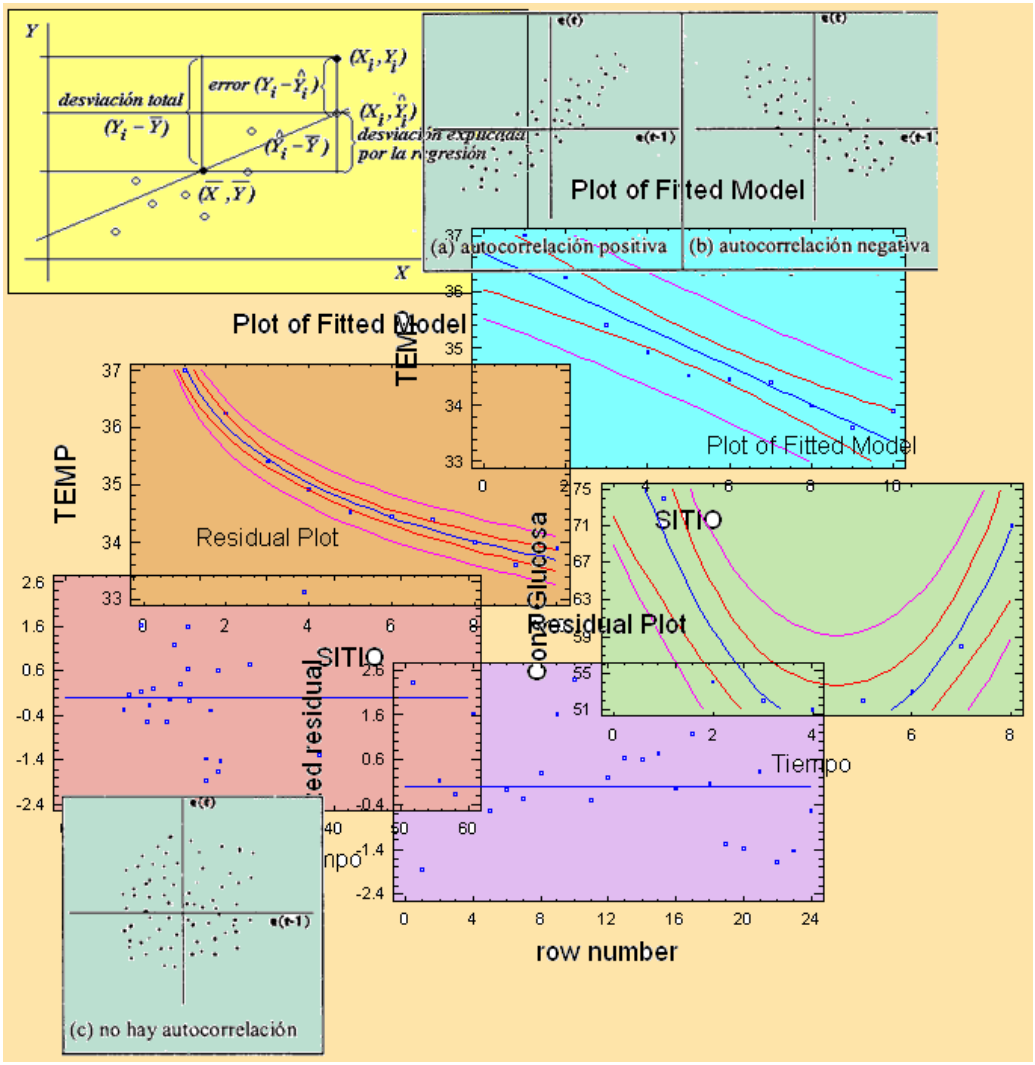
UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
FACULTAD DE ESTUDIOS SUPERIORES ZARAGOZA

DIVISIÓN DE CIENCIAS QUÍMICO BIOLÓGICAS
ACADEMIA DE ESTADÍSTICA Y CÓMPUTO CIENTÍFICO
Serie: Comunicaciones en Estadística y Cómputo Científico



ANÁLISIS DE REGRESIÓN UN ENFOQUE PRÁCTICO

MARÍA JOSÉ MARQUES DOS SANTOS
MARÍA DEL CARMEN GALINDO DE SANTIAGO
ARMANDO CERVANTES SANDOVAL

2007

Análisis de Regresión
Un Enfoque Práctico

María José Marques Dos Santos

Ma. del Carmen Galindo de Santiago

Armando Cervantes Sandoval

Publicado con apoyo del proyecto PAPIME PE100606

FACULTAD DE ESTUDIOS SUPERIORES ZARAGOZA

Av. Guelatao N° 66, Colonia Ejército de Oriente,

Delegación Iztapalapa, México, D.F. 09230

ISBN: 978-970-32-4722-6

PRÓLOGO

El presente material se elaboró con la finalidad de dar a conocer las herramientas básicas de la Regresión, no como un libro especializado en la materia, de los cuales existen muchos, sino de mostrar la forma de resolver problemas prácticos con ayudar de un software estadístico, sin perder de vista el rigor de los diferentes modelos. Se pretende que este material no sólo sirva a los estudiantes del DIPLOMADO EN ESTADÍSTICA EN LÍNEA de la FES ZARAGOZA, UNAM, sino a todos aquellos que necesitan de esta herramienta para analizar sus datos derivados de su investigación.

El material consta principalmente de las técnicas de REGRESIÓN LINEAL SIMPLE, CURVILÍNEA, MÚLTIPLE y POLINÓMICA con especial énfasis en el DIAGNÓSTICO de la regresión por medio del ANÁLISIS DE RESIDUOS, AUTOCORRELACIÓN y MULTICOLINEALIDAD; se apoya con el uso del software de análisis estadístico STATGRAPHICS para enfatizar en el análisis e interpretación de los problemas y sus resultados más que en el cálculo.

Consta de tres capítulos, uno de Análisis de Regresión y Correlación Lineal simple, otro de Análisis de Regresión Curvilínea y otro de Regresión Lineal Múltiple y Polinómica. No se incluyen ejercicios porque éstos se propondrán como actividades a realizar en el curso en línea.

Los autores agradecen cualquier comentario, corrección o sugerencia que enriquezca este material.

María José Marques Dos Santos
marques@servidor.unam.mx

Ma. del Carmen Galindo de Santiago
cgalind@servidor.unam.mx

Armando Cervantes Sandoval
arpacer@servidor.unam.mx

TABLA DE CONTENIDOS

CAPÍTULO 1

ANÁLISIS DE REGRESIÓN LINEAL SIMPLE

1.1 ANÁLISIS DE REGRESIÓN Y CORRELACIÓN

A menudo en investigación se está interesado en estudiar **la relación entre** dos **variables** como cantidad de fertilizante y producción, concentración de un fármaco inyectado a un animal de laboratorio y latidos del corazón, dureza de los plásticos tratados con calor durante diferentes períodos. La naturaleza y grado de relación entre este tipo de variable se puede analizar mediante dos técnicas: *regresión* y c*orrelación,* que aunque de alguna manera, están relacionadas tienen propósitos e interpretaciones diferentes.

La diferencia entre ambos procedimientos no permite que se sustituya uno por el otro en una situación experimental dada.

El análisis de regresión es útil para **determinar la forma probable de la relación entre las variables** (la ecuación que relaciona a ambas variables) cuando hay un fenómeno de *causa y efecto*; y su objetivo principal es el de predecir o estimar el valor de una variable (*respuesta o dependiente* (*Y*)), correspondiente al valor dado de la otra variable (*explicativa o independiente* (*X*)). En otras palabras, el investigador decide cuáles valores tomará la variable independiente, mientras que los valores de la variable dependiente están determinados por la relación que existe, si la hay, entre la variable dependiente y la independiente. Por lo tanto, debe emplearse el análisis de regresión en situaciones experimentales en las cuales el investigador controla la variable independiente.

El análisis de correlación, por otra parte, consiste en la **medición del grado o intensidad de *asociación* entre dos variables sin importar cuál es la causa y cuál el efecto**. Cuando se puede demostrar que la variación de una variable está de algún modo asociada con la variación de otra, entonces se puede decir que las variables están *correlacionadas*. Una correlación puede ser *positiva* (cuando al aumentar una variable la otra también aumenta), o *negativa* (cuando al aumentar una variable la otra disminuye). Por otro lado, si la variación de una variable no corresponde en absoluto a la variación de la otra, entonces no existe ninguna asociación y, por consiguiente, *ninguna correlación* entre las dos variables.

1.2 DIAGRAMAS DE DISPERSIÓN

El primer paso a realizar en el estudio de la relación entre dos variables es el diagrama de dispersión que consiste en representar los pares de valores (X_i, Y_i) como puntos en un sistema de ejes cartesianos *X Y*. Debido a la variación del muestreo los puntos estarán dispersos.

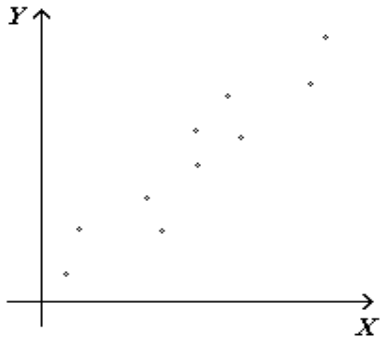
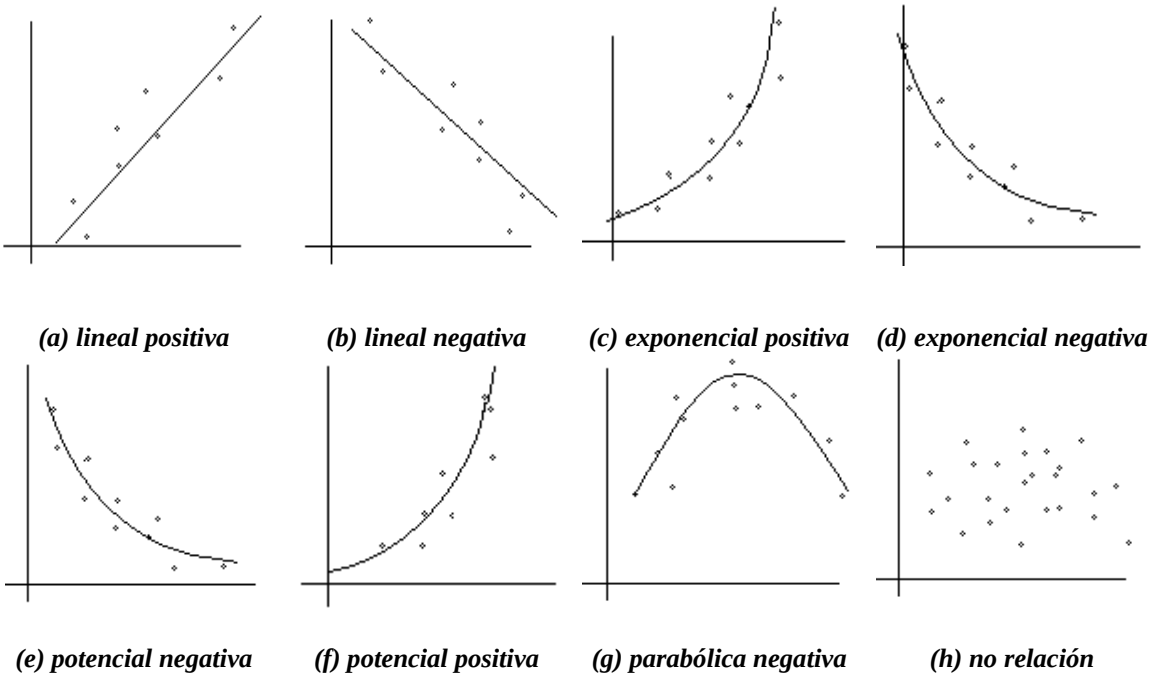


Fig. 1 Diagrama de dispersión

Si los puntos muestran una tendencia lineal positiva o negativa se puede ajustar una línea recta que servirá entre otras cosas para predecir valores de *Y* correspondientes a valores de *X*. Según como se presenten los puntos en el diagrama se puede, intuitivamente, imaginar el tipo de relación que hay entre las dos variables; sin embargo, esto no será definitivo, sólo permitirá establecer si la relación es lineal o no lineal, y si es positiva o negativa.

Después de dibujar los puntos un examen del diagrama puede revelar que estos siguen un patrón, e indicar el modelo matemático a utilizarse en el análisis. Otra forma de llegar al modelo puede ser por consideraciones teóricas o porque se sabe por experiencia o por referencia cómo se comportan las variables.

El análisis de regresión puede ser lineal, no lineal (curvilíneo), lineal simple o lineal múltiple; el lineal simple se ocupa sólo de dos variables y el múltiple de tres o más variables.



(a) lineal positiva ***(b) lineal negativa*** ***(c) exponencial positiva*** ***(d) exponencial negativa***

(e) potencial negativa ***(f) potencial positiva*** ***(g) parabólica negativa*** ***(h) no relación***

Fig. 2 Tipos de relación entre dos variables

1.3 REGRESIÓN LINEAL SIMPLE

Después de determinar el modelo matemático a utilizar y establecer que la relación es lineal se procede a ajustar una recta llamada Recta de *Regresión* o Recta *de Ajuste*, lo cual se puede realizar por varios métodos. Uno de los más potentes es el método de *mínimos cuadrados*, pues si se cumplen sus suposiciones se pueden realizar inferencias acerca de los parámetros. Otro método muy eficaz, en caso de que hayan casos extraordinarios, es el método de regresión resistente (véase Velleman y Hoaglin, 1981).

1.4 SUPUESTOS PARA LA REGRESIÓN LINEAL SIMPLE

1. Los valores de la variable independiente *X* son fijos, a *X* se le llama variable no aleatoria.

2. Para cada valor de *X* hay una subpoblación de valores de *Y,* y cada subpoblación de valores de *Y* debe estar normalmente distribuida (*Normalidad*).

3. Las varianzas de las subpoblaciones de *Y* deben ser iguales (*Homoscedásticidad*).

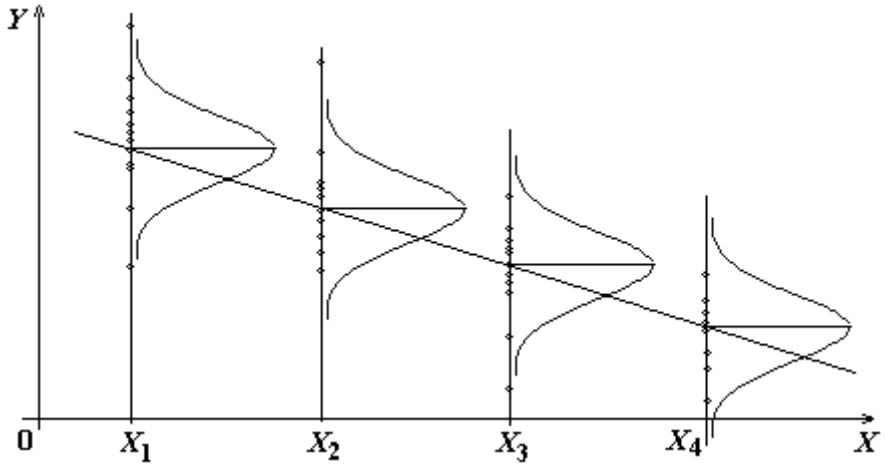


Fig. 3 Para cada valor de X hay una subpoblación de valores de Y, normalmente distribuida con varianzas iguales (normalidad y homoscedásticidad)

4. Las medias de las subpoblaciones de *Y* todas están sobre una recta. (*Linealidad*).

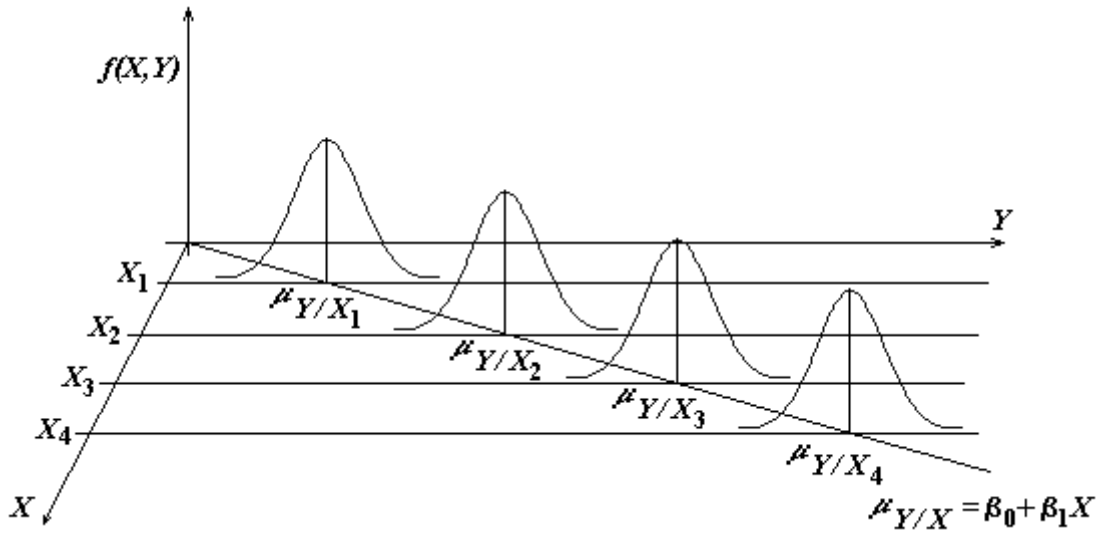


Fig. 4 Las medias de las subpoblaciones de Y están todas sobre una recta (Linealidad)

5. Los valores de *Y* son estadísticamente independientes; es decir, los valores de *Y* correspondientes a un valor de *X* no dependen de los valores de *Y* para otro valor de *X*. (*Independencia*).

Bajo estos supuestos la relación que se quiere estimar es:

$$\mu_{Y/X} = \beta_0 + \beta_1 X \qquad [1]$$

Esto significa que el valor medio de *Y* para un valor fijo de *X* es igual a $\beta_0 + \beta_1 X$. Las constantes β_0 y β_1 son la ordenada al origen y la pendiente, respectivamente.

El problema consiste en utilizar la información en una muestra de tamaño *n* para estimar los valores de los parámetros β_0 y β_1.

Una *Y* seleccionada aleatoriamente se representa de la siguiente manera:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad [2]$$

Donde ε_i es el error aleatorio. Ya que la recta de ajuste es una recta *probabilística* no *determinística*, (en Álgebra y Geometría Analítica se utiliza la recta determinística mientras que en estadística siempre se presenta un *error de ajuste* o *de estimación*).

Despejando ε_i de la ecuación [2], se tiene:

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i) = Y_i - \mu_{Y/X}$$

Donde se puede ver que ε_i es la desviación de cada valor de Y_i observado con respecto a la media de la población de valores de Y.

1.5 MÍNIMOS CUADRADOS

La ecuación de regresión de la población [1] se estima con la ecuación:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad \text{o} \quad \hat{y} = b_0 + b_1 x$$

donde: $\hat{y}$, $\hat{\beta}_0 = b_0$ y $\hat{\beta}_1 = b_1$ son estimadores de $\mu_{Y/X}$, β_0 y β_1, respectivamente. Para obtener los estimadores b_0 y b_1 se puede utilizar el método de *mínimos cuadrados.*

El método de *mínimos cuadrados* consiste en ajustar la recta que cumpla con la condición de que la suma de los cuadrados de las desviaciones de cada valor observado Y de su correspondiente valor de predicción $\hat{Y}$, sea mínima. En otras palabras, la suma:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \text{, debe ser mínima.}$$

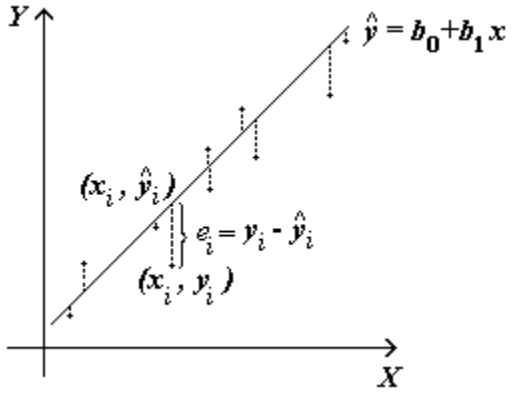


Fig. 5 Desviaciones de los valores observados con respecto a los valores de predicción

Derivando parcialmente la suma de cuadrados, con respecto a b_0 y b_1, e igualando a cero para minimizarla (cero es el valor mínimo que puede tomar cualquier cantidad elevada al cuadrado), se tiene:

$$\frac{\partial \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2}{\partial b_0} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$

$$\frac{\partial \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2}{\partial b_1} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)x_i = 0$$

Lo que da origen a las *ecuaciones normales* para la recta de mínimos cuadrados

$$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum x_i \qquad [3]$$

$$\sum_{i=1}^{n} x_i y_i = b_0 \sum x_i + b_1 \sum x_i^2 \qquad [4]$$

Despejando b_0 de [3] se tiene

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum x_i}{n} = \bar{y} - b_1 \bar{x} \qquad [5]$$

Sustituyéndola en [4], se tiene b_1.

