o y el F de tablas y decida si la inclusión de la variable carbohidratos es significativa o no.

Solución a problemas impares

13.1 c; b; a y d

13.3



13.5 a) Desarrollando y agrupando:

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^n (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x} \bar{y}) = \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i - \bar{x} \sum_{i=1}^n y_i + n \bar{x} \bar{y} \\ &= \sum_{i=1}^n x_i y_i - n \bar{y} \bar{x} - n \bar{x} \bar{y} + n \bar{x} \bar{y} = \sum_{i=1}^n x_i y_i - n \bar{y} \bar{x} = \sum_{i=1}^n x_i y_i - n \sum_{i=1}^n \frac{y_i}{n} \sum_{i=1}^n \frac{x_i}{n} \\ &= \sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i \end{aligned}$$

b)

$$\begin{aligned} r &= \frac{1}{n-1} \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{\frac{\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i\right)^2}{n-1}} \sqrt{\frac{\sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i\right)^2}{n-1}}} = \frac{1}{n-1} \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\frac{1}{n-1} \sqrt{\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i\right)^2} \sqrt{\sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i\right)^2}} \\ &= \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i\right)^2} \sqrt{\sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i\right)^2}} = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sqrt{\left(\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i\right)^2\right) \left(\sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i\right)^2\right)}} \\ &= \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{n \left(\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i\right)^2\right) n \left(\sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i\right)^2\right)}} = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \sqrt{n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i\right)^2}} \end{aligned}$$

13.7 Es una asociación lineal negativa fuerte con $r = -0.966$

13.9 Sustituyendo lo indicado resulta:

$$\begin{aligned}
 r &= \frac{\sum_{i=1}^n (x_i - \bar{x})(mx_i + b - m\bar{x} - b)}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (mx_i + b - m\bar{x} - b)^2}} = \frac{m \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{m^2 \sum_{i=1}^n (x_i - \bar{x})^2}} \\
 &= \frac{m \sum_{i=1}^n (x_i - \bar{x})^2}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{m^2}} = \frac{m}{\sqrt{m^2}}
 \end{aligned}$$

13.11 En el caso de una recta horizontal la ecuación está dada por $y = b$, donde b es la intersección con el eje y o ordenada al origen. Sustituyendo y_i por b en la ecuación 13.3:

$$r = \frac{\sum_{i=1}^n x_i b - \sum_{i=1}^n x_i \sum_{i=1}^n b}{\sqrt{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \sqrt{n \sum_{i=1}^n b^2 - \left(\sum_{i=1}^n b\right)^2}} = \frac{nb \sum_{i=1}^n x_i - nb \sum_{i=1}^n x_i}{\sqrt{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \sqrt{n^2 b^2 - (nb)^2}} = \frac{0}{0}$$

13.13 a) V; b) F; c) V; d) F; e) F (Ya que en una asociación curva r puede ser 0, sin embargo, hay asociación).

13.15 La pendiente de una recta expresa el ángulo de inclinación de la misma y por ende está vinculado al coeficiente de correlación r . Dado que las dos pendiente son positivas, puede razonarse de la siguiente manera: Al tender a cero la pendiente, la recta tiende a ser horizontal; si en cambio, la pendiente tiende a infinito, la recta tiende a ser vertical. De acuerdo con esto, al fenómeno 2 le correspondería un coeficiente de correlación más alto que al fenómeno 1.

13.17 b_0 es la intersección de la recta con el eje x (ordenada al origen) y b_1 es la pendiente de la recta (cambio de y por unidad de x)

13.19 No, ya que la ecuación de la recta que pasa *por* ellos puede obtenerse de manera más sencilla usando sólo dos puntos.

13.21 a) No es simétrica. En general, si se intercambian los valores de x y y en la ecuación 13.6, se obtendrán diferentes resultados.

b) No es simétrica. En general si se intercambian los valores de x y y en la ecuación 13.5, se obtendrán diferentes resultados.