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \sum x_i \sum y_i}{\sum x_i^2 - n} = \frac{\sum x_i y_i - n\bar{x}\,\bar{y}}{\sum x_i^2 - n\bar{x}^2} \qquad [6]$$

Al dividir el numerador y denominador de [6] entre $(n-1)$ se puede escribir b_1 en términos del estimador de la varianza, s_x^2 y de $\frac{\sum x_i y_i - n\bar{x}\,\bar{y}}{n-1}$, que es el estimador de la varianza "combinada" de las dos variables *X* y *Y*, denominada covarianza. Es decir,

$$s_{xy} = Cov(x,y) = \frac{\sum x_i y_i - n\bar{x}\,\bar{y}}{n-1} \quad \text{y} \quad b_1 = \frac{s_{xy}}{s_x^2} \qquad [7]$$

En términos de desviaciones de las medias, las sumas de cuadrados y de los productos cruzados pueden definirse como sigue:

$$m_{11} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - n\,\bar{y}^2$$
$$m_{22} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - n\,\bar{x}^2$$
$$m_{12} = \sum (y_i - \bar{y})(x_i - \bar{x}) = \sum y_i x_i - n\,\bar{y}\,\bar{x}$$

Con estas definiciones, la fórmula [6] para calcular b_1, se puede escribir en forma compacta como:

$$b_1 = \frac{m_{12}}{m_{22}}$$

Al obtener los valores de b_0 y b_1 por las ecuaciones [5] y [6], ya se pueden sustituir en la ecuación de la recta: $\hat{y} = b_0 + b_1 x$. Para representar esta ecuación sobre un diagrama de dispersión basta con tomar dos puntos, seleccionando valores arbitrarios de *x*, que se sustituyen en la ecuación para obtener los correspondientes valores de $\hat{y}$.

Ejemplo 1: Se realizó un experimento para estudiar el efecto de cierto fármaco en la disminución del ritmo cardíaco en adultos. La variable independiente es la dosis del fármaco en miligramos, y la variable dependiente es la diferencia entre el ritmo más bajo registrado después de la administración del fármaco y el ritmo antes (control) de la administración del mismo.

Se puede observar que la variable *x* cumple con la suposición de valores fijos. Los valores de *x,* se fijaron de antemano y no se les permitió variar aleatoriamente.

Dosis (*mg*)	Reducción del ritmo cardíaco (*latidos/min*)
0.50	10
0.75	8
1.00	12
1.25	12
1.50	14
1.75	12
2.00	16
2.25	18
2.50	17
2.75	20
3.00	18
3.25	20
3.50	21

Como se mencionó anteriormente, el primer paso en un análisis de regresión es representar los puntos en un diagrama de dispersión.

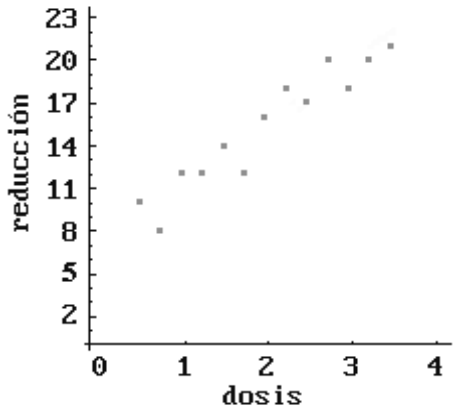


Fig. 6 Diagrama de dispersión para los datos del ejemplo 1

Como se puede observar los puntos siguen una relación lineal positiva, por lo tanto, se procede a determinar la recta de ajuste.

x	y	x^2	xy
0.50	10	0.2500	5.0
0.75	8	0.5625	6.0
1.00	12	1.0000	12.0
1.25	12	1.5625	15.0
1.50	14	2.2500	21.0
1.75	12	3.0625	21.0
2.00	16	4.0000	32.0
2.25	18	5.0625	40.5
2.50	17	6.2500	42.5
2.75	20	7.5625	55.0
3.00	18	9.0000	54.0
3.25	20	10.5625	65.0
3.50	21	12.2500	73.5
$\bar{x} = 2.00$	$\bar{y} = 15.2308$	$\sum x^2 = 63.37$	$\sum xy = 6442.50$
$s_x = 0.9736$	$s_y = 4.1864$		
$s_x^2 = 0.9479$	$s_y^2 = 17.5256$		

$$b_1 = \frac{\sum x_i y_i - n\bar{x}\,\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{6442.5 - (13)(2)(15.2308)}{63.3750 - (13)(2)^2} = 4.0879$$

$$b_0 = \bar{y} - b_1\bar{x} = 15.2308 - 4.0879(2) = 7.0549$$

La ecuación de la recta de ajuste es: $\hat{y} = 7.0549 + 4.0879x$

Para representarla, se sustituyen **dos valores cualesquiera** de x en la ecuación de la recta obtenida, por ejemplo, con los valores 1.00 y 3.00 se obtienen los puntos: (1.00, 11.14) y (3.00, 19.32).

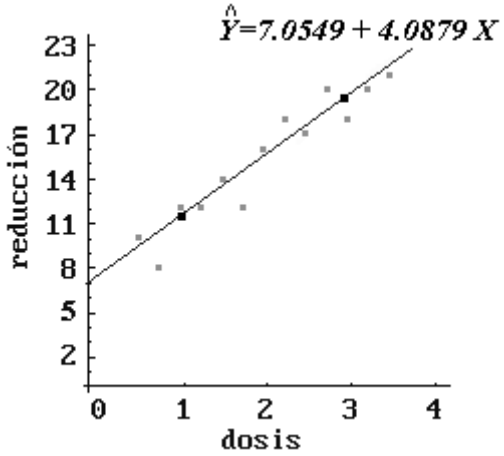


Fig. 7 Recta de regresión por mínimos cuadrados del ejemplo 1

Aunque la recta obtenida se desvía muchos, de los puntos observados, se tiene la seguridad de que para cualquier otra recta que se trace, la suma de los cuadrados de las desviaciones verticales a ella será mayor que la suma de cuadrados de las desviaciones a la recta de mínimos cuadrados. Otro método para encontrar los valores que mejor estimen a los parámetros β_0 y β_1 es a través de matrices, como se indica a continuación:

Si la ecuación de una línea recta es:

$$Y = f(X) = \beta_0 + \beta_1 X$$

Donde:
β_0 ; ordenada al origen
β_1 : pendiente de la recta

El modelo de regresión lineal es:

$Y_i = \mu_{Y/X} + \varepsilon_i = \beta_0 + \beta_1 X + \varepsilon_i$, con $i = 1,2,3,\cdots,n$

Entonces, para cada observación se tiene:

$$Y_1 = \beta_0 + \beta_1 X_1 + \varepsilon_1$$
$$Y_2 = \beta_0 + \beta_1 X_2 + \varepsilon_2$$
$$\vdots$$
$$Y_n = \beta_0 + \beta_1 X_n + \varepsilon_n$$

El modelo se puede escribir como:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ & \vdots \\ 1 & X_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} = \mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$$

Donde:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ & \vdots \\ 1 & X_n \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Realizando operaciones con matrices con el fin de despejar la matriz $\boldsymbol{\beta}$, para las estimaciones de la muestra, se tiene

$$\begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$$

$$\mathbf{X^t X\beta = X^t Y}$$

De donde la solución matricial para calcular los parámetros de la ecuación de regresión es:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X^t X})^{-1}(\mathbf{X^t Y}) \qquad [8]$$

1.6 ANÁLISIS DE VARIANZA EN REGRESIÓN LINEAL

Geométricamente se tiene

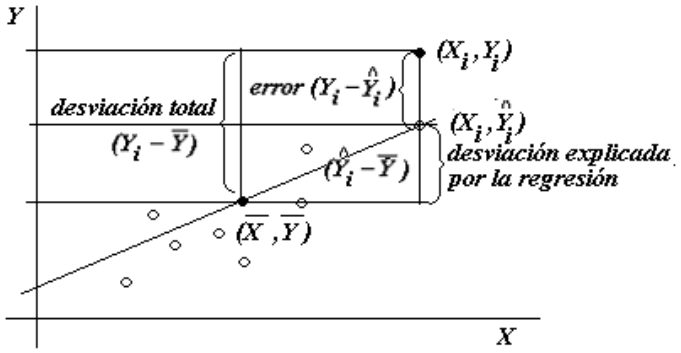


Fig. 8 Desviaciones en el modelo de regresión: 1) Total, 2) Explicada por la Regresión y 3) Error

1. $y_i - \bar{y}$ desviación total
2. $\hat{y}_i - \bar{y}$ desviación explicada por la regresión
3. $y_i - \hat{y}_i$ error

$$\underbrace{y_i - \bar{y}}_{\text{Total}} = \underbrace{(\hat{y}_i - \bar{y})}_{\text{Regresión}} + \underbrace{(y_i - \hat{y}_i)}_{\text{Error}}$$

Al elevar al cuadrado y aplicar sumatorias se tiene:

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

Note que el término $2\sum(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = 2\sum(\hat{y}_i - \bar{y})e_i$ es cero porque $\sum(\hat{y}_i - \bar{y}) = 0$ por ser el primer momento (Ver Marques 2004).

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{SC}_{\text{Total}}}=\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\text{SC}_{\text{Regresión}}}+\underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{\text{SC}_{\text{Error}}} \qquad [9]$$

Cantidades que permiten realizar un ANOVA, para contrastar las hipótesis:

$$Ho: \beta_i = 0$$
$$Ha: \beta_i \neq 0$$

Fuente de variación	**g.l.**	**SC**	**CM**	$\mathbf{F_C}$	$\mathbf{F_t}$
Regresión	1	SC_{Reg}	CM_{Reg}	$\frac{CM_{\text{Reg}}}{CM_{Error}}$	$F_{1-\alpha,1,n-2}$
Error	$n-2$	SC_{Error}	CM_{Error}		
Total	$n-1$	SC_{Total}			

Este ANOVA tiene el siguiente par de hipótesis:

Ho: $\beta_i = 0$, es decir que todos los coeficientes del modelo son iguales a cero y por lo tanto **no hay un modelo lineal** que describa el comportamiento de los datos.

Ha: $\beta_i \neq 0$, de que al menos uno de los coeficientes es diferente de cero y entonces **si hay un modelo lineal.**

INTERPRETANDO a β_0 *y* β_1

Ho: $\beta_1 = 0$

Caso 1: **No se rechaza** *Ho*: $\beta_1 = 0$; **es decir, que la pendiente es cero o que no hay pendiente,** entonces se tienen dos opciones de interpretación.

a) Si la suposición de línea recta es correcta significa que *X* no proporciona ayuda para predecir *Y*, esto quiere decir que $\overline{Y}$ predice a *Y*.

$$Ho: \beta_1 = 0$$

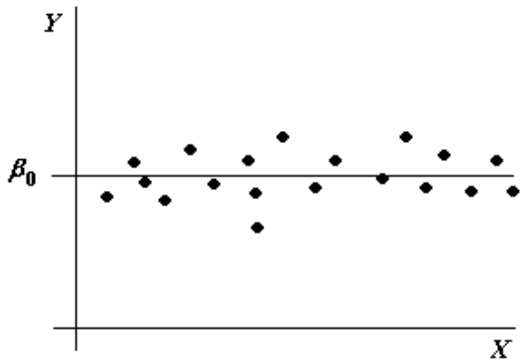


Fig. 9 X no proporciona ayuda para predecir Y

b) La verdadera relación entre *X* y *Y* no es lineal, esto significa que el modelo puede involucrar funciones cuadráticas, cúbicas o funciones más complejas.

Fig. 10 La relación ente X y Y es no lineal

NOTA: Si hay una curvatura se requiere un elemento cuadrático en el modelo, si hay dos curvaturas entonces se requiere un cúbico y así sucesivamente.

Caso 2: Se rechaza *Ho*: $\beta_1 = 0$, **es decir, sí hay pendiente o en otras palabras si hay un modelo lineal que describe el comportamiento de los datos.**

a) *X* proporciona información significativa para predecir *Y*

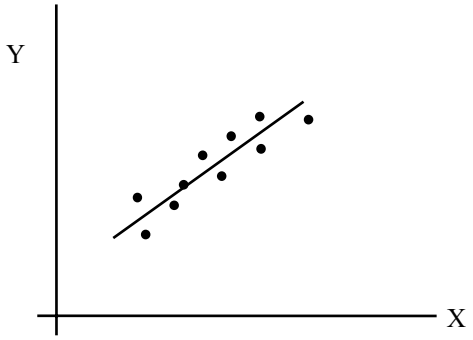


Fig. 11 La relación ente X y Y es totalmente lineal

b). El modelo puede tener un término lineal más, quizás un término cuadrático.

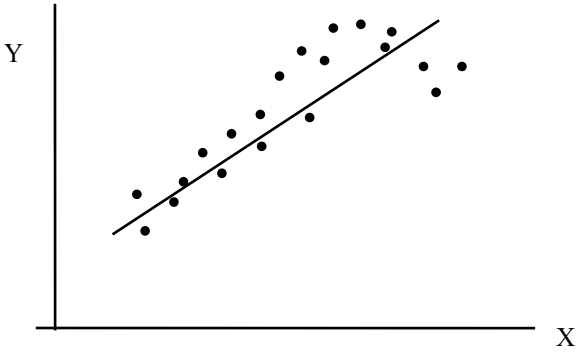


Fig. 12 La relación ente X y Y es parcialmente lineal

Caso 3. Prueba. *Ho*: $\beta_0 = 0$, Si **NO** se rechaza esta Hipótesis, puede ser apropiado ajustar un modelo sin β_0 , siempre y cuando exista experiencia previa o teoría que sugiera que la recta ajustada debe pasar por el origen y que existan datos alrededor del origen para mejorar la información sobre β_0 .

1.7 COEFICIENTE DE DETERMINACIÓN r^2

Retomando las sumas de cuadrados vistas anteriormente,

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

$$SC_{Total} = SC_{Regresión} + SC_{Error}$$

Despejando la suma de cuadrados de la regresión (explicada) y dividiendo ambos miembros entre la suma de cuadrados total, se tiene el coeficiente de determinación r^2.

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

$$r^2=\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2-\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=\frac{SC_{Reg}}{SC_{Total}}=1-\frac{SC_{Error}}{SC_{Total}}$$

Donde $r^2 \in (0.1)$

A diferencia del coeficiente de correlación, r, a esta r-cuadrada se le interpreta como una medida de la variación de la variable *Y* explicada por los cambios o variación en la variable *X*. Es común leerla como el porcentaje de variación en *Y* explicado por los cambios en *X*.

1.8 ANÁLISIS DE CORRELACIÓN

El análisis de regresión se usa cuando se busca establecer el tipo de relación que hay entre dos variables; pero cuando sólo interesa establecer *el grado de asociación lineal entre dos variables aleatorias* se usa el *análisis de correlación*.

La medida del grado de relación entre dos variables se llama coeficiente de correlación y se representa universalmente por ρ. En el modelo de correlación se asume que *X* y *Y* varían en una distribución conjunta. Si esta distribución está distribuida normalmente se llama *distribución normal bivariada*. Las suposiciones para un modelo de correlación lineal bivariada, para el cual se estima ρ, se describen a continuación:

1. *Y* y *X* son variables aleatorias. Como tales, no deben ser designadas como dependiente e independiente, cualquier designación dará el mismo resultado.

2. La población bivariada es normal. Una población normal bivariada es, entre otras cosas, aquella en la que *Y* y *X* están normalmente distribuidas.

3. La relación entre *Y* y *X* es, en cierto sentido, lineal. Este supuesto implica que todas las

medias de *Y* asociadas con valores de *X,* $\mu_{Y/X}$ caen sobre una línea recta, que es la línea de regresión de *Y* contra *X*. De la misma manera todas las medias de *X* asociadas con valores de *Y*, $\mu_{X/Y}$ caen sobre una línea recta, que es la línea de regresión de *X* contra *Y*.

Si se cumplen los supuestos antes descritos el coeficiente de correlación de Pearson de la población se define de la siguiente forma:

$$\rho = \frac{C_{X,Y}}{\sigma_X \sigma_Y} = \frac{\sum X_i Y_i - n\mu_X \mu_Y}{\sqrt{(\sum X_i^2 - n\mu_X^2)(\sum Y_i^2 - n\mu_Y^2)}}$$

Observaciones acerca de la definición anterior.

1) Es una ecuación que contiene los cinco parámetros de la población normal bivariada: μ_X, σ_X, μ_Y, σ_Y y ρ. El último, como se mencionó anteriormente, es el coeficiente de correlación para la población normal bivariada.

2) ρ es simétrico con respecto a *Y* y a *X*; es decir, el intercambio entre *X* y *Y* no cambia a ρ.

3) Cuando la covarianza es cero, ρ es cero, esto indica que no hay relación entre las variables.

4) Cuando hay covarianza perfecta entre *X* y *Y*, y ambas varían en la misma dirección, $\rho = 1$. De manera similar, cuando hay covarianza perfecta, pero *Y* y *X* varían en sentidos opuestos, $\rho = -1$ Por otra parte, cuando existe cierto grado de covarianza entre *X* y *Y*, se tiene

$-1 < \rho < 0$ y $0 < \rho < 1$ y, en general: $-1 < \rho < 1$

Cuando se extrae una muestra de *n* pares de valores, donde cada valor *x* es una observación al azar de la población *X* y cada valor *y* es una observación al azar de la población *Y*; pero las dos *no son* necesariamente *independientes*. Además, cuando se cumple el supuesto de una población normal bivariada el estimador de ρ está dado por *r*, el cual queda definido por:

$$r = \frac{\sum x_i y_i - n\bar{x}\,\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)(\sum y_i^2 - n\bar{y}^2)}} = \frac{\sum x_i y_i - n\bar{x}\,\bar{y}}{(n-1)s_x s_y}$$

Al igual que ρ puede variar entre –1 y 1. ($-1 < r < 1$), pero su interpretación se hará por medio de su cuadrado (r^2) **el *coeficiente de determinación***. Este valor puede emplearse como interpretación de la intensidad de la asociación entre las dos variables que parecen estar correlacionadas. De manera específica, el coeficiente de determinación indica el porcentaje de *la variación de Y* que *está asociada con* (o *"es explicada por*") la variación de *X,* o viceversa. Por ejemplo, si la correlación muestral entre dos variables tales como la clorofila y la biomasa es $r = 0.50$, elevando al cuadrado este coeficiente da un coeficiente de determinación $r^2 = 0.25$. Esto

sugiere que el 25% de la variación de una de las dos variables está asociada con o "es explicada por" la variación de la otra. No podemos decir cuál explica a cuál, porque ambas se consideran variables aleatorias, en otras palabras: éste coeficiente no establece relaciones causales.

Observe que *r* se puede calcular utilizando como base la fórmula de cálculo de b_1, esto es

$$r = \frac{b_1 s_x}{s_y}$$

¿Cómo encontrar la recta de regresión o de mínimos cuadrados en Statgraphics?

Sólo hay que seguir los pasos que se indican a continuación:

1. Crear un archivo de datos con dos columnas, una para la variable independiente (*X*) y otra para la variable dependiente (*Y*).

2. Del menú seguir la secuencia

Relate -> Simple Regresión

3. En el diálogo que aparece colocar en su lugar la variable dependiente (*Y*) y la independiente (*X*).

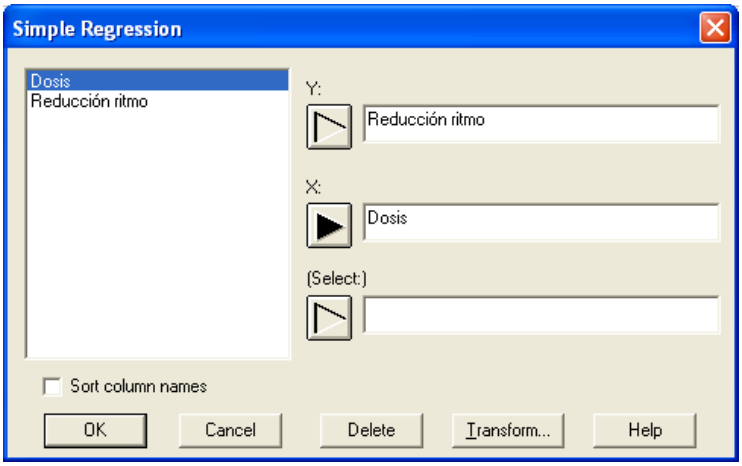


Fig. 13 Introducción de variables en el modelo de Regresión lineal simple

4. Dar OK.

RESULTADOS

Simple Regression - Reducción ritmo vs. Dosis

```
Regression Analysis - Linear model: Y = a + b*X
------------------------------------------------------------------------------
Dependent variable: Reducción ritmo
Independent variable: Dosis
------------------------------------------------------------------------------
                             Standard          T
Parameter       Estimate       Error       Statistic        P-Value
------------------------------------------------------------------------------
Intercept        7.05495      0.887572        7.94859         0.0000
Slope            4.08791      0.401991        10.1692         0.0000
------------------------------------------------------------------------------
```

INTERPRETACIÓN:
En la primera parte de los resultados se puede ver que se ajustó una recta Y = a + b*X, en esa tabla se observan los valores de los estimadores de la ordenada al origen (7.05495) (intercept) y de la pendiente (4.08791) (slope), los cuales coinciden con los obtenidos manualmente. Así mismo, se pueden ver los valores de los errores estándar de estos estimadores, los valores de la prueba t-student para los parámetros de regresión y los p-values, que como son 0.0000 < 0.05, indican que se rechazan las hipótesis ($H_0 : \beta_0 = 0$ y $H_0 : \beta_1 = 0$), por lo cual, podemos afirmar que Y depende de X, y la ordenada al origen es diferente de cero.

Esto quiere decir que la ecuación de la recta es:

Reducción ritmo = 7.05495 + 4.08791*Dosis

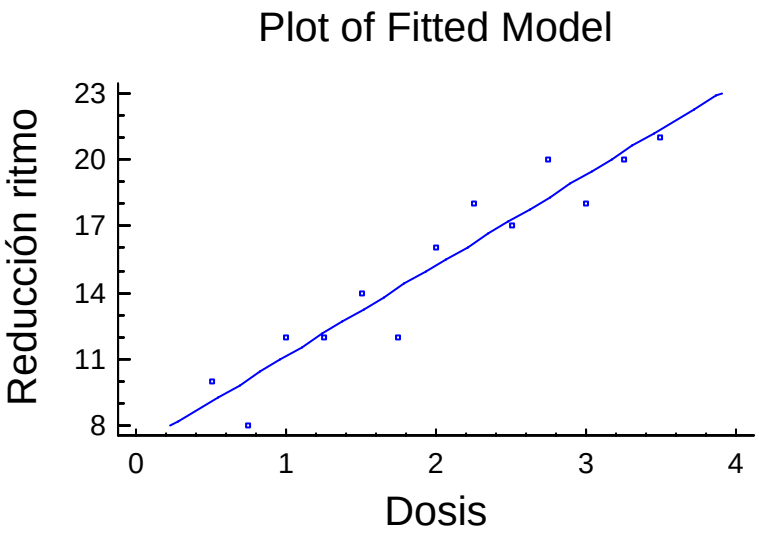
```
                          Analysis of Variance
------------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
------------------------------------------------------------------------------
Model                     190.088      1      190.088     103.41       0.0000
Residual                  20.2198     11      1.83816
------------------------------------------------------------------------------
Total (Corr.)             210.308     12

Correlation Coefficient = 0.950714
R-squared = 90.3856 percent
R-squared (adjusted for d.f.) = 89.5116 percent
Standard Error of Est. = 1.35579
Mean absolute error = 1.0448
Durbin-Watson statistic = 2.82148 (P=0.0237)
Lag 1 residual autocorrelation = -0.434072
```

INTERPRETACIÓN:
En la tabla de análisis de varianza se ve que p-value = 0.0000 < 0.05, lo que indica que el modelo lineal es significativo. También vemos que el valor de r^2 ajustada es 89.5116, indicando que 89.51% de la variación de Y es explicada por la variación o los cambios en la variable X (Dosis).

A continuación de esta tabla están los valores de los coeficientes de correlación lineal, r, de determinación r^2, el error estándar de la regresión y otros valores que posteriormente se explican.

También se puede ver el "StatAdvisor" que es justamente la interpretación de las dos tablas descritas.

```
The StatAdvisor
---------------
   The output shows the results of fitting a linear model to describethe relationship between
Reducción ritmo and Dosis. The equation of the fitted model is

   Reducción ritmo = 7.05495 + 4.08791*Dosis

Since the P-value in the ANOVA table is less than 0.01, there is a statistically significant
relationship between Reducción ritmo and Dosis at the 99% confidence level.

   The R-Squared statistic indicates that the model as fitted explains 90.3856% of the
variability in Reducción ritmo. The correlation coefficient equals 0.950714, indicating a
relatively strong relationship between the variables. The standard error of the estimate shows
the standard deviation of the residuals to be 1.35579. This value can be used to construct
prediction limits for new observations by selecting the Forecasts option from the text menu.

   The mean absolute error (MAE) of 1.0448 is the average value of the residuals.  The Durbin-
Watson (DW) statistic tests the residuals to determine if there is any significant correlation
based on the order in which they occur in your data file.  Since the P-value is less than 0.05,
there is an indication of possible serial correlation. Plot the residuals versus row order to see
if there is any pattern which can be seen.
```



La tercera parte del análisis es el diagrama de dispersión con la recta ajustada. Esta gráfica muestra que si hay una tendencia lineal en los datos. En los resultados de correlación y determinación se tiene que el modelo explica un 95.07% de la variación en los valores de *Y*, por efecto de los cambios en *X*.

La r^2 *ajustada* indica un 89.512% de variación explicada de la variable *Y* por efecto de los cambios en la variable *X*.

NOTA: Se pueden ordenar otros análisis al paquete, sin embargo, éstos se van a ir realizando a medida que se vayan explicando los conceptos.

1.9 DIAGNÓSTICO DEL MODELO DE REGRESIÓN

Los procedimientos o técnicas de diagnóstico se aplican esencialmente para detectar desacuerdos entre el modelo y los datos para los cuales se ajusta.

Las técnicas de diagnóstico más utilizadas son las que verifican los supuestos del modelo, los casos extraordinarios (outlires) y la colinealidad, ésta última utilizada principalmente en la regresión lineal múltiple.

Los supuestos que se hacen del estudio del análisis de regresión se determinan a través de los residuos y son:

1. La relación entre *Y* y *X* es lineal.
2. Los errores tienen media cero
3. Los errores tienen varianza constante σ^2.
4. Los errores no están correlacionados.
5. Los errores se distribuyen normalmente.

Los supuestos 4 y 5 implican que los errores son variables aleatorias independientes.

1.10 EXAMINANDO LOS RESIDUOS

Los *residuos o errores* se definen como las n diferencias $e_i = y_i - \hat{y}_i$, i = 1, 2, …, n, donde y_i es una observación y $\hat{y}_i$ es el valor ajustado obtenido al usar la ecuación de regresión ajustada. De acuerdo con esta definición, los residuos e_i son la diferencia entre lo observado y lo ajustado por la ecuación de regresión. Este error es una medida de la variabilidad no explicada por el modelo de regresión.

Como se mencionó anteriormente, los residuos se usan para detectar si se viola algún supuesto al ajustar un modelo específico a los datos del problema que se esté trabajando. Las posibles violaciones al modelo se pueden tipificar como:

- Presencia de casos extraordinarios (outliers) en los datos.
- Evidencia que sugieren varianza no constante.
- Evidencia de que la distribución de los errores no proviene de una distribución normal.
- Evidencias que sugieren que la forma del modelo no es la apropiada
- Autocorrelación, que se define como la falta de independencia de los errores.

El análisis de los residuos se basa en gráficas que despliegan las características generales de los residuos, y con base en éstas y a ciertos patrones que deben seguir se puede determinar la violación de los supuestos.

1.11 TIPOS DE RESIDUOS

a) Residuos ordinarios

Los residuos tienen varias propiedades importantes, como: media cero y su varianza aproximada es:

$$CM_E = \frac{SC_E}{n-2} = \frac{\sum_{i=1}^{n}(e_i - \overline{e}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

Nota: Observe que $\overline{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$

Los residuos no son independientes, sin embargo tienen sólo $n - 2$ grados de libertad asociados a ellos; esto es que los residuos ordinarios involucran las unidades del problema original. Esta no independencia de los residuos tienen un pequeño efecto en la investigación de la adecuación del modelo siempre que n no sea pequeño.

b) Residuos estandarizados

Una forma de expresión de los residuos es dividirlos entre su desviación estándar de la siguiente forma:

$$d_i = \frac{e_i}{\sqrt{CM_e}} \qquad [10]$$

Los residuos estandarizados tienen media cero y varianza aproximadamente unitaria (valor cercano a 1).

c) Residuos estudentizados

Dado que los residuos ordinarios involucran las unidades del problema y los resultados que arroja el análisis de los mismos se ven alterado por estas unidades, se busca otra forma de expresarlos, la cual no dependa de escala alguna.

En la expresión [10], los residuos se dividen por su desviación estándar promedio, sin embargo, en algunos grupos de datos, las desviaciones estándar pueden discrepar altamente. En regresión lineal simple:

$$\begin{aligned} V(e_i) &= V(y_i - \hat{y}_i) \\ &= V(y_i) + V(\hat{y}_i) - 2Cov(y_i - \hat{y}_i) \\ &= \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(x_i - \overline{x})^2}{s_{xx}}\right] - 2Cov(y_i - \hat{y}_i) \end{aligned}$$

Se puede mostrar

$$Cov(y_i - \hat{y}_i) = Cov\left[y_i, \bar{y} + \frac{s_{xy}}{s_{xx}}(x_i - \bar{x}) \right] = \sigma^2 \left[\ddot{1}{n} + \frac{(x_i - \bar{x})^2}{s_{xx}} \right]$$

Consecuentemente, la varianza del i-ésimo residuo es:

$$V(e_i) = \sigma^2 \left[1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{s_{xx}} \right) \right]$$

Por lo tanto, los residuos estudentizados se definen como:

$$r_i = \frac{e_i}{\sqrt{CM_e \left[1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_x^2} \right) \right]}} \qquad [11]$$

Es importante notar que en la expresión [11], el residuo e_i ordinario de mínimos cuadrados se ha dividido por el error estándar exacto, más que por un valor promedio como en [10]. Los residuos estudentizados son extremadamente útiles en diagnóstico de regresión lineal simple y múltiple.

En un grupo pequeño de datos los residuos estudentizados tienen con frecuencia una escala más apropiada que los residuos estandarizados debido a que las diferencias de las varianzas residuales de los dos tipos de residuos son más amplias. Cuando *n* es grande, entonces la diferencia de escala de los residuos de mínimos cuadrados ente los dos métodos será pequeña.

1.12 TIPOS DE GRÁFICAS

Los residuos se pueden usar en una variedad de gráficos para identificar si se presenta alguna violación de supuestos.

Generalmente se hacen diferentes gráficas para comprobar la idoneidad del modelo o para detectar sus inadecuaciones. Estos métodos son simples y efectivos y frecuentemente recomendados en el análisis de regresión.

El tipo de residuo utilizado frecuentemente son los estudentizados, debido a que estos no dependen de las unidades del contexto del problema y por lo que es mejor trabajar con ellos y hacer un buen análisis.

a) Gráficas de normalidad

Un *método simple para probar el supuesto de normalidad* es el de graficar los residuos contra una escala especial que está relacionada con los percentiles de una distribución normal estándar.

Al graficar los residuos, si éstos se comportan de acuerdo a una línea recta se cumple el supuesto de normalidad de los errores. Los residuos ordenados $e_{(i)}$ están usualmente graficados contra los

"valores normales esperados" $\Phi^{-1}[(i-½)/n]$, donde Φ denota la función distribución acumulada normal estándar.

En la figura 14*a*, se despliega una gráfica de probabilidad normal "ideal", en las figuras 14*b* a 14*e* se presentan algunos problemas típicos.

En algunos estudios se ha establecido que las gráficas de normalidad realizadas con muestras pequeñas ($n \leq 16$), frecuentemente se desvían sustancialmente de la linealidad. Para muestras grandes ($n \geq 32$), las gráficas se comportan mucho mejor. Usualmente se requiere alrededor de 20 puntos para obtener gráficos de probabilidad normal estables y lo suficientemente fáciles de interpretar.

Aunque pequeñas desviaciones de la normalidad no afectan grandemente al modelo, la no normalidad bruta es potencialmente más seria si se considera que los estadísticos *t* y *F*, así como los intervalos de confianza y de predicción dependen del supuesto de normalidad.

Además, si los errores vienen de una distribución con colas más pronunciadas que la normal, el ajuste por mínimos cuadrados puede ser sensible a un subgrupo pequeño de datos. La distribución de errores de colas pronunciadas con frecuencia genera casos extraordinarios (outliers) que "jalan" el ajuste por mínimos cuadrados demasiado en su dirección. En esos casos se puede considerar alguna otra técnica de estimación.

Un defecto común que se presenta en las gráficas de normalidad es la ocurrencia de uno o 2 residuos grandes. Algunas veces éste es un indicativo de que las observaciones correspondientes son casos extraordinarios.

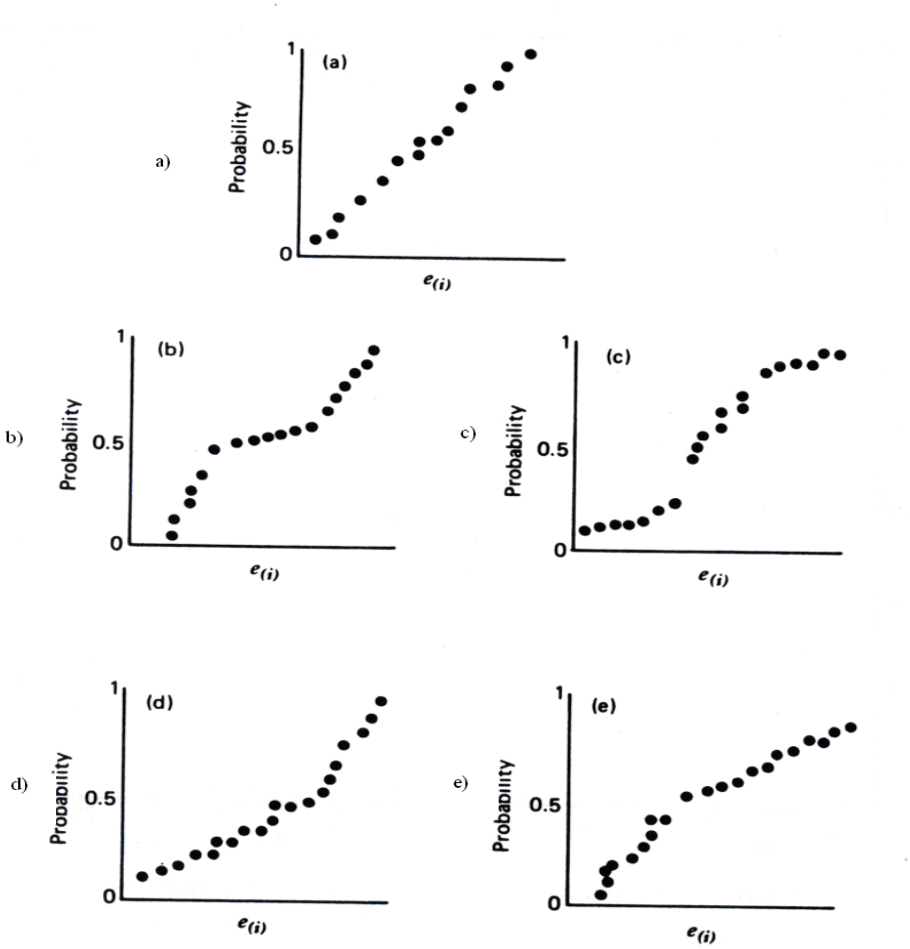


Fig. 14 Curvas de probabilidad normal: (a) ideal; (b) distribución con colas pronunciadas; (c) distribuciones con colas ligeras; (d) sesgo positivo; (e) sesgo negativo.

El supuesto de normalidad también puede comprobarse construyendo un histograma de residuos; sin embargo, frecuentemente el número de residuos es demasiado pequeño para permitir una fácil identificación visual de la forma de la distribución normal.

Los residuos estandarizados o estudentizados también se usan en la detección de desviaciones de la normalidad. Si los errores están normalmente distribuidos, entonces aproximadamente el 68% de los residuales estandarizados se encuentren entre -1 y +1, aproximadamente el 95% de ellos se encuentran entre -2 y +2. Una desviación sustancial de estos límites indica una violación potencial del supuesto de normalidad. Si *n* es pequeña, se pueden sustituir los límites ± 1 y ± 2 por los valores correspondientes a la distribución t_{n-2}. El examen de los residuos estandarizados es, de esta manera, también útil en la identificación de "outliers".

b) Gráfica de residuos contra los valores ajustados ($\hat{y}_i$).

Esta gráfica contrasta los residuos contra los valores ajustados y puede dar indicio de diferentes violaciones de los supuestos. Los residuos e_i, se grafican vs. los valores ajustados $\hat{y}_i$ **,** y NO los

observados y_i, debido a que los e_i y los $\hat{y}_i$ no están correlacionados, mientras que los e_i y los y_i están usualmente correlacionados. Si al hacer la gráfica, la nube de puntos se concentra alrededor de una línea horizontal (Fig. 15*a*), esto indica que no hay anormalidad en el modelo y que el análisis de mínimos cuadrados es idóneo.

En caso de que la gráfica no se presente de esta manera (Figs. 15*b*-15*d*), se puede pensar en una deficiencia del modelo.

El patrón de la Fig. 15*b* indica que la varianza de los errores no es constante. Este patrón de embudo de apertura hacia fuera implica que la varianza es una función creciente de y (un embudo de apertura hacia adentro es también posible, e indica que la $V(e)$ se incrementa cuando y decrece). En este caso se puede aplicar un análisis de mínimos cuadrados ponderados o una transformación de las observaciones y_i antes de "correr" la regresión.

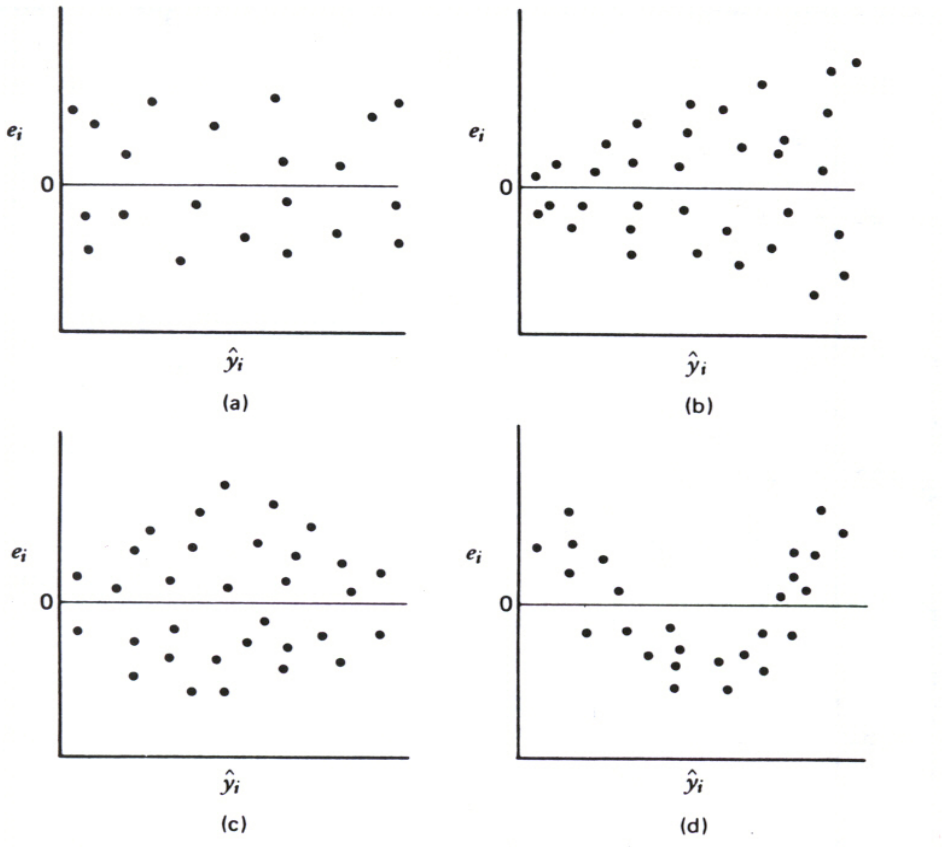


Fig. 15 Patrones de residuos (a) satisfactorio; (b) embudo; (c) doble arco; (d) no lineal

El patrón de doble arco (Fig. 15*c*) frecuentemente se presenta cuando y es una proporción entre 0 y 1 (Binomial). La varianza de una proporción binomial cerca de 0.5 es más grande que una cercana a 0 o a 1.

Un tratamiento usual para corregir la desigualdad de la varianza es aplicar una adecuada transformación a *una variable explicativa, a la variable respuesta* o *utilizar el método de mínimos cuadrados ponderados*. En la práctica es frecuente hacer las transformaciones sobre las variables de respuestas, para estabilizar varianzas.

Un gráfico curveado como el de la Fig. 15*d*, indica no linealidad. Esto significa que necesita otra variable como variable explicativa en el modelo que puede ser un término cuadrático o la transformación de la variable explicativa.

Un gráfico de residuos vs. $\hat{y}_i$ puede también revelar uno o más residuos inusualmente grandes. Estos puntos serán "outliers" potenciales. Los residuos grandes que se presentan en el extremo de los valores $\hat{y}_i$ pueden indicar que cualquiera de las varianzas no son constantes o que la verdadera relación entre y y x es no lineal.

c) Gráficas de residuales contra cada variable x_i.

Este tipo de gráficas al igual que la anterior sirve para *diagnosticar no linealidad y varianza no constante en los errores*. Al graficar contra cada x_i, el análisis que se realiza se refiere a cada x_i por separado de tal manera que la información que arrojan estas gráficas sirve para transformar el modelo en caso de que dependa de una variable en particular.

Estas gráficas frecuentemente muestran patrones tales como los de las figuras 10, a excepción de que la escala horizontal es x_i en vez de $\hat{y}_i$. De la misma manera, una distribución al azar y alrededor de la recta horizontal muestra un comportamiento idóneo. Los patrones de embudo o doble arco indican varianza no constante. La banda curva implica que posiblemente es necesario incluir otra variable explicativa o que es necesaria una transformación.

d) Otras gráficas de residuos

Si *se conoce la secuencia de tiempo* en la cual los datos fueron colectados, se puede realizar una gráfica de residuos contra el tiempo ordenado. Esto puede indicar que la varianza cambia con el tiempo o que el término lineal o cuadrático en tiempo se puede adicionar al modelo.

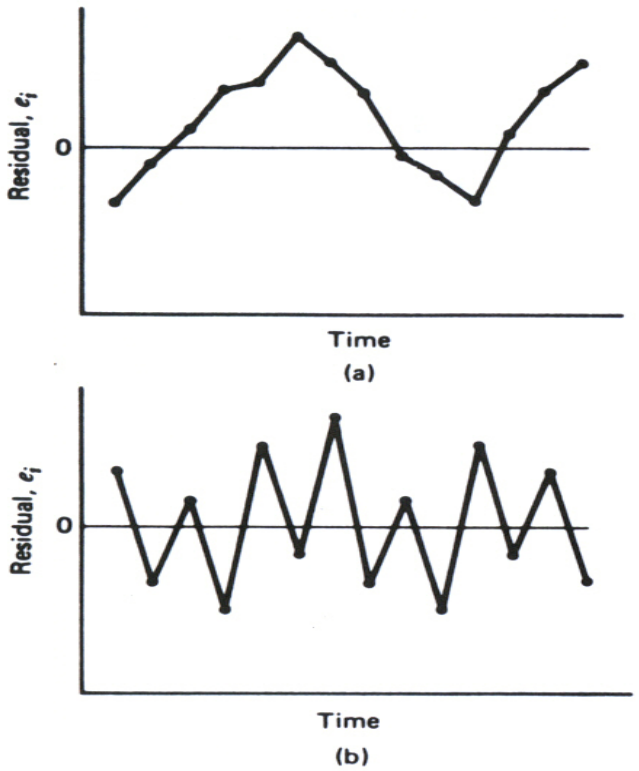


Fig. 16 Prototipos de gráficas de residuos contra el tiempo mostrando autocorrelación de los errores; (a) autocorrelación positiva; (b) autocorrelación negativa.

La gráfica de residuos de secuencia de tiempo puede indicar que los errores en un período de tiempo están correlacionados con éste u otros períodos de tiempo. La correlación entre errores del modelo y los diferentes períodos de tiempo se llama *autocorrelación*.

Una figura como 16*a*, indica autocorrelación positiva, mientras que la figura 16*b* es típica de la autocorrelación negativa. La presencia de autocorrelación en los errores indica una seria violación de los supuestos básicos de regresión.

1.13 DETECCIÓN Y TRATAMIENTO DE "OUTLIERS"

Un "outlier" es una observación extrema, entre residuales es una observación que está más lejana que el resto. Los residuos que son más grandes, en valores absolutos, que el resto, aproximadamente 3 o 4 desviaciones estándar de la media son potenciales outliers. Los outliers son puntos atípicos al resto de los datos. Dependiendo de su localización en el espacio *X*, los outliers pueden tener efectos en el modelo de regresión de *moderados a graves.*

Las gráficas de residuos vs. $\hat{y}_i$ y las gráficas de probabilidad normal son de ayuda para identificar outliers. Examinar los residuales escalados (estandarizados o estudentizados) es una excelente vía para identificar potenciales outliers.

Los outliers pueden analizarse cuidadosamente para ver si se puede encontrar una razón para su comportamiento inusual. Algunas veces los outliers son "malos" valores que se presentan como resultados inusuales pero como eventos explicables. Por ejemplo, se incluyen medidas o análisis defectuosos, registros de datos incorrectos o fallas de algún instrumento de medición. Si éste es el caso, entonces el outlier se puede corregir (si es posible) o borrado del grupo de datos. Evidentemente descartar malos valores es deseable porque la ecuación ajustada por mínimos cuadrados sin este dato extraordinario minimiza la suma de cuadrados de los residuos.

Algunas veces encontramos que los outliers son observaciones inusuales pero perfectamente admisibles. La eliminación de estos puntos para "mejorar el ajuste de la ecuación" puede dar un falso sentido de precisión en la estimación o predicción. Ocasionalmente encontramos que los outliers son más importantes que el resto de los datos porque pueden controlar algunas propiedades claves del modelo detectando condiciones o alteraciones experimentales importantes. Los outliers pueden ser puntos fuera de las inadecuaciones del modelo, tales como fallas para ajustar bien los datos en una cierta región del *X*-espacio.

Existen varias pruebas estadísticas propuestas para detectar y rechazar outliers. Aunque estas pruebas pueden utilizarse para identificar outliers, no pueden utilizarse para determinar que estos puntos deban ser automáticamente rechazados.

El efecto de los outliers en el modelo de regresión se puede verificar fácilmente mediante un análisis detallado de esos puntos y reajuste de la ecuación de regresión. Se sabe que los valores de los coeficientes de regresión o los resúmenes estadísticos, tales como t, F, R^2 y los cuadrados medios de los residuos pueden ser muy sensibles a la presencia de outliers.

1.14 TRANSFORMACIONES QUE ESTABILIZAN LA VARIANZA

Cuando la varianza de los errores no es constante sobre todas las observaciones, se dice que el error es heterocedástico. La heterocedasticidad se puede remover por medio de una adecuada transformación y puede tener una distribución en la que la varianza dependa de los datos.

Una razón común de la violación de este supuesto es porque la variable respuesta *Y*, sigue una distribución de probabilidad en la cual la varianza está funcionalmente relacionada con la media. Por ejemplo, si *Y* es una variable aleatoria Poisson, entonces la varianza de *Y* es igual a la media. Puesto que la media de *Y* está relacionada con la variable explicativa *X*, la varianza de *Y* será proporcional a *X*. Las transformaciones para estabilizar la varianza son utilizadas frecuentemente en estos casos. De esta manera, si la distribución de *Y* es Poisson podemos aplicar antes de hacer la regresión de $Y' = \sqrt{Y}$, contra *X*, puesto que la varianza de la raíz cuadrada de una variable aleatoria Poisson es independiente de la media. Otro ejemplo es, si la variable respuesta es una

proporción ($0 \le Y_i \le 1$) y la gráfica de residuos vs. $\hat{y}_i$ tiene un patrón de doble arco (Fig. 10*c*), entonces la transformación $Y' = arcsen\sqrt{Y}$ es la apropiada.

Algunas transformaciones comúnmente usadas que estabilizan la varianza se resumen a continuación:

Relación de σ^2 a E(Y)	**Transformación**
$\sigma^2 \propto$ constante	$Y' = Y$ (Sin transformación)
$\sigma^2 \propto E(Y)$	$Y' = \sqrt{Y}$ (Raíz cuadrada; datos Poisson)
$\sigma^2 \propto E(Y)\,[1 - E(Y)]$	$Y' = arcsen\sqrt{Y}$, (arco seno; proporción binomial $0 \le Y_i \le 1$)
$\sigma^2 \propto [E(Y)]^2$	$Y' = Ln(Y)$ (logaritmo)
$\sigma^2 \propto [E(Y)]^3$	$Y' = \frac{1}{\sqrt{Y}})$ (Raíz cuadrada recíproca)
$\sigma^2 \propto [E(Y)]^4$	$Y' = \frac{1}{Y})$ (Recíproca)

La fortaleza de una transformación depende de la cantidad de curvatura que ésta induce. Las transformaciones se dan desde una relativa suave raíz cuadrada hasta una relativamente fuerte recíproca. Generalmente una transformación suave se aplica en un intervalo estrecho de valores (por ejemplo, $Y_{máx}/Y_{min} < 2.3$) y se tiene un efecto pequeño, mientras que una transformación fuerte en un amplio intervalo de valores tendrá efectos más dramáticos en el análisis.

A veces se pueden usar experiencias anteriores o consideraciones teóricas como guía para seleccionar transformaciones apropiadas. Sin embargo, en algunos casos no tenemos razones *a priori* para sospechar que al varianza del error no es constante. Una primera indicación del problema se realiza desde *la inspección del diagrama de dispersión* o *en el análisis de residuos,* en estos casos la transformación aplicada se puede seleccionar empíricamente.

Esto es importante para detectar y corregir varianzas del error no constantes. Si este problema no se elimina, los estimadores de mínimos cuadrados serán aún imparciales pero no podrán ser más grandes que la varianza mínima.

En el STATGRAPHICS se pueden ver los residuos, tanto con las opciones tabulares como con las gráficas, que se pueden obtener al dar clic en los **cuadros amarillo y azul**, en el lado superior izquierdo de la barra de menú

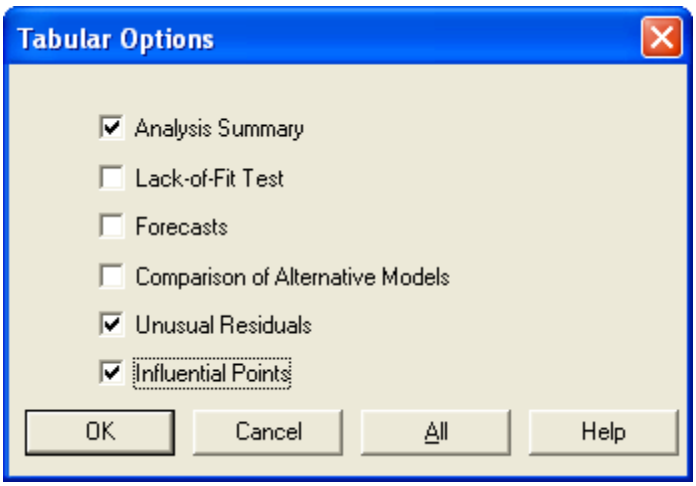


Fig. 17 Opciones Tabulares

- Forecasts sirve para determinar los intervalos de confianza de predicción y de valores esperados con el objeto de hacer pronósticos.

- Lack-of-Fit Test, es una prueba para probar la falta de ajuste del modelo. Funciona solamente cuando se tienen repeticiones de la variable *Y* en cada uno de los valores de *X*. Aquí se prueba la Ho: No hay falta de ajuste (el modelo es adecuado) contra Ha: Hay falta de ajuste (el modelo no es el adecuado).

- Comparision of Alternative Models. Con base en las correlaciones se define si hay algún otro modelo que explique mejor el comportamiento de los datos.

En las opciones gráficas se recomienda seleccionar todos los gráficos de residuales, ya que permiten analizar rápidamente el cumplimiento de supuestos.

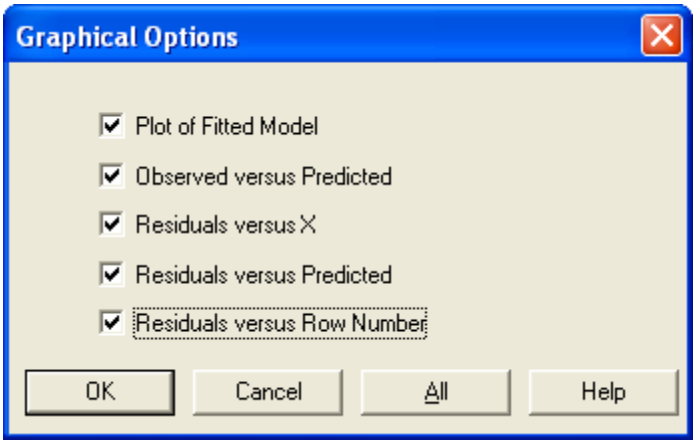


Fig. 18 Opciones Gráficas

Continuando con el ejemplo 1, se examinan las gráficas de los residuos y con ello el cumplimiento de los supuestos.

RESULTADOS

Unusual Residuals

Row	X	Y	Predicted Y	Residual	Studentized Residual

The StatAdvisor

The table of unusual residuals lists all observations which have Studentized residuals greater than 2.0 in absolute value. Studentized residuals measure how many standard deviations each observed value of Reducción ritmo deviates from a model fitted using all of the data except that observation. In this case, there are no Studentized residuals greater than 2.0.

INTERPRETACIÓN:
En este ejemplo no hay residuos studentizados en valor absoluto que sean mayores que 2.00, por lo que no hay de que preocuparnos.

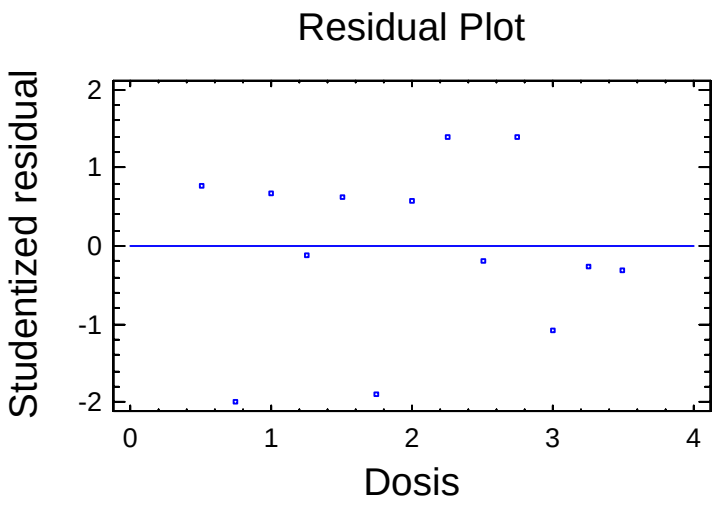
Influential Points

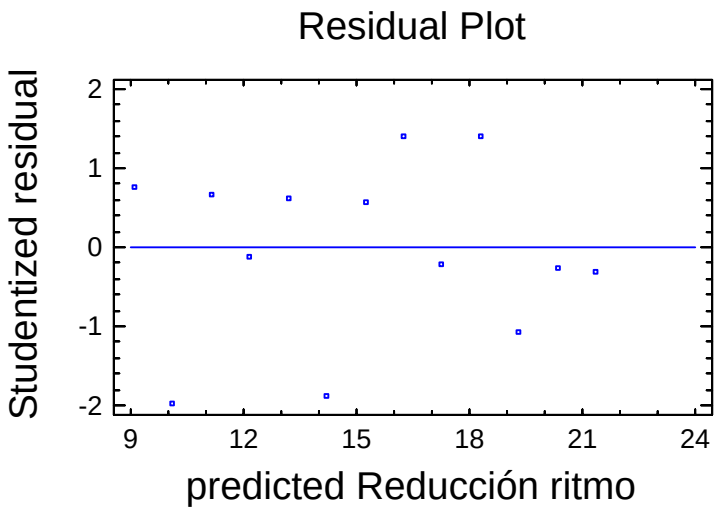
Row	X	Y	Predicted Y	Studentized Residual	Leverage

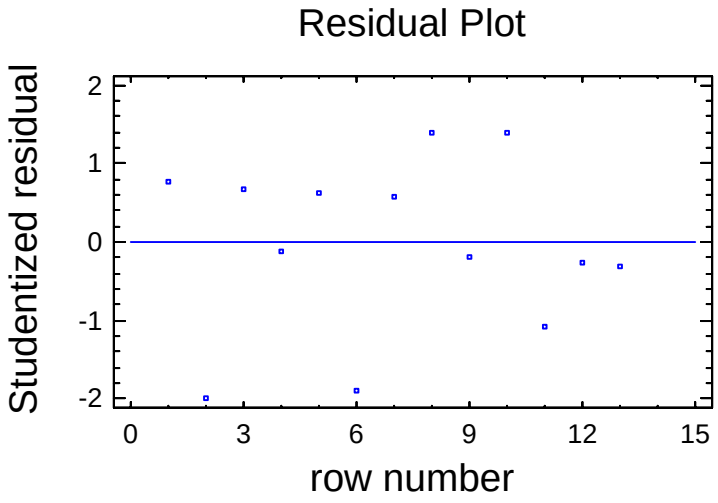
Average leverage of single data point = 0.153846

The StatAdvisor

The table of influential data points lists all observations which have leverage values greater than 3 times that of an average data point. Leverage is a statistic which measures how influential each observation is in determining the coefficients of the estimated model. In this case, an average data point would have a leverage value equal to 0.153846. There are no data points with more than 3 times the average leverage.







INTERPRETACIÓN:

Los gráficos de residuos muestran que la nube de puntos se concentra alrededor de una línea horizontal, esto indica que no hay anormalidad en el modelo y que el análisis de mínimos cuadrados es idóneo. Por otra parte, el patrón no muestra ninguna tendencia de embudo o arco, lo que indica que la varianza de los errores es constante.

1.15 INFERENCIA EN EL ANÁLISIS DE REGRESIÓN LINEAL

Una vez que se confirma que el modelo cumple con los supuestos de la regresión lineal, se puede realizar la inferencia. Cabe señalar que las inferencias sólo son válidas para la regresión lineal (no curvilínea, a menos que se haya linealizado) y cuando se cumplen los supuestos.

Aunque la recta de mínimos cuadrados es la recta que mejor se ajusta a los puntos, todavía

muchos de éstos se desvían de ella. La medida numérica de tales desviaciones es el estimador insesgado de la **varianza de los errores** de ajuste por la regresión, que se define como:

$$s^2_{y/x} = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

Sustituyendo $\hat{y} = b_0 + b_1 x$

$$s^2_{y/x} = \frac{\sum (y_i - b_0 - b_1 x)^2}{n-2} \qquad [12]$$

Sustituyendo la ecuación. $b_0 = \overline{y} - b_1 \overline{x}$

$$s^2_{y/x} = \frac{\sum (y_i - \overline{y} + b_1 \overline{x} - b_1 x_i)^2}{n-2}$$

Agrupando

$$s^2_{y/x} = \frac{\sum [(y_i - \overline{y}) - b_1 (x_i - \overline{x})]^2}{n-2} \qquad [13]$$

Al desarrollar el cuadrado, aplicar propiedades de las sumatorias y sustituir la ecuación [7] en la expresión resultante se llega a:

$$s^2_{y/x} = \frac{n-1}{n-2}(s^2_y - b_1^2 s^2_x) \qquad [14]$$

La raíz cuadrada positiva de la expresión [11] es llamada *Error Estándar de Estimación;* esto es:

$$s_{y/x} = \sqrt{\frac{n-1}{n-2}(s^2_y - b_1^2 s^2_x)}$$

1.15.1 ESTIMACIÓN Y PRUEBA DE HIPÓTESIS PARA LA PENDIENTE DE LA RECTA DE REGRESIÓN POBLACIONAL β_1

Para probar la hipótesis $H_0 : \beta_1 = (\beta_1)_0$, se utiliza la distribución *t*-student con (n – 2) grados de libertad ya que se desconoce la varianza poblacional $\sigma^2_{b_1}$; el estimador de esta varianza se define por:

$$s^2_{b_1} = \frac{s^2_{y/x}}{(n-1)s^2_x}$$

Por lo tanto, el estadístico de prueba se define como:

$$t = \frac{b_1 - (\beta_1)_0}{\frac{s_{y/x}}{s_x\sqrt{n-1}}}, \text{ con } g.l. = n - 2$$

El intervalo de confianza de nivel 1 - α está dado por:

$$b_1 - t_{1-\frac{\alpha}{2},n-2}\frac{s_{y/x}}{s_x\sqrt{n-1}} < \beta_1 < b_1 + t_{1-\frac{\alpha}{2},n-2}\frac{s_{y/x}}{s_x\sqrt{n-1}}$$

1.15.2 PRUEBA DE INDEPENDENCIA

Una prueba muy sencilla y útil es la prueba de *independencia* entre las variables. Si la hipótesis nula $H_0 : \beta_1 = 0$ no se rechaza, *Y* no depende linealmente de *X*, en caso contrario se dice que *Y* depende de *X*.

1.15.3 ESTIMACIÓN Y PRUEBA DE HIPÓTESIS PARA LA ORDENADA AL ORIGEN β_0

Para probar la hipótesis $H_0 : \beta_0 = (\text{valor supuesto})$, se utiliza la distribución *t*-student con *n* - 2 grados de libertad y cuya expresión está dada por:

$$t = \frac{b_0 - (\beta_0)_0}{s_{y/x}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}}}, \text{ con } g.l. = n - 2$$

El intervalo de confianza de nivel 1 - α está dado por:

$$b_0 - t_{1-\frac{\alpha}{2},n-2}s_{y/x}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}} < \beta_0 < b_0 + t_{1-\frac{\alpha}{2},n-2}s_{y/x}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}}$$

1.15.4 ESTIMACIÓN DE LA MEDIA $\mu_{y/x}$

A veces es conveniente estimar el valor medio o *esperado* de *Y* para un valor dado de *X*, tal estimación se hace con el intervalo de confianza siguiente:

$$\hat{y}_0 - t_{1-\frac{\alpha}{2},n-2} s_{Y/X} \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{(n-1)s_x^2}} < \mu_{y/x} < \hat{y}_0 + t_{1-\frac{\alpha}{2},n-2} s_{y/x} \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{(n-1)s_x^2}}$$

Donde $\hat{y}_0 = b_0 + b_1 x_0$

Los límites de confianza para los valores medios, se deben calcular para cada valor de x_0; tales límites serán más estrechos a medida que se aproximen a la media de la variable independiente y más amplios a medida que se alejan de ella. Por esta razón se obtienen límites curvos llamados bandas de confianza dentro de los cuales queda comprendida la recta verdadera para un nivel de significación α.

Para trazar las bandas de confianza se elegirán cuando menos tres valores de x_0, dos valores extremos y uno intermedio, y se harán las estimaciones por intervalo. Estos intervalos se dibujan sobre el diagrama de dispersión uniendo todos los puntos generados por los límites inferiores y, por otro lado, todos los correspondientes a los límites superiores.

1.15.5 INTERVALO DE CONFIANZA PARA UN VALOR DE PREDICCIÓN O DE PRONÓSTICO $\hat{Y}$

Como las bandas de confianza se abren a medida que x_0 se aleja de la media $\overline{x}$ resulta aventurado hacer predicciones sobre *Y* para valores fuera del intervalo de valores de *X* empleado para establecer la ecuación de regresión de la muestra. A menos que se esté razonablemente seguro que existe la misma función de regresión sobre un amplio rango de valores de *X*, por haberse experimentado en una muestra, se podrá usar valores de x_0 alejados de $\overline{x}$ para predecir valores de *Y*.

Cuando se hacen predicciones, éstas, más que ser sobre valores medios de *Y*, ($\mu_{y/x}$ **),** son para valores individuales $\hat{Y}$, por lo tanto, el error al azar (error de estimación ε_i) es una fuente adicional de variación, que produce el aumento de la varianza y que el intervalo de confianza para un valor de predicción $\hat{Y}$ sea más amplio que el de $\mu_{y/x}$**.**

$$\hat{y} - t_{1-\frac{\alpha}{2},n-2} s_{y/x} \sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)s_x^2}} < \hat{Y} < \hat{y} + t_{1-\frac{\alpha}{2},n-2} s_{y/x} \sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)s_x^2}}$$

1.15.6 INTERVALO DE CONFIANZA PARA LA VARIANZA DEL ERROR DE REGRESIÓN $\sigma^2_{Y/X}$

Otra estimación por intervalo que es conveniente realizar es la estimación de la varianza, $\sigma^2_{Y/X}$ de los errores de regresión. Esta estimación con (1 - α)100 de confianza, está dada por:

$$\frac{s^2_{y/x}(n-2)}{\chi^2_{1-\frac{\alpha}{2},n-2}} < \sigma^2_{y/x} < \frac{s^2_{y/x}(n-2)}{\chi^2_{1-\frac{\alpha}{2},n-2}}$$

Si se desea estimar el error estándar de estimación $\sigma_{Y/X}$ se trabaja con el intervalo para la varianza y al final se extrae la raíz cuadrada a los límites de confianza obtenidos para la varianza.

Continuado con el ejemplo 1, calculemos ahora en el STATGRAPHICS los intervalos de confianza del 95% para: la pendiente, la ordenada al origen, la media de regresión y la predicción, estos últimos cuando la dosis es $x_0 = 2.4$

Con el archivo de datos que ya se tenía en STATGRAPHICS, con terminación (.sf3) ir a opciones tabulares y checar "**Forecasts**", dar OK. Luego dar botón derecho y seleccionar PANE OPTIONS, En esta opción cambiar Forecast at x, por 2.4 y dar OK

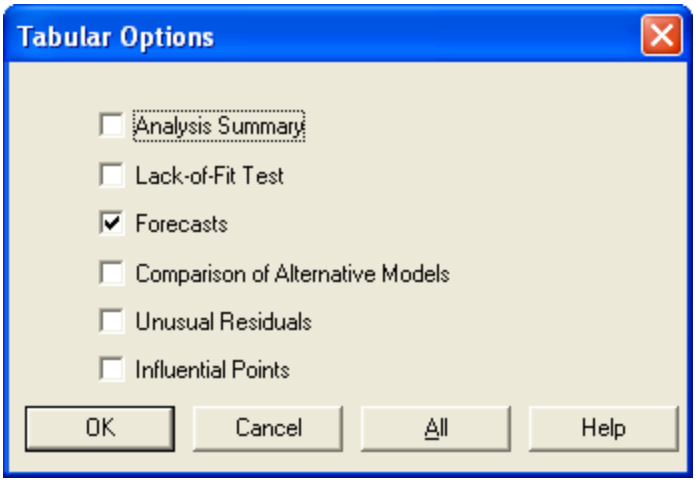


Fig. 19 Opciones Tabulares

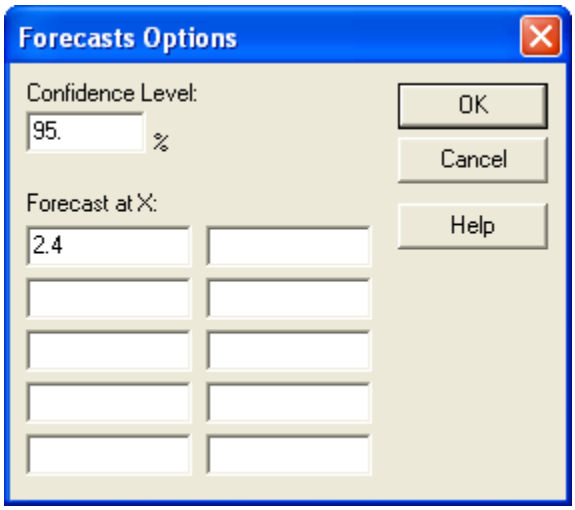


Fig. 20 Opciones de pronóstico

RESULTADOS:

Simple Regression - Reducción ritmo vs. Dosis

```
Regression Analysis - Linear model: Y = a + b*X
------------------------------------------------------------------------------
Dependent variable: Reducción ritmo
Independent variable: Dosis
------------------------------------------------------------------------------
                              Standard          T
Parameter       Estimate         Error       Statistic        P-Value
------------------------------------------------------------------------------
Intercept        7.05495       0.887572        7.94859         0.0000
Slope            4.08791       0.401991        10.1692         0.0000
------------------------------------------------------------------------------

Predicted Values
------------------------------------------------------------------------------
                                      95.00%                   95.00%
               Predicted        Prediction Limits         Confidence Limits
          X            Y        Lower        Upper        Lower        Upper
------------------------------------------------------------------------------
        2.4      16.8659      13.7491      19.9828      15.9658      17.7661
------------------------------------------------------------------------------

The StatAdvisor
---------------
   This table shows the predicted values for Reducción ritmo using the fitted model. In addition
to the best predictions, the table shows:
   (1) 95.0% prediction intervals for new observations
   (2) 95.0% confidence intervals for the mean of many observations
The prediction and confidence intervals correspond to the inner and outer bounds on the graph of
the fitted model.
```

INTERPRETACIÓN

De la primera tabla (del modelo de regresión, obtenida anteriormente) se tienen los valores de los estimadores de los parámetros de la recta con sus errores estándar, por lo que se pueden calcular los intervalos de confianza del 95% para β_1 y β_0.

$$4.08791 - t_{0.975,11}0.401991 < \beta_1 < 4.08791 + t_{0.975,11}0.401991$$

$$4.08791 - (2.2010)(0.401991) < \beta_1 < 4.08791 + (2.2010)(0.01991)$$

$$4.08791 - 0.884782 < \beta_1 < 4.08791 + 0.884782$$

$$3.203128 < \beta_1 < 4.972692 \text{, con 95\% de confianza}$$

Por lo tanto, la pendiente de la población está entre 3.20 y 4.97 con 95% de confianza.

$$7.05495 - t_{0.975,11}0.887572 < \beta_0 < 7.05495 + t_{0.975,11}0.887572$$

$$7.05495 - (2.2010)(0.887572) < \beta_0 < 7.05495 + (2.2010)(0.887572)$$

$$7.05495 - 1.953546 < \beta_0 < 7.05495 + 1.953546$$

$$5.101404 < \beta_0 < 9.008496 \text{, con 95\% de confianza}$$

Por lo que la ordenada al origen de la población está entre 5.10 y 9.01 con 95% de confianza.

De la segunda tabla (Forecasts), se tienen los intervalos siguientes para la media y para la predicción:

$$15.9658 < \mu_{y/x} < 17.7661 \text{, con 95\% de confianza}$$

$$13.7491 < \hat{Y} < 19.9828 \text{, con 95\% de confianza}$$

CAPÍTULO 2

ANÁLISIS DE REGRESIÓN NO LINEAL

2.1 ANÁLISIS DE REGRESIÓN NO LINEAL

A partir del diagrama de dispersión, por datos previos, por conocimientos teóricos o por un análisis de residuos, se puede conocer que la relación entre las dos variables es no lineal y puede representarse adecuadamente por cierta función matemática curvilínea; por ejemplo, la tendencia general del crecimiento poblacional sigue un modelo exponencial positivo, el decaimiento radiactivo sigue un modelo exponencial negativo, la expectativa de vida en función del producto interno bruto es una función logarítmica, pues no se espera un promedio de vida cada vez más grande a medida que aumenta el producto interno bruto.

En algunos casos una función no lineal se puede expresar como una línea recta usando transformaciones adecuadas. Dichos modelos no lineales son llamados *linealmente transformables.*

2.2 REGRESIÓN EXPONENCIAL O SEMILOGARÍTMICA

Cuando se sospecha que la relación es de tipo exponencial, se propone una ecuación de regresión de la forma:

$$\hat{y} = cd^{x}$$

Como sugiere el nombre exponencial, la variable independiente *x* aparece en el exponente.

Por necesidad teórica y conveniencia práctica *se transforma* nuestra ecuación de regresión a otra, tomando logaritmos a ambos lados

$$Ln\ \hat{y} = Ln\ c + x\ Ln\ d \qquad [15]$$

Observe que sólo hay logaritmos en la variable *Y*. Si se hace

$$b_0 = Ln\ c$$
$$b_1 = Ln\ d$$

Entonces, la ecuación [15] se transforma en:

$$Ln\ \hat{y} = b_0 + b_1 x \qquad [16]$$

Ésta es una ecuación lineal en *Ln(y)* y *x*, que es una función semilogarítmica, de manera que si se llevan los puntos a papel semilogarítmico se obtiene una recta.

Para transformar de nuevo la ecuación [16] a la forma original sólo se necesita tomar la función inversa del logaritmo; es decir, la función exponencial de la base adecuada según haya sido la base de los logaritmos con que se esté trabajando; o bien si se toman logaritmos naturales (de base *e*), entonces:

$$c = e^{b_0}$$
$$d = e^{b_1}$$
$$\hat{y} = (e^{b_0})(e^{b_1})^x$$

Se pueden calcular los coeficientes b_0 y b_1 resolviendo el sistema de ecuaciones normales que se obtiene de las siguientes expresiones:

$$b_0 = ant\log\frac{\sum \log y \sum x^2 - \sum x \sum x * \log y}{n * \sum x^2 - (\sum x)^2}$$

$$b_1 = ant\log\frac{n * \sum x * \log y \sum x \sum \log y}{n * \sum x^2 - (\sum x)^2}$$

2.3 REGRESIÓN POTENCIAL O DOBLE LOGARÍTMICA

En ocasiones se tiene que una función potencial, como

$$\hat{y} = c\, x^b \qquad [17]$$

puede representar la relación entre *y* y *x* en la muestra. Aquí se desea hallar *c* y *b*. Para determinar *c* y *b*, se toman logaritmo a ambos lados de la ecuación [17] y se obtiene:

$$Ln\ \hat{y} = Ln\ c + b\ Ln\ x \qquad [18]$$

Este resultado es una relación doble logarítmica, porque tanto la variable *y* como la variable *x* se expresan ahora en logaritmos.

Si se hace:

$$b_0 = Ln\ c$$
$$b_1 = b$$

entonces la ecuación [18] se transforma en:

$$Ln\ \hat{y} = b_0 + b_1\ Ln\ x \qquad [19]$$

que es una ecuación lineal en *Ln(y)* y *Ln(x)*, la cual es una función doble logarítmica, de manera que si se llevan los puntos a papel *log - log* se obtiene una recta.

Por lo tanto, para hacer los cálculos primero se toman logaritmos a *x* y a *y,* y después se procede como en la regresión lineal.

Para transformar de nuevo la ecuación [19] a la forma original sólo se necesita tomar la función inversa del logaritmo, esto es la función exponencial de la base adecuada según haya sido la base de los logaritmos con que se esté trabajando; esto es, si se toman logaritmos naturales (de base *e*), entonces:

$$c = e^{b_0}$$

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

2.4 OBSERVACIONES ACERCA DE LA REGRESIÓN NO LINEAL

Como se expuso al tratar las regresiones curvilíneas, generalmente se hace una transformación inicial de datos por medio de logaritmos de tal forma que la relación entre las variables transformadas sea lineal. Las transformaciones logarítmicas no son las únicas, pudiera darse el caso de que las transformaciones recíprocas ($U = 1/X$ o $W = 1/Y$ o ambas) fuesen más adecuadas.

Las transformaciones en regresión lineal se hacen principalmente por dos razones, una de ellas es que este procedimiento puede facilitar los cálculos, la otra es que las transformaciones en regresión lineal son convenientes porque la teoría estadística está especialmente bien desarrollada bajo el supuesto de linealidad.

El análisis curvilíneo, al igual que el lineal, se realiza descomponiendo la variación por la variable o las variables independientes y el error. El método de mínimos cuadrados tiene la función de minimizar la suma de cuadrados del error y maximizar la suma de cuadrados debida a la regresión.

El grado de ajuste, en cada caso es medido por r^2, el coeficiente de determinación que es la razón de la suma de cuadrados explicada por la regresión entre la suma de cuadrados total.

La siguiente tabla muestra otras funciones linealizables, con sus transformaciones y las formas lineales resultantes y la Fig. 22 su representación gráfica.

Figura	Función linealizable	Transformación	Forma lineal
22 a, b	$Y = \beta_0 X^{\beta_1}$	$Y' = \log Y, X' = \log X$	$Y' = \log \beta_0 + \beta_1 X'$
22 c, d	$Y = \beta_0 e^{\beta_1 x}$	$Y' = \ln Y$	$Y' = \ln \beta_0 + \beta_1 X$
22 e, f	$Y = \beta_0 + \beta_1 \log X$	$X' = \log X$	$Y' = \beta_0 + \beta_1 X'$
22 g, h	$Y = \frac{X}{\beta_0 X - \beta_1}; \frac{1}{Y} = \frac{\beta_0 X - \beta_1}{X} = \beta_0 - \beta_1 \frac{1}{X}$	$Y' = \frac{1}{Y}, \quad X' = \frac{1}{X}$	$Y' = \beta_0 - \beta_1 X'$

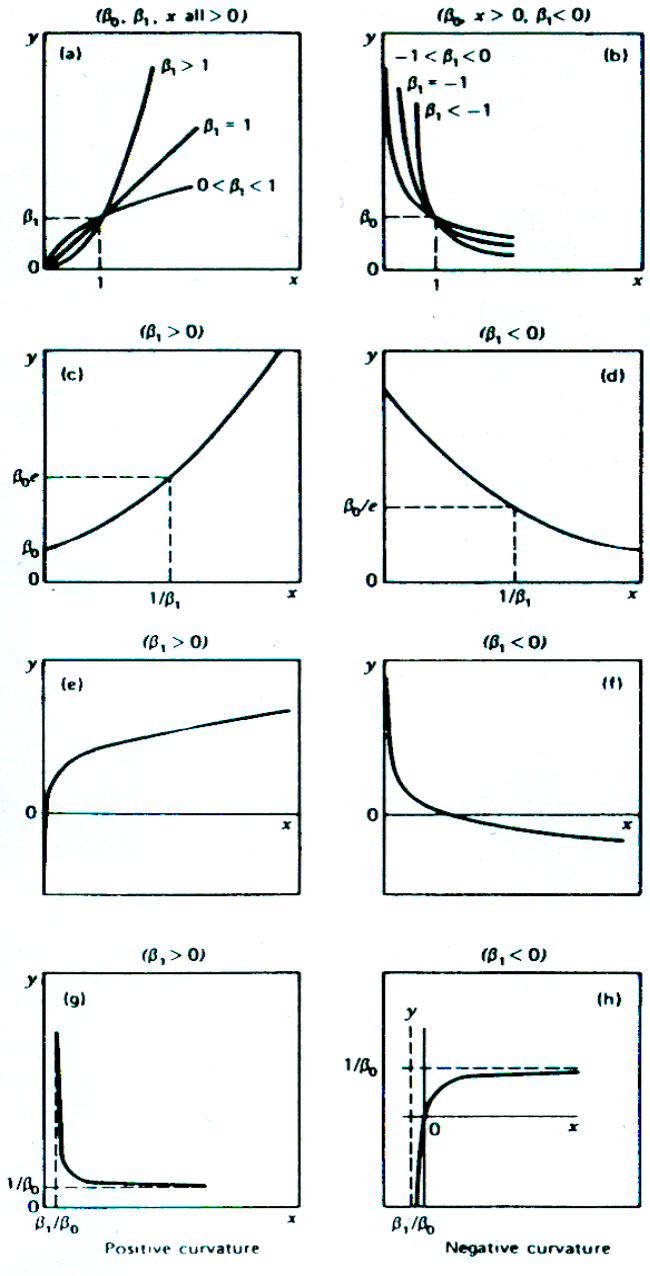


Fig. 22 Funciones linealizables (tomadas de Daniel and Wood, 1980)

Ejemplo 2: Se sabe que los grandes depósitos de agua tienen un efecto regulador de la temperatura de las masas de tierra que los rodean. En una noche fría en Florida central se registraron las temperaturas en puntos equidistantes de un gran lago. Los datos que se obtuvieron aparecen en la siguiente tabla. El modelo sugerido para estos datos es

$$\hat{y} = b_0 * e^{b_1 x}$$

Convertir el modelo dado a lineal y estimar los parámetros mediante el método de mínimos cuadrados. Además, encontrar los intervalos de 90% de confianza para β_0. (Ejemplo tomado de Wackerly, et al, 2002).

Sitio (x)	Temperatura (y)
1	37.00°F
2	36.25°F
3	35.41°F
4	34.92°F
5	34.52°F
6	34.45°F
7	34.40°F
8	34.00°F
9	33.62°F
10	33.90°F

Solución: En primer lugar linealizar el modelo, tomando logaritmos a ambos lados:

$$Ln\,\hat{y} = Ln\,b_0 + b_1 x$$

En segundo término introducir los datos en el modo de regresión lineal de una calculadora donde las variables son x y $Ln\,y$

x	$y^* = Ln\,y$
1	3.610918
2	3.590439
3	3.566994
4	3.553060
5	3.541539
6	3.539509
7	3.538057
8	3.526361
9	3.515121
10	3.523415

Los valores de los estimadores de los parámetros de regresión lineal son:

$$b_0 = 3.602707 \quad \text{y} \quad b_1 = -0.009484618$$

Por lo tanto, el modelo linealizado es:

$$Ln\,\hat{y} = 3.602707 + (-0.009484618)x$$

$$Ln\,\hat{y} = 3.602707 - 0.009484618x$$

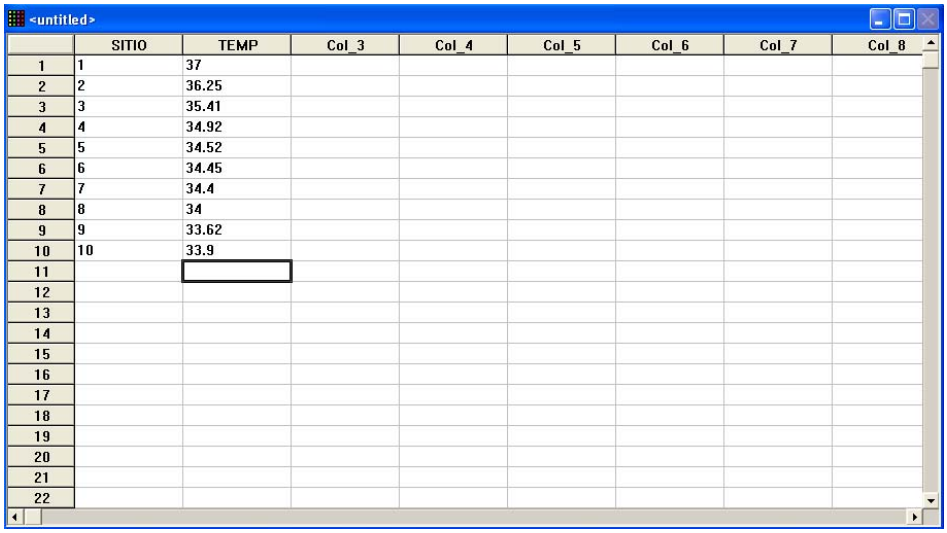
El modelo exponencial es:

$$\hat{y} = (e^{3.602707})(e^{-0.0094846})^x$$

$$\hat{y} = (36.697423)(0.9906)^x$$

Para hacerlo en STATGRAPHICS, los pasos a seguir son:

1. Crear el archivo de datos con dos columnas, una para la variable independiente (sitio) y otra para la variable dependiente (temp).



	SITIO	TEMP	Col_3	Col_4	Col_5	Col_6	Col_7	Col_8
1	1	37						
2	2	36.25						
3	3	35.41						
4	4	34.92						
5	5	34.52						
6	6	34.45						
7	7	34.4						
8	8	34						
9	9	33.62						
10	10	33.9						
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								
22								

Fig. 23 Datos Del ejemplo 2 en STATGRAPHICS (Tomados de Wackerly, 2002)

2. Del menú seguir la secuencia

Relate → Simple Regresión

3. En el diálogo que aparece colocar en su lugar la variable dependiente (temp) y la independiente (sitio).

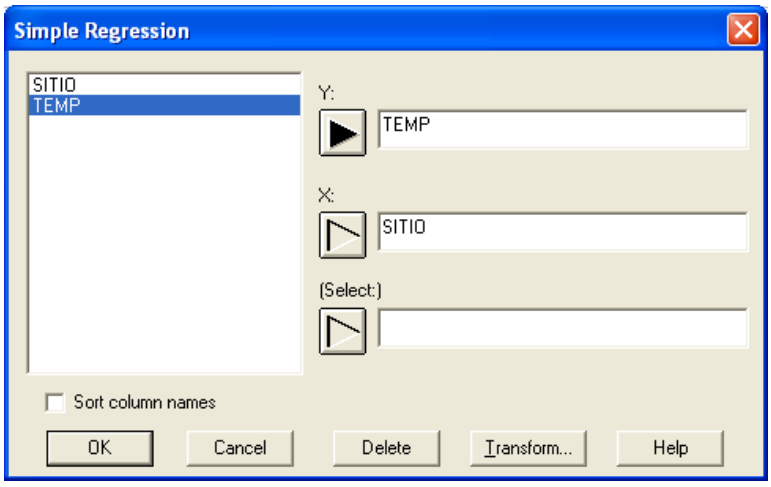


Fig. 24 Datos Del ejemplo 2 en STATGRAPHICS (Tomados De Wackerly, 2002)

4. Dar OK y llamar a las opciones tabulares y gráficas de esta opción.

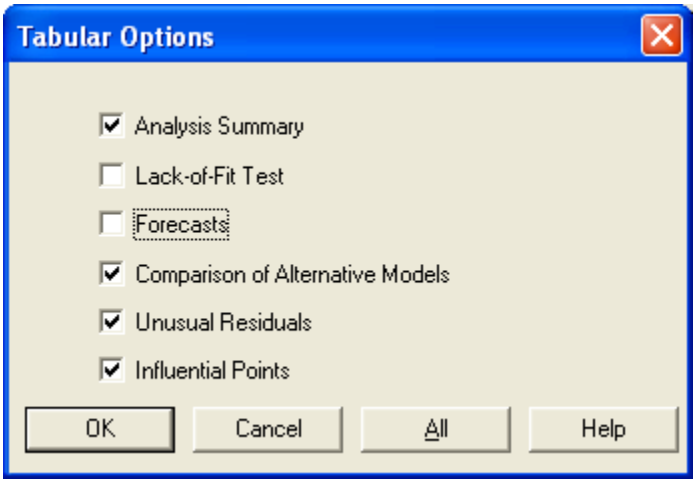


Fig. 25 Opciones Tabulares

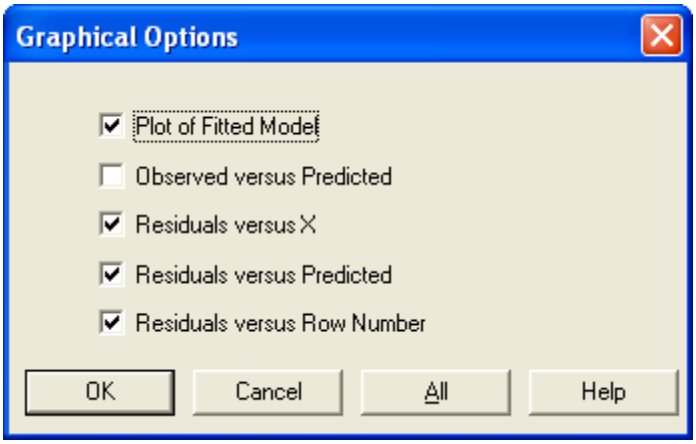


Fig. 26 Opciones Gráficas

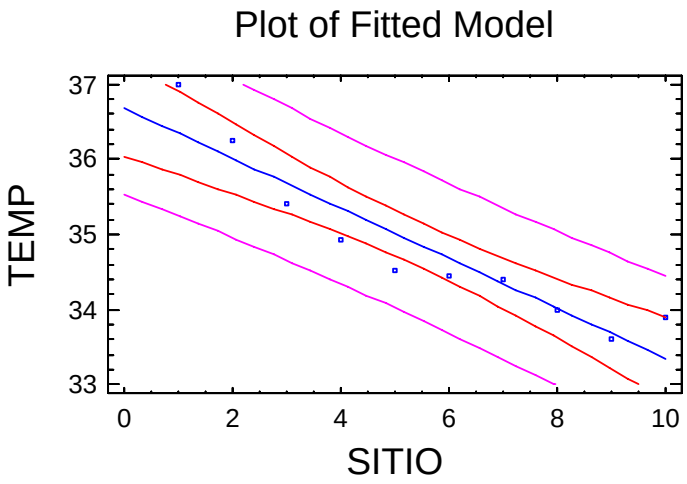
RESULTADOS

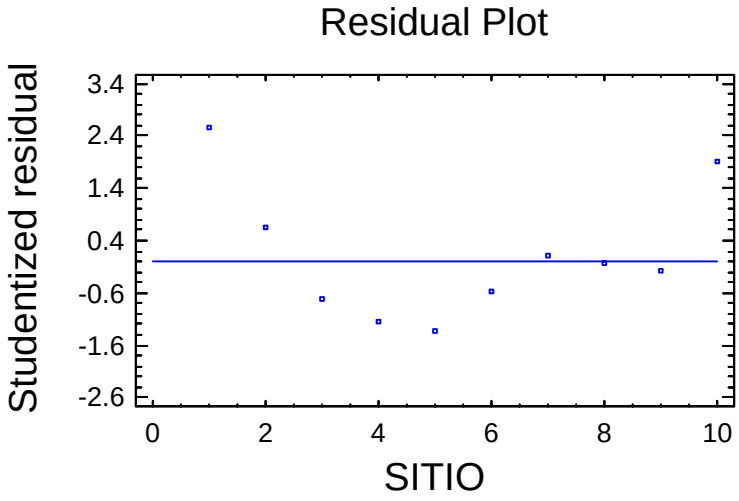
Simple Regression - TEMP vs. SITIO

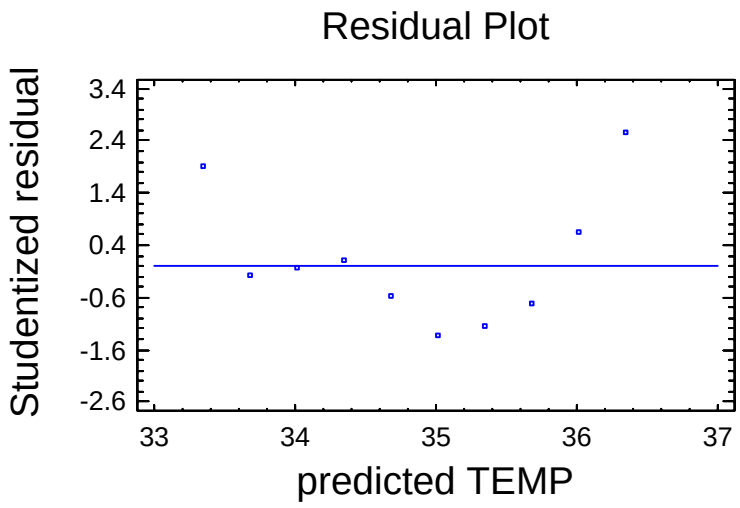
```
Regression Analysis - Linear model: Y = a + b*X
------------------------------------------------------------------------------
Dependent variable: TEMP
Independent variable: SITIO
------------------------------------------------------------------------------
                             Standard          T
Parameter       Estimate         Error       Statistic        P-Value
------------------------------------------------------------------------------
Intercept          36.68      0.280299         130.86         0.0000
Slope          -0.333273     0.0451743       -7.37748         0.0001
------------------------------------------------------------------------------


                          Analysis of Variance
------------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
------------------------------------------------------------------------------
Model                     9.16333      1      9.16333      54.43       0.0001
Residual                  1.34688      8      0.16836
------------------------------------------------------------------------------
Total (Corr.)             10.5102      9

Correlation Coefficient = -0.933729
R-squared = 87.1851 percent
R-squared (adjusted for d.f.) = 85.5832 percent
Standard Error of Est. = 0.410316
Mean absolute error = 0.299091
Durbin-Watson statistic = 0.73633 (P=0.0019)
Lag 1 residual autocorrelation = 0.359994

The StatAdvisor
---------------
   The output shows the results of fitting a linear model to describe the relationship between
TEMP and SITIO. The equation of the fitted model is

   TEMP = 36.68 - 0.333273*SITIO

Since the P-value in the ANOVA table is less than 0.01, there is a statistically significant
relationship between TEMP and SITIO at the 99% confidence level.

   The R-Squared statistic indicates that the model as fitted explains 87.1851% of the
```

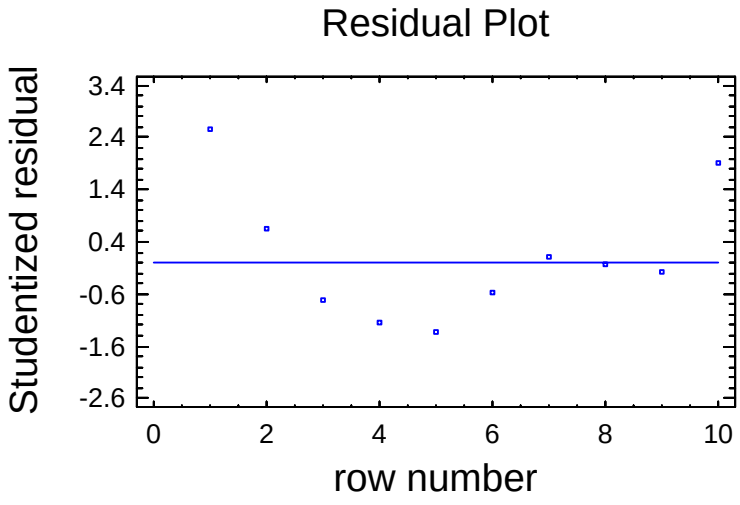
variability in TEMP. The correlation coefficient equals -0.933729, indicating a relatively strong relationship between the variables. The standard error of the estimate shows the standard deviation of the residuals to be 0.410316. This value can be used to construct prediction limits for new observations by selecting the Forecasts option from the text menu.

The mean absolute error (MAE) of 0.299091 is the average value of the residuals. The Durbin-Watson (DW) statistic tests the residuals to determine if there is any significant correlation based on the order in which they occur in your data file. Since the P-value is less than 0.05, there is an indication of possible serial correlation. Plot the residuals versus row order to see if there is any pattern which can be seen.









Unusual Residuals

Row	X	Y	Predicted Y	Residual	Studentized Residual
1	1.0	37.0	36.3467	0.653273	2.56

The StatAdvisor

The table of unusual residuals lists all observations which have Studentized residuals greater than 2.0 in absolute value. Studentized residuals measure how many standard deviations each observed value of TEMP deviates from a model fitted using all of the data except that observation. In this case, there is one Studentized residual greater than 2.0, but none greater than 3.0.

Influential Points

Row	X	Y	Predicted Y	Studentized Residual	Leverage

```
Average leverage of single data point = 0.2

The StatAdvisor
---------------
   The table of influential data points lists all observations which have leverage values greater
than 3 times that of an average data point.  Leverage is a statistic which measures how
influential each observation is in determining the coefficients of the estimated model. In this
case, an average data point would have a leverage value equal to 0.2. There are no data points
with more than 3 times the average leverage.
```

INTERPRETACIÓN

Es fácil observar que el modelo lineal no se ajusta, pues los puntos tienen una tendencia curva, tanto en el diagrama de dispersión como en los gráficos de residuos, por otra parte, se ve claramente que existe un residuo studentizado muy grande de 2.56, lo cual es una clara violación a los supuestos de normalidad, de homocedasticidad y de residuos fuera del intervalo [-2,2].

La siguiente tabla de comparación de los modelos se puede ver que el modelo logarítmico en x es el que tiene los coeficientes de correlación y de determinación más altos. Sin embargo, el sugerido por Wackerly fue el exponencial negativo, por lo que se va a utilizar el exponencial, aunque éste no tenga el coeficiente de determinación más alto.

```
Comparison of Alternative Models
-------------------------------------------------
Model                   Correlation       R-Squared
-------------------------------------------------
Logarithmic-X              -0.9893          97.87%
Multiplicative             -0.9890          97.81%
Square root-X              -0.9722          94.52%
Reciprocal-Y                0.9407          88.49%
Reciprocal-X                0.9375          87.89%
Exponential                -0.9373          87.85%
Square root-Y              -0.9355          87.52%
Linear                     -0.9337          87.19%
S-curve                     0.9332          87.09%
Double reciprocal          -0.9288          86.26%
Logistic                          <no fit>
Log probit                        <no fit>
-------------------------------------------------

The StatAdvisor
---------------
   This table shows the results of fitting several curvilinear models to the data.  Of the models
fitted, the logarithmic-X model yields the highest R-Squared value with 97.8735%. This is
10.6885% higher than the currently selected linear model. To change models, select the
Analysis Options dialog box.
```

5. **Sobre la pantalla de la regresión lineal simple, apretamos botón derecho y damos clic en ANALYSIS OPTIONS, al aparecer la tabla de los diferentes modelos, damos clic en el modelo EXPONENCIAL, luego OK.**

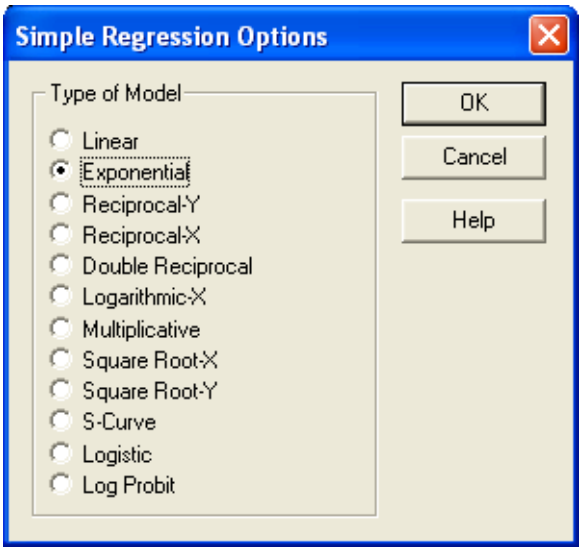


Fig. 27 Analysis Options

Simple Regression - TEMP vs. SITIO

```
Regression Analysis - Exponential model: Y = exp(a + b*X)
-----------------------------------------------------------------------------
Dependent variable: TEMP
Independent variable: SITIO
-----------------------------------------------------------------------------
                              Standard          T
Parameter       Estimate        Error       Statistic        P-Value
-----------------------------------------------------------------------------
Intercept        3.60271     0.00773801        465.586         0.0000
Slope        -0.00948462     0.00124709       -7.60538         0.0001
-----------------------------------------------------------------------------


                          Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                  0.00742153      1   0.00742153      57.84       0.0001
Residual               0.00102646      8  0.000128307
-----------------------------------------------------------------------------
Total (Corr.)          0.00844799      9

Correlation Coefficient = -0.937281
R-squared = 87.8497 percent
R-squared (adjusted for d.f.) = 86.3309 percent
Standard Error of Est. = 0.0113273
Mean absolute error = 0.00833894
Durbin-Watson statistic = 0.763638 (P=0.0024)
Lag 1 residual autocorrelation = 0.347791

The StatAdvisor
---------------
   The output shows the results of fitting a exponential model to describe the relationship
between TEMP and SITIO. The equation of the fitted model is

   TEMP = exp(3.60271 - 0.00948462*SITIO)

Since the P-value in the ANOVA table is less than 0.01, there is a statistically significant
relationship between TEMP and SITIO at the 99% confidence level.

   The R-Squared statistic indicates that the model as fitted explains 87.8497% of the
```

variability in TEMP after transforming to a logarithmic scale to linearize the model. The correlation coefficient equals -0.937281, indicating a relatively strong relationship between the variables. The standard error of the estimate shows the standard deviation of the residuals to be 0.0113273. This value can be used to construct prediction limits for new observations by selecting the Forecasts option from the text menu.

The mean absolute error (MAE) of 0.00833894 is the average value of the residuals. The Durbin-Watson (DW) statistic tests the residuals to determine if there is any significant correlation based on the order in which they occur in your data file. Since the P-value is less than 0.05, there is an indication of possible serial correlation. Plot the residuals versus row order to see if there is any pattern which can be seen.

The StatAdvisor

This table shows the results of fitting several curvilinear models to the data. Of the models fitted, the logarithmic-X model yields the highest R-Squared value with 97.8735%. This is 10.0239% higher than the currently selected exponential model. To change models, select the Analysis Options dialog box.

Unusual Residuals

Row	X	Y	Predicted Y	Residual	Studentized Residual
1	1.0	37.0	36.351	0.648989	2.47

The StatAdvisor

The table of unusual residuals lists all observations which have Studentized residuals greater than 2.0 in absolute value. Studentized residuals measure how many standard deviations each observed value of TEMP deviates from a model fitted using all of the data except that observation. In this case, there is one Studentized residual greater than 2.0, but none greater than 3.0.

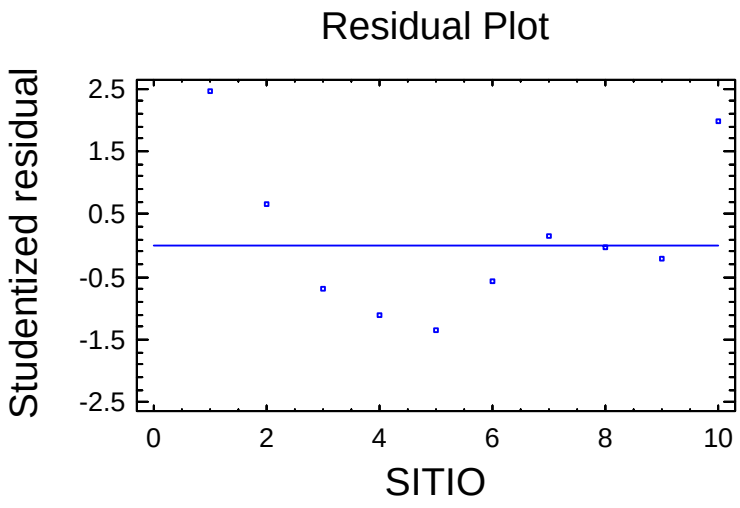
Influential Points

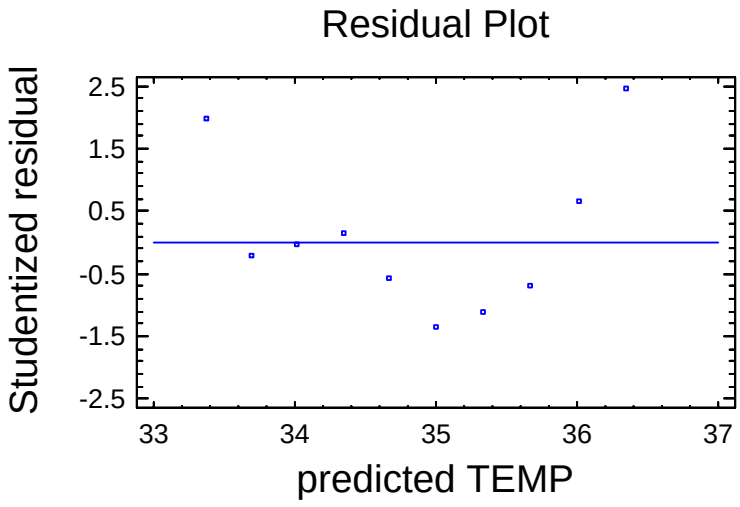
Row	X	Y	Predicted Y	Studentized Residual	Leverage

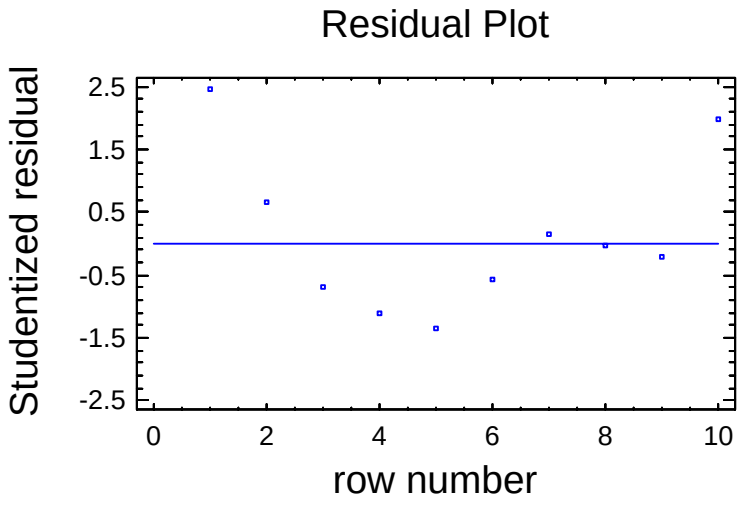
Average leverage of single data point = 0.2

The StatAdvisor

The table of influential data points lists all observations which have leverage values greater than 3 times that of an average data point. Leverage is a statistic which measures how influential each observation is in determining the coefficients of the estimated model. In this case, an average data point would have a leverage value equal to 0.2. There are no data points with more than 3 times the average leverage.


Residual Plot
Studentized residual
2.5
1.5
0.5
-0.5
-1.5
-2.5
0
2
4
6
8
10
SITIO


Residual Plot
Studentized residual
2.5
1.5
0.5
-0.5
-1.5
-2.5
33
34
35
36
37
predicted TEMP


Residual Plot
Studentized residual
2.5
1.5
0.5
-0.5
-1.5
-2.5
0
2
4
6
8
10
row number

INTERPRETACIÓN

El modelo exponencial que se obtiene es el mismo que se obtuvo con la calculadora. Es decir:

$y = \exp(a + bx)$ **o equivalentemente** $y = e^{a+bx} = e^{a} * (e^{b})^{x}$

Con valores de:

a = 3.60271

b **= -0.00948462**

$$y = exp(3.60271 + 0.00948462\,x)$$

Es decir:

$$y = e^{3.60271+0.00948462x} = e^{3.60271} * e^{0.00948462x} = 36.697423 * 1.009530^{x}$$

El intervalo del 90% para β_0 **se obtiene usando el estimador** b_0**, su error estándar y el valor t-student de tablas con 8 grados de libertad.**

$$3.602707 \mp 1.8595(0.00773801)$$

$$3.602707 - 0.014389 < Ln\,\beta_0 < 3.602707 + 0.014389$$

$$3.588318 < Ln\,\beta_0 < 3.617095$$

$$36.173171 < \beta_0 < 37.229273$$

Es importante hacer notar que en este ejemplo, se utilizó el modelo exponencial, aunque con el STATGRAPHICS se pudo observar que el modelo logarítmico tiene un coeficiente de determinación mayor. A continuación se presenta este modelo.

Sobre la pantalla de la regresión lineal simple, apretamos botón derecho y damos clic en ANALYSIS OPTIONS, al aparecer la tabla con los diferentes modelos, damos clic en el modelo LOGARITHMIC-X, luego OK.

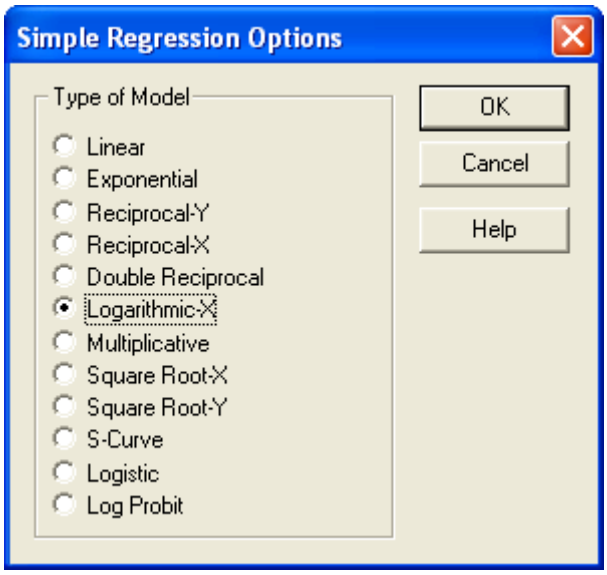


Fig. 28 Analysis Options

RESULTADOS

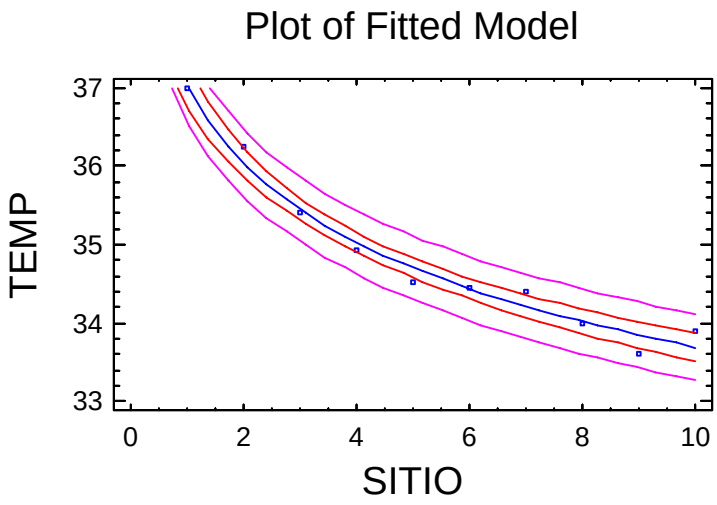
Simple Regression - TEMP vs. SITIO

```
Regression Analysis - Logarithmic-X model: Y = a + b*ln(X)
-----------------------------------------------------------------------------
Dependent variable: TEMP
Independent variable: SITIO
-----------------------------------------------------------------------------
                              Standard          T
Parameter       Estimate        Error       Statistic        P-Value
-----------------------------------------------------------------------------
Intercept        37.0499       0.126386        293.149         0.0000
Slope           -1.45848      0.0760062       -19.1889         0.0000
-----------------------------------------------------------------------------


                          Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                      10.2867      1      10.2867     368.21       0.0000
Residual                  0.223495      8    0.0279368
-----------------------------------------------------------------------------
Total (Corr.)              10.5102      9

Correlation Coefficient = -0.989311
R-squared = 97.8735 percent
R-squared (adjusted for d.f.) = 97.6077 percent
Standard Error of Est. = 0.167143
Mean absolute error = 0.124147
Durbin-Watson statistic = 2.16076 (P=0.2474)
Lag 1 residual autocorrelation = -0.183051

The StatAdvisor
---------------
   The output shows the results of fitting a logarithmic-X model to describe the relationship
between TEMP and SITIO. The equation of the fitted model is

   TEMP = 37.0499 - 1.45848*ln(SITIO)
```

Since the P-value in the ANOVA table is less than 0.01, there is a statistically significant relationship between TEMP and SITIO at the 99% confidence level.

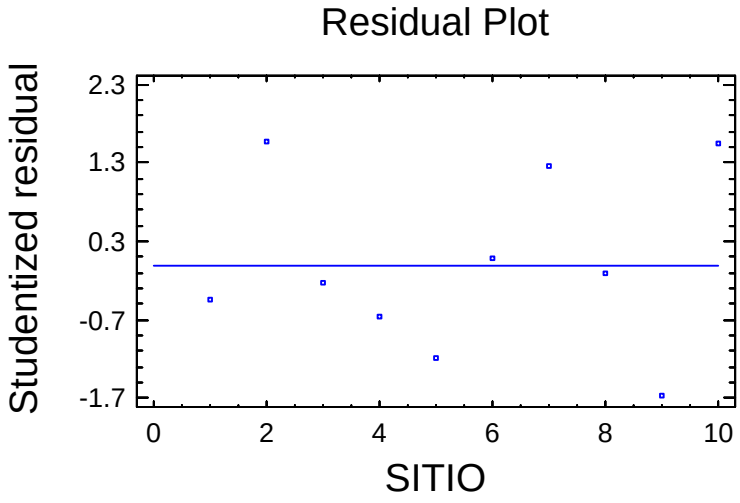
The R-Squared statistic indicates that the model as fitted explains 97.8735% of the variability in TEMP. The correlation coefficient equals -0.989311, indicating a relatively strong relationship between the variables. The standard error of the estimate shows the standard deviation of the residuals to be 0.167143. This value can be used to construct prediction limits for new observations by selecting the Forecasts option from the text menu.

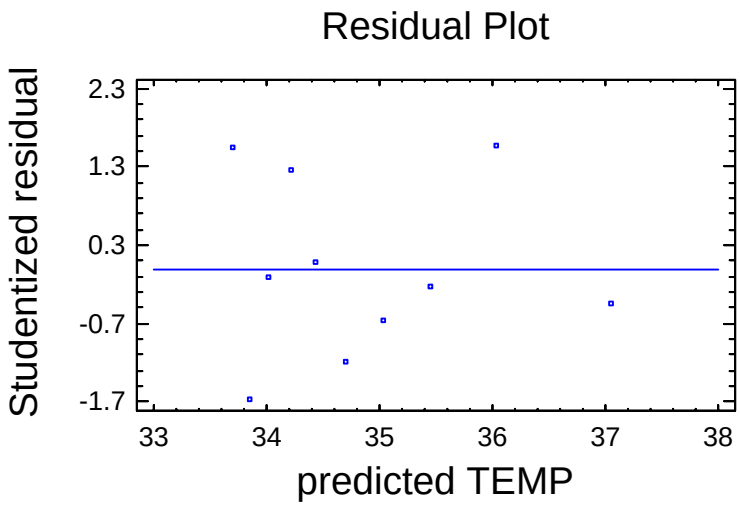
The mean absolute error (MAE) of 0.124147 is the average value of the residuals. The Durbin-Watson (DW) statistic tests the residuals to determine if there is any significant correlation based on the order in which they occur in your data file. Since the P-value is greater than 0.05, there is no indication of serial autocorrelation in the residuals.

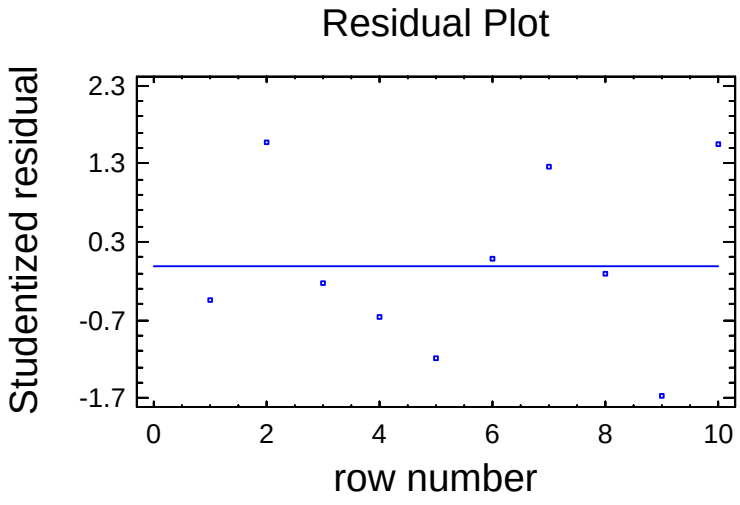


The StatAdvisor

This table shows the results of fitting several curvilinear models to the data. Of the models fitted, the logarithmic-X model yields the highest R-Squared value with 97.8735%. This is the currently selected model.







Unusual Residuals

Row	X	Y	Predicted Y	Residual	Studentized Residual

The StatAdvisor

The table of unusual residuals lists all observations which have Studentized residuals greater than 2.0 in absolute value. Studentized residuals measure how many standard deviations each observed value of TEMP deviates from a model fitted using all of the data except that observation. In this case, there are no Studentized residuals greater than 2.0.

Influential Points

Row	X	Y	Predicted Y	Studentized Residual	Leverage

Average leverage of single data point = 0.2

```
The StatAdvisor
---------------
   The table of influential data points lists all observations which have leverage values greater
than 3 times that of an average data point.  Leverage is a statistic which measures how
influential each observation is in determining the coefficients of the estimated model. In this
case, an average data point would have a leverage value equal to 0.2. There are no data points
with more than 3 times the average leverage.
```

INTERPRETACIÓN

Se ve claramente en el diagrama de dispersión que este modelo sí se ajusta, además las gráficas de residuos no tienen ninguna tendencia y no existen residuos studentizados que en valor absoluto sean mayores que 2. Así que el modelo logarítmico es el mejor, esto es:

$$Temp = 37.0499 - 1.45848 * Ln(sitio)$$

La estimación de β_0 se realiza de la misma forma, sustituyendo el valor estimado b_0, su eror estándar y el valor t-student de tablas con 8 grados de libertad.

$$37.0499 \mp 1.8595(0.126386)$$

$$37.0499 - 0.235015 < \beta_0 < 37.0499 + 0.235015$$

$$36.814885 < \beta_0 < 37.284915$$

CAPÍTULO 3

ANÁLISIS DE REGRESIÓN LINEAL MÚLTIPLE Y POLINÓMICA

3.1 REGRESIÓN LINEAL MÚLTIPLE

En la mayoría de los casos de la vida real, para poder predecir un comportamiento o fenómeno son necesarias más de una variable independiente o explicativa, por lo tanto, se tiene una función lineal múltiple.

Los modelos de regresión estudiados en las secciones anteriores son modelos donde sólo existe una variable independiente X, ahora se estudiará el modelo de regresión múltiple; es decir, donde se tiene una variable dependiente Y y dos o más independientes X_1, X_2,..., X_k así en vez de obtener una función lineal en dos variables (línea recta), tendremos una función lineal en tres o más variables.

Las variables independientes numéricas tales como área, presión, volumen, longitud, peso, etc. podrán ser introducidas de manera directa a las ecuaciones de regresión para obtener el modelo de predicción. Otros tipos de variables, que no estén expresadas de manera numérica, tales como tiempos transcurridos o fechas, deberán ser transformadas a una expresión numérica y de esta manera involucrarlas en el análisis.

$$Y = f(X_1, X_2, \cdots, X_k) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon_i$$

Este modelo describe un hiperplano en un espacio k–dimensional de las variables independientes X_i. Los parámetros β_j representan los cambios esperados en la variable respuesta Y por cambio unitario en X_i, cuando todas las variables independientes restantes X_j ($i \neq j$) se mantienen constantes. Por esta razón los parámetros β_j (j= 0, 1, 2, ..., k) son frecuentemente llamados *coeficientes de regresión parcial.* En regresión lineal triple todas las medias deben encontrarse en un plano, el punto de intersección del plano con el eje Y es la constante de regresión β_0 .

Comparando la regresión simple contra la múltiple se tiene que:

1. Es más difícil la elección del mejor modelo, ya que casi siempre hay varias opciones razonables.

2. Se dificulta visualizar el modelo, por la dificultad de “pintar” más de tres dimensiones.

3. Requiere cálculos complejos, generalmente se realiza con recursos computacionales y software especializado.

4. Además de los supuestos arriba mencionados para la regresión lineal simple, en la regresión lineal múltiple se debe cumplir la **no colinealiadad** de las variables explicativas.

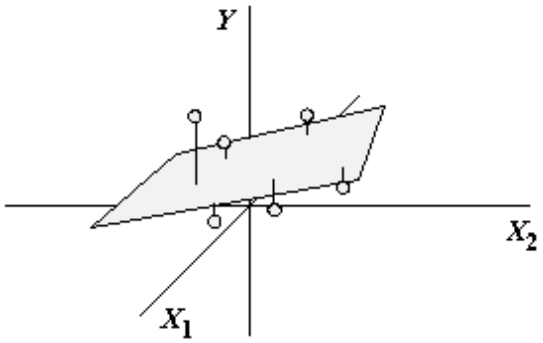


Fig. 28 Ajuste de un plano lineal con dos variables independientes

3.2 MÍNIMOS CUADRADOS

Al igual que en la regresión lineal simple, se puede trabajar el método de mínimos cuadrados**.** Para esto:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon_i$$

Este es el modelo de regresión lineal múltiple con k **regresores o variables explicativas y** $k+1$ **parámetros.**

Donde: $\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)$

Con base en los datos muestrales

$$\varepsilon_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k)$$

Al elemento de la derecha se le conoce como residual o residuo y refleja la desviación de los datos observados con respecto al plano ajustado.

Suma de cuadrados, elevando al cuadrado y sumando los elementos de la ecuación anterior.

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left[y_i - (b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k) \right]^2$$

El método consiste en encontrar los valores $b_0, b_1, b_2, \ldots$ llamados estimadores de mínimos cuadrados, para los cuales la suma de cuadrados es mínima. De tal manera que, se pueda construir la siguiente tabla de ANOVA.

Tabla de ANOVA para las hipótesis:
Ho: $\beta_i = 0$
Ha: Al menos un $\beta_i \neq 0$

Fuente de variación	**g.l.**	**SC**	**CM**	**F_C**	**F_t**
Regresión	k	$SC_{Reg} = SC_{Total} - SC_{Error}$	$CM_{Reg} = \frac{SC_{Reg}}{k}$	$\frac{CM_{Reg}}{CM_{Error}}$	$F_{1-\alpha,1,n-2}$
Error o Residual	$n-k-1$	$SC_{Error} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$CM_{Error} = \frac{SC_{Error}}{n-k-1}$	—	
Total	$n-1$	$SC_{Total} = \sum_{i=1}^{n}(y_i - \bar{y})^2$	—	—	

Donde los supuestos del análisis de Regresión se pueden resumir en la siguiente expresión.

$$\varepsilon \approx \mathrm{NI}(\mu_{Y/X_1,X_2,\cdots,X_k}, \sigma^2)$$

Los errores o residuales se distribuyen normal e independientemente con desviaciones al ajuste lineal (Media) igual a cero y varianza σ^2.

En forma más explícita, el modelo clásico de regresión lineal múltiple debe cumplir con los siguientes **supuestos**.

1. Las variables independientes $X_1, X_2, \ldots, X_k$ son fijas o no aleatorias.
2. Los errores ε_i tienen distribuciones normales con $\mu_{\varepsilon_i} = 0$.
3. La varianza de la regresión es constante e igual a la varianza de los errores ε_i
$$\sigma^2_{Y/X_1,\cdots,X_k} = \sigma^2_{\varepsilon_i} = \sigma^2$$
4. Los errores son estadísticamente independientes; es decir, los residuos (ε_i) no están correlacionados $[Cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j]$.
5. Pueden existir relaciones significativas de dependencia lineal entre dos cualesquiera de las variables independientes, pero su correlación debe ser pequeña (colinealidad).
6. El número de observaciones de la muestra debe superar al número de coeficientes de regresión que se deben estimar para garantizar que el número de grados de libertad sea diferente de cero.

Después de obtener la ecuación de regresión de la muestra, se debe establecer si los resultados de la muestra son estadísticamente significativos para poder usar la ecuación de regresión como instrumento de predicción.

Como ya se expuso esta inferencia se realiza con un análisis de varianza

3.3 COEFICIENTE CORRELACIÓN PARCIAL y PARCIAL MÚLTIPLE

La correlación es la medida de la fuerza de relación lineal entre dos variables, después de controlar los efectos de otras variables en el modelo; es decir, el grado de asociación entre Y y una variable explicativa, eliminando el efecto lineal de todas las otras variables explicativas. Mide la fuerza de la relación entre Y y una sola variable independiente, considerando la cantidad en que se reduce la variación explicada al incluir esta variable en la ecuación de regresión. Esta correlación se representa por:

$$R_{Y,X_1/X_2} \qquad R_{Y,X_1/X_2,X_3} \qquad R_{Y,(X_3,X_4,X_5)/X_1,X_2}$$

Expresiones que se leen:

$R_{Y,X_1/X_2}$ Correlación de las variables Y–X_1, cuando se tiene controlado el efecto de X_2 en un modelo. También se puede leer: correlación de Y–X_1, cuando X_2 ya está en el modelo.

$R_{Y,X_1/X_2,X_3}$ Correlación de las variables Y–X_1, cuando se tienen controlados los efectos de X_2 y X_3 en un modelo.

$R_{Y,(X_3,X_4,X_5)/X_1,X_2}$ Correlación de las variables X_3, X_4 y X_5 con Y, cuando se tienen controlados los efectos de X_1 y X_2 en un modelo.

3.4 COEFICIENTE DE DETERMINACIÓN MÚLTIPLE

Al igual que en la regresión simple, el coeficiente de correlación elevado al cuadrado es el coeficiente de determinación parcial.

$$
\begin{aligned}
R^2_{Y/(X_1,X_2,\cdots,X_k)} &= \frac{SC_{Reg}}{SC_{Total}} \\
&= \frac{SC_{total} - SC_{Error}}{SC_{Total}} \\
&= \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2 - \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \\
&= 1 - \frac{SC_{Error}}{SC_{Total}}
\end{aligned}
$$

Donde ***r* y *r*^{2} representan la correlación y determinación simple,** mientras que ***R* y *R*^{2} se utilizan para la correlación y determinación múltiple.**

Es común pensar en R^2 como una medida de la reducción de la variabilidad de Y, obtenida por usar las variables explicativas X_1, X_2, … , X_k. Como en regresión simple, se puede tener $0 < R^2 < 1$, sin embargo, un valor grande de R^2 no necesariamente implica que el modelo de regresión es bueno, ya que adicionar una variable explicativa al modelo siempre incrementará R^2 a pesar de que la variable explicativa contribuya muy poco al modelo. Lo cual puede llevar a interpretar, predecir o estimar mal la correlación múltiple.

Algunos analistas prefieren el uso del estadístico R^2 ajustada porque la R^2 ordinaria siempre incrementará (al menos no decrece) cuando se adiciona un término nuevo al modelo de regresión.

En los métodos de selección de variables y construcción de modelos R^2 ajustada se usa frecuentemente para proteger al modelo de un sobreajuste, como sería incluir variables innecesarias al modelo, la R^2 ajustada, penaliza al analista cuando esto ocurre. Esta medida es calculada rutinariamente por la mayoría del software de análisis estadístico.

Se define R^2 ajustada como R^2_{ajus}, reemplazando SC_{error} (suma de cuadrados del error) y SC_{total} (suma de cuadrados total) por sus correspondientes Cuadrados Medios

$$
R^2_{ajus} = 1 - \frac{\frac{SC_{Error}}{(n-k-1)}}{\frac{SC_{Total}}{(n-1)}} = 1 - \frac{n-1}{n-k-1}(1-R^2)
$$

Si R^2_{ajus} y R^2 difieren dramáticamente, entonces indica que el modelo ha sido sobreespecificado, esto es, hay términos que no contribuyen significativamente al ajuste.

3.5 CONTRASTES DE HIPÓTESIS DE LA REGRESIÓN LINEAL MÚLTIPLE

A menudo se desea probar que tan significantes son los parámetros del modelo de regresión, lo cual se logra al contrastar si dichos coeficientes son iguales a cero; las hipótesis son:

$$H_0 : \beta_0 = \beta_1 = \cdots = \beta_k = 0$$

$$H_a : \beta_i \neq 0$$

Rechazar *Ho* implica que al menos una de variables del modelo contribuye significativamente al ajuste. El parámetro para probar esta hipótesis es una generalización del utilizado en regresión lineal simple. La suma total de cuadrados (SC_{Total}) se descompone en la suma de cuadrados de regresión (SC_{Reg}) y en la sumas de cuadrados del error (SC_{Error}).

$$SC_{Total} = SC_{Reg} + SC_{Error}$$

Consecuentemente el valor de *F* estimado se obtiene de la ecuación:

$$F_C = \frac{\dfrac{SC_{Reg}}{k}}{\dfrac{SC_{Error}}{n-k-1}} = \frac{CM_{Reg}}{CM_{Error}}$$

Valor que se compara con una $F_{1-\frac{\alpha}{2},k,n-k-1}$ de tablas. La regla de decisión es: Se Rechaza *Ho,* si $F_C > F$ de tablas.

3.6 INTERVALOS DE CONFIANZA EN REGRESIÓN MÚLTIPLE

Los intervalos de confianza para los coeficientes individuales de regresión y para la media de la variable de respuesta, juegan un papel importante en la regresión múltiple.

Para construir intervalos de confianza para los coeficientes de regresión β_i, es necesario asumir que los errores ε_i, están normal e independientemente distribuidos con media cero y varianza σ^2. Por lo tanto las observaciones Y_i están normal e independientemente distribuidas con media $\beta_0 + \sum_{i=1}^{k} \beta_i X_{ij}$ y varianza σ^2. Entonces los estimadores por mínimos cuadrados b_i son una combinación lineal de las observaciones, resultando que b_i está normalmente distribuida con media del vector β y matriz de covarianza $\sigma^2(\mathbf{X'X})^{-1}$. Esto implica que la distribución marginal de algunos coeficientes de regresión b_i es normal con media β_i y varianza $\sigma^2 C_{ij}$, donde

$\sigma^2 = \frac{SC_{Error}}{n-k-1}$ (n = número de observaciones y k es el número de coeficientes estimados) y C_{ij} es el j-ésimo elemento diagonal de la matriz $(\mathbf{X'X})^{-1}$.

De esta manera, el intervalo de confianza para β_i es el siguiente:

$$\beta_i \pm t_{\alpha/2,n-k-1}\sqrt{\sigma^2 C_{ij}}$$

El intervalo de confianza para la media de Y, $E[Y]$ se calcula de la siguiente manera:

$$\hat{Y}_0 \pm t_{\alpha/2,n-k-1}\sqrt{\hat{\sigma}^2 x_0^{'}(X'X)^{-1}x_0}$$

donde x_0 es el vector de puntos particulares X.

3.7 AUTOCORRELACIÓN (NO INDEPENDENCIA DE LOS RESIDUOS)

El problema a estudiar ahora es aquel que se presenta cuando los términos de error en el modelo de regresión no son independientes.

La falta de independencia puede presentarse en datos de estudios de corte transversal o en datos en el tiempo. En el primer caso supóngase que se tiene interés en estudiar el consumo de familias en diferentes barrios o zonas geográficas. Se puede pensar que dentro de la misma zona o barrio los errores en el modelo de estimar el consumo estén correlacionados, debido por ejemplo al hecho de que los vecinos quieren mantener un mismo nivel de consumo. En este caso se habla de correlación espacial.

La presencia de autocorrelación en los errores tiene varios efectos en el procedimiento ordinario de regresión por mínimos cuadrados. Estos son resumidos a continuación:

- Los coeficientes de regresión por mínimos cuadrados ordinarios son más imparciales. Podemos decir que estos estimadores son ineficientes.

- Cuando los errores están positivamente autocorrelacionados, los cuadrados medios residuales subestiman seriamente a σ^2. Consecuentemente el error estándar del coeficiente de regresión puede ser también pequeño. Entonces los intervalos de confianza resultan más pequeños que lo que en realidad son y las pruebas de hipótesis de los coeficientes individuales de regresión pueden indicar que uno o más variables explicativas pueden contribuir significativamente al modelo cuando en realidad no es así. Generalmente la subestimación de σ^2 da al analista una falsa impresión de adecuación del modelo.

- Los intervalos de confianza y la pruebas de hipótesis basadas en las distribuciones *t* y *F* son, estrictamente hablando, no apropiadas.

Dos problemas generales de autocorrelación son:

- Si la autocorrelación se presenta porque una variable explicativa fue omitida y si esta variable explicativa se identifica y se incluye en el modelo, la aparente autocorrelación puede desaparecer. Si el problema de autocorrelación no puede ser resuelto incluyendo previamente factores omitidos entonces el analista puede regresar al modelo que específicamente incorpora la estructura de autocorrelación. Tales modelos usualmente requieren técnicas especiales de estimación de parámetros.
- Otra situación que ocurre frecuentemente sobre todo en negocios y economía es tener datos en series de tiempo

3.7.1 DETECTANDO LA PRESENCIA DE AUTOCORRELACIÓN

Los gráficos de residuos pueden ser usados para la detección de autocorrelación (Fig 31). La dispersión más significativa es el gráfico de residuos contra tiempo, que puede mostrar las siguientes tendencias:

- Autocorrelación positiva: cuando los puntos se encuentran predominantemente en el primer y tercer cuadrante, lo que significa que los residuos sucesivos tienden a tener el mismo signo.

- Autocorrelación negativa: cuando la mayor parte de los puntos están en el segundo y cuarto cuadrante, y por lo tanto los residuos consecutivos tienden a tener signos opuestos.

- Ausencia de autocorrelación: cuando los puntos se extienden sobre los cuatro cuadrantes.

Varias pruebas estadísticas pueden ser usadas para detectar la presencia de autocorrelación. La prueba desarrollada por Durbin y Watson es ampliamente utilizada.

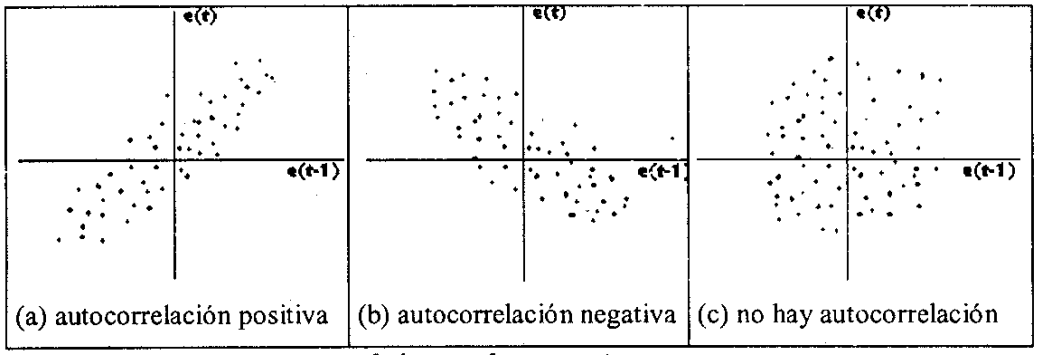


Fig. 31 Representación gráfica de los residuos sucesivos

3.7.2 PRUEBA DE DURBIN WATSON

Esta prueba es útil para detectar problemas de autocorrelación de primer orden en los residuos. El modelo más sencillo que relaciona los errores es el modelo lineal, en el que los errores poblacionales ε_t y ε_{t-1} tienen una correlación ρ. Una estimación de esta correlación estará dada por la correlación entre los residuos de mínimos cuadrados e_t y e_{t-1}.

Las hipótesis nula y alterna que se plantean en esta prueba son:

*H*o: los residuos no están correlacionados
Ha: existe autocorrelación de primer orden entre los residuos.

Estas hipótesis se pueden expresar como:

*H*o: $\rho = 0$
Ha: $\rho > 0$

El estadístico de la prueba está dado por

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

Donde los e_t, t =1, 2, …, n, son los residuos de un análisis de mínimos cuadrados ordinarios aplicados a los datos (Y_t, X_t). Desafortunadamente la distribución de *d*, depende de la matriz ***X***.

Sin embargo, Durbin y Watson muestran que *d* se encuentra entre 2 límites, *dL* y *dU*, tal que si *d* cae fuera de estos límites, se puede llegar a una conclusión respecto a las hipótesis planteadas. El procedimiento de decisión es el siguiente:

Si $d < d_L$	se rechaza *H*o *:* $\rho = 0$
Si $d > d_U$	no se rechaza *H*o *:* $\rho = 0$
Si $d_L \leq d \leq d_U$	La prueba es no concluyente

Es claro que pequeños valores de *d* implica que Ho*:* $\rho = 0$ puede ser rechazada porque la autocorrelación positiva indica que las condiciones de los errores sucesivos tienen magnitudes similares, y la diferencia entre los residuos $e_t - e_{t-1}$ será pequeña. Durbin y Watson sugieren varios procedimientos para resolver resultados no concluyentes.

Los valores de los límites d_L y d_U, para varios tamaños de muestra, número de variables explicativas y tres niveles de significación (error tipo I) se pueden consultar en tablas.

La autocorrelación negativa no se encuentra con frecuencia, sin embargo, si desea probar autocorrelación negativa, se puede usar 4 – *d*, donde *d* es definida como el estadístico de prueba.

Entonces el criterio de decisión para Ho*:* $\rho = 0$ vs Ha *:* $\rho \neq 0$ son los mismos usados en la prueba de autocorrelación positiva.

3.8 REGRESIÓN POLINÓMICA

Es frecuente encontrar que la relación no lineal más obvia entre dos variables es una en la que la variable dependiente *Y* puede ser aproximada por medio de un polinomio en la variable independiente *X*.

Una función de regresión polinómica de grado k (una variable explicativa) se define por la ecuación:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k + \varepsilon$$

Una función de regresión polinómica comúnmente empleada es la de segundo grado (cuadrática) que es una parábola que viene dada por:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2$$

Para hallar los estimadores de β_0, β_1 y β_2 se utiliza la ecuación de regresión lineal múltiple, pues haciendo los cambios de variable $X_1 = X$, $X_2 = X^2, \ldots, X_k = X^k$ y usando la fórmula $\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}(\mathbf{X}^t\mathbf{Y})$ para calcularlos, es como si se trabajara con regresión lineal múltiple.

3.9 DIAGNÓSTICO DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE Y MEDIDAS DE ADECUACIÓN DEL MODELO

La evaluación de la adecuación del modelo es parte importante de un problema de regresión múltiple. Algunas de las técnicas que aquí se presentan son extensiones de las usadas en regresión lineal simple.

3.10 GRÁFICAS DE RESIDUOS

Los residuos e_i en el modelo de regresión lineal múltiple juegan un papel importante en el juicio de la adecuación del modelo, justo como en la regresión lineal simple. Las gráficas de residuos utilizadas para la regresión lineal simple se pueden aplicar directamente en regresión múltiple.

Específicamente se obtienen las siguientes gráficas:

- Residuos en gráficos de normalidad
- Residuos contra cada variable explicativa X_i.
- Residuos contra $\hat{Y}_i$ ajustada.
- Residuos en secuencia de tiempo (si se conoce)

Estas gráficas también se usan para detectar desviaciones de la normalidad, casos extremos, desigualdad de varianzas y la equivocada función de una variable explicativa. En todas las gráficas pueden utilizarse residuos estandarizados o studentizados.

3.10.1 GRÁFICAS DE RESIDUOS CONTRA VARIABLES EXPLICATIVAS OMITIDAS POR EL MODELO

Si estas variables explicativas son candidatas que no fueron incluidos en el modelo, entonces graficando los residuos vs. los niveles de esta variable explicativa (asumiendo que éstas son conocidas) puede revelarse alguna dependencia de las respuesta *Y* en los factores omitidos. Si se muestra alguna estructura en esta gráfica, puede indicar que la incorporación de este factor puede mejorar el modelo.

3.10.2 GRÁFICAS DE REGRESIÓN X_j vs VARIABLE EXPLICATIVA X_i.

Esta gráfica se puede usar en el estudio de la relación entre variables explicativas y la disposición de los datos entre variables explicativas y la disposición de los datos en el espacio *X*. Considerando el despliegue de la gráfica de X_i vs X_j

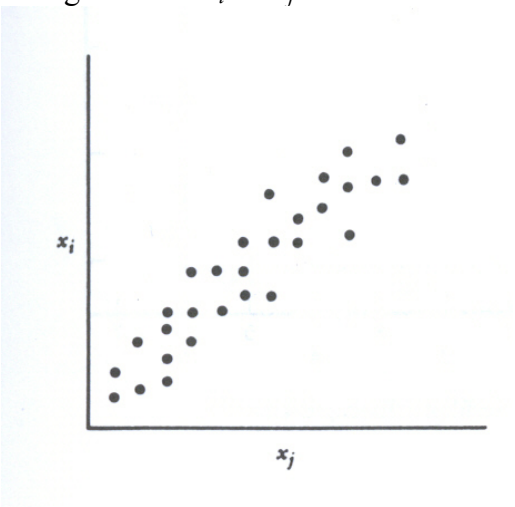


Fig. 29 Gráfica de X_i contra X_j

Indica que X_i y X_j están altamente correlacionadas. Consecuentemente esto puede dar pauta a incluir alguna de las variables en el modelo pero no las dos. Si dos o más variables están altamente correlacionadas, se dice que se presenta *multicolinealidad* en los datos. La multicolinealidad puede perturbar seriamente el ajuste de mínimos cuadrados y en alguna situación hacer el modelo de regresión casi inútil.

Las gráficas X_i vs X_j se pueden usar también para descubrir puntos que están aislados del resto de los datos y que pueden influenciar potencialmente las propiedades claves del modelo.

3.11 DIAGNÓSTICOS DE INFLUENCIA

Ocasionalmente se encuentra que un pequeño subconjunto de datos ejercen una influencia desproporcionada en el ajuste del modelo de regresión. Esto es, los parámetros estimados o las predicciones pueden depender más de la influencia de este subconjunto de puntos que de la mayoría de los datos. Será de importancia localizar estos puntos influyentes y evaluar su impacto en el modelo. Si estos puntos influyentes son "malos" valores (mal registro, o equivocaciones al transcribir), entonces pueden ser eliminados. Por otro lado, pueden ser datos no equivocados pero si éstos controlan propiedades claves del modelo, sería conveniente conocer la magnitud de influencia en el uso del modelo. ***Las siguientes medidas de influencia son las más usadas.***

3.11.1 PUNTOS DE INFLUENCIA (LEVERAGE POINTS)

La disposición de los puntos en el *X*-espacio es importante en la determinación de las propiedades del modelo. En particular, observaciones aisladas que potencialmente tienen una influencia desproporcionada en la estimación de parámetros, valores de predicción y el usual resumen estadístico. Daniel y Wood (1980) sugieren que la suma de cuadrados ponderada, que es la distancia desde el punto i-ésimo al centro del conjunto de datos es:

$$WSSD_i = \sum_{j=1}^{k} \left[\frac{\beta_j (X_{ij} = \bar{X}_j)}{\sqrt{CM_{Error}}} \right]^2, \qquad \text{i} = 1, 2, \ldots, n$$

y, este punto puede considerarse como aislado en el *X*-espacio.

El procedimiento general es que el total de puntos $WSSD_i$ se ordena de manera creciente y ocurre una concentración de puntos de donde se pueden identificar los valores grandes. Es difícil dar una pauta formal para identificar un valor "grande" de $WSSD_i$, generalmente si los valores aumentan suavemente del más pequeño al más grande, es probable que no haya puntos extremos. En cambio, saltos bruscos en la magnitud de $WSSD_i$ frecuentemente indican *que uno o más puntos extremos están presentes.*

3.11.2 DISTANCIA DE COOK

La distancia de Cook detecta si el i-ésimo caso es influyente y consiste en buscar la distancia entre los parámetros estimados si incluyen la observación i-ésima y si no la incluyen. Cada observación tiene su distancia y se considera significativa (influyente) si es mayor que 1, esto es, mide el cambio en la estimación de los coeficientes de regresión si el i-ésimo caso es eliminado. Si se remueve el caso con el valor Cook más grande, la estimación de los coeficientes puede cambiar más que por otro caso cualquiera. Para cada caso, Cook puede ser visto como una medida Euclidiana escalada entre dos vectores de valores ajustados, en el primero el caso es incluido, en el segundo el caso es excluido. Algebraicamente, esta medida puede escribirse como una función de residuos de influencia. Para cada caso:

$$C_i^2 = \frac{\sum (\hat{Y}_i - \hat{Y}_{j(i)})^2}{k * CM_{error}}$$

donde $\hat{Y}j_{(i)}$ son los valores de predicción cuando se ha eliminado una observación, k es el número de variables explicativas (X_1, X_2, ... ,X_k). C_i^2 es la distancia ponderada estandarizada ente los coeficientes de regresión obtenidos de los datos completos y los coeficientes de regresión obtenidos por la eliminación de la i-ésima observación. La estandarización está dada por $k*CM_{error}$.

Si un punto es influyente, esta eliminación causará grandes cambios y el valor de C_i^2 será grande. Un valor grande de C_i^2 indica que el punto es potencialmente influyente (leverage).

Se ha sugerido que los puntos de C_i^2 con valores más grandes que el 50% del valor de F con k y ($n - k$–1) grados de libertad se clasifican como puntos influyentes. Una operación práctica es clasificar los puntos de C_i^2 más grandes que 1 como influyentes. En vez de usar un corte rígido, se sugiere que todos los C_i^2 sean examinados. Un gráfico de todos los C_i^2 (unit plot) o un diagrama de tallo-hoja es un buen gráfico. Cuando los valores son más o menos los mismos esta acción no se considera.

3.11.3 DFFITS

Una medida similar a la distancia de Cook es la llamada DFFITS

$$DFFITS_{(i)} = \frac{\hat{Y}_i - \hat{Y}_{(i)i}}{CM_{error\,(i)}\sqrt{p_{ii}}} = \frac{\varepsilon_i \sqrt{p_{ii}}}{CM_{error\,(i)}\sqrt{1 - p_{ii}}}$$

donde p_{ii} son los elementos de la diagonal de la matriz $\mathbf{P}$ ($\mathbf{P} = \mathbf{X(X^tX)^{-1}X^t}$) y $CM_{error(i)}$ son los cuadrados medios residuales cuando se elimina la observación i.

Las DFFITS corresponden a las Ci (como opuesto a C_i^2) cuando la normalización es dada por $CM_{error(i)}$ en vez de CM_{error}. Los puntos con (DFFITS) más grandes, en valor absoluto, que $2\sqrt{\frac{k}{n-k}}$ son clasificados usualmente como puntos influyentes.

3.11.4 DFBETA(S)

Es una medida de influencia del i-ésimo caso en *cada* estimación de coeficientes de regresión separadamente. Esto es, como cambia cada coeficiente de regresión estimado b_j si el caso fuera excluido. Se puede calcular este estadístico para un dato, corriendo dos regresiones, una con el

dato y la otra sin él. Entonces, para cada coeficiente, encontramos la diferencia en el valor entre las dos estimaciones.

$$DFBETA_{j,i} = \frac{b_j - b_{j(i)}}{CM_{Error(i)} C_{ii}}$$

donde C_{ii} son los elementos de la diagonal de la matriz $(\mathbf{XX})^{-1}$ y $b_{j(i)}$ es la estimación del *j*-ésimo coeficiente de regresión calculado sin la *i*-ésima observación.

Un valor calculado grande de $DFBETA_{j,i}$ indica que la observación *i* tiene influencia considerable en el coeficiente de regresión j-ésimo. Observe que $DFBETA_{j,i}$ es un vector $n \times (k\text{-}1)$ que da información similar a la distancia de Cook.

El dato se considera influyente si DFBETA$> \frac{2}{\sqrt{n}}$

3.12 INCUMPLIMIENTO DE LOS SUPUESTOS

En esta sección se presentarán brevemente los problemas que surgen cuando los supuestos del modelo de regresión múltiple no se cumplen. Los problemas que se plantearán son de Multicolinealidad y detección de Correlación Serial o Autocorrelación en los residuos.

3.13 MULTICOLINEALIDAD

Este problema surge cuando no se cumple la condición que señala que ninguna de las variables explicativas puede ser una combinación lineal exacta de las otras variables explicativas, es decir las variables independientes están altamente correlacionadas entre si, de tal manera que unas dependen de otras.

Si las variables independientes están perfectamente relacionadas entre si en forma lineal, se dice que son ***linealmente dependientes***. En estos casos no se pueden obtener estimaciones de los coeficientes de la ecuación de regresión ya que no se pueden resolver las ecuaciones normales.

Cuando se presenta el problema de **Multicolinealidad** ente las variables independientes, el sistema de ecuaciones normales (que permiten obtener los coeficientes de regresión) se ve afectado y no da una solución única para cada uno de los parámetros de la regresión. El problema de la Multicolinealidad afecta a la descripción del modelo de regresión múltiple, ya que significa que todos los datos se encuentran sobre una misma línea recta y por lo tanto no existe un plano óptimo, sino los infinitos que pasan por dicha recta.

En la práctica, rara vez se encuentran casos de dependencia perfecta ya que los errores de muestreo y de medición son inevitables. Sin embargo, se habla de un problema de Multicolinealidad cuando dos o más variables independientes están altamente correlacionadas entre sí, o cuando hay bajas correlaciones entre dos variables pero altas entre tres o más.

Los efectos de la Multicolinealidad llevan a que los errores estándares de los coeficientes sean elevados, es decir tienden a ser mayor de lo que serían si no hubiera Multicolinealidad. Como consecuencia el valor del estadístico *t* en la prueba de hipótesis de significación de los β_j es más pequeño de lo que debería ser, y por lo tanto, es posible llegar a la conclusión errónea de que la variable independiente *Xi* no es importante en el modelo.

La Multicolinealidad se puede medir a través de la matriz de correlación, la cual permite conocer la tendencia y magnitud de la relación lineal o asociación entre las variables independientes. El modelo de regresión se vuelve cada vez menos confiable a medida que aumenta la correlación entre dichas variables independientes.

En general de acuerdo al valor de correlación se propone la siguiente clasificación:

Valor de r	Intensidad de asociación
$0.00 < r \leq 0.30$	Correlación débil
$0.30 < r \leq 0.75$	Correlación media
$0.75 < r \leq 1.00$	Correlación fuerte

Se considera que existe Multicolinealidad entre dos variables independientes cuando la correlación entre ambas es fuerte ($r > 0.75$). Una manera de corregir este problema es eliminando del modelo una de las variables independientes involucradas en la Multicolinealidad. La pregunta es ¿cuál de las dos?, la respuesta es la menos significativa. Otra manera sería tratar de reemplazar la variable multicolineal por otra menos colineal pero sin alterar el contenido teórico del modelo.

Identificar la variable menos significativa en una regresión no es fácil, ni se puede deducir empíricamente de una simple observación a los datos. Para lograr la identificación de la variable menos significativa de las que están altamente correlacionadas, se puede utilizar un procedimiento estadístico denominado "Análisis Factorial", el cual trata de agrupar aquellas variables que se encuentran muy relacionadas entre si en un único factor, bajo el criterio de que éstas están poco relacionadas con el resto de las variables independientes no incluidas en este factor; de tal manera que se logre pasar de un modelo inicial de "n" variables independientes a otro modelo con "$n - 1$" variables independientes, eliminando de esta manera una de las dos variables con alta correlación.

***Ejemplo 3*:** Un embotellador de refrescos está analizando las rutas de servicio de las máquinas expendedoras en su sistema de distribución. Él está interesado en predecir el tiempo requerido por el conductor de la ruta para mantener las máquinas expendedoras listas. Esta actividad de servicio incluye el abastecimiento