13.23 a) Sustituyendo x_i por ax_i y y_i por by_i en la expresión 13.3

$$r = \frac{n \sum_{i=1}^n ax_i by_i - \sum_{i=1}^n ax_i \sum_{i=1}^n by_i}{\sqrt{n \sum_{i=1}^n (ax_i)^2 - \left(\sum_{i=1}^n ax_i\right)^2} \sqrt{n \sum_{i=1}^n (by_i)^2 - \left(\sum_{i=1}^n by_i\right)^2}} = \frac{ab \left(n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i \right)}{a \sqrt{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} b \sqrt{n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i\right)^2}}$$

$$= \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \sqrt{n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i\right)^2}}$$

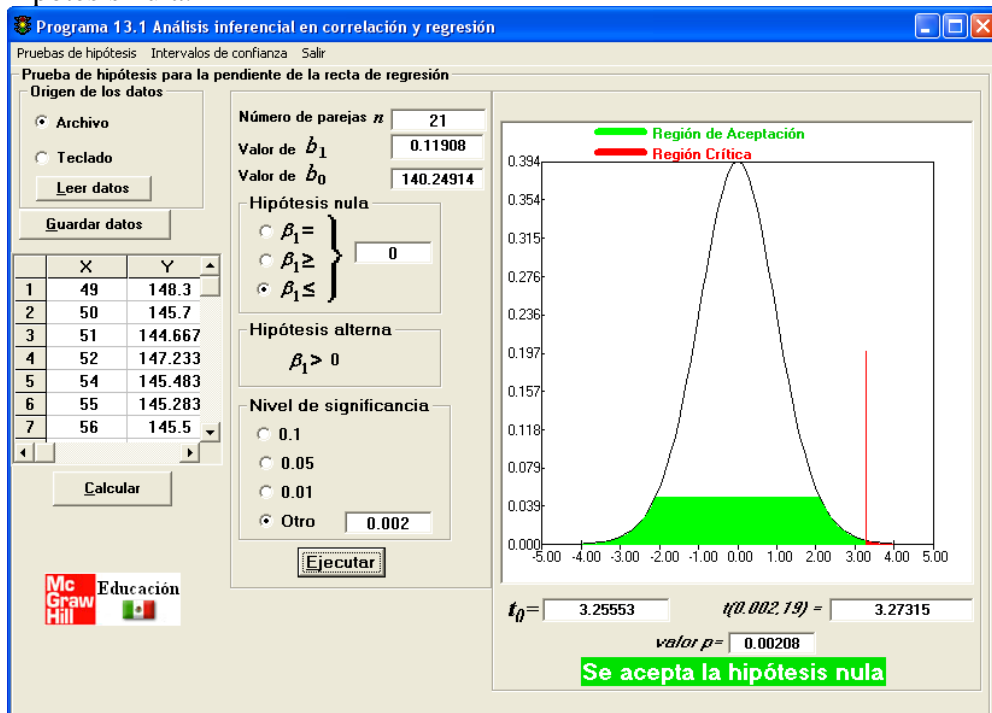
b) Siga las ideas mostradas en el inciso anterior.

c) Se trata de la propiedad de la independencia de r de las unidades en que se midan x y y . Por ejemplo, el cálculo del coeficiente de correlación r para el caso del maratón da el mismo valor si se emplean grados Celsius o grados Fahrenheit.

13.25 a) Se acepta H_0

b) Se puede concluir, con cualquiera de los valores predeterminados de α (0.1, 0.05, 0.01), que no hay correlación entre las variables en estudio. Lo anterior le daría al investigador un margen amplio de seguridad en su conclusión pero no certidumbre.

13.29 Al ejecutar el programa 13.1 pero utilizando para el nivel de significancia valores menores de 0.00208 (por ejemplo 0.002) se obtiene como resultado la aceptación de la hipótesis nula.



13.31 $y = 3 - 3e^{-0.2386x}$

13.33 Grado = 3

13.35 $P = 481.03743v^{-1.06533}$

13.37 a) Al aumentar x “aumenta” y por lo que se trata de una correlación lineal positiva

b) Sí se confirma.

c) $r = 0.9356$. Se trata de una correlación lineal positiva fuerte.

d) $\hat{y} = 3.08961 + 0.96247x_1 - 2.64627x_2$

Deseamos expresar nuestro agradecimiento a la M. C. Leticia Cañedo Suárez de la Escuela Superior de Cómputo del Instituto Politécnico Nacional por sus valiosos comentarios, sugerencias y motivación.

De igual forma agradecemos al Ing. César Gustavo Gómez Sierra su apoyo técnico para finalizar exitosamente este libro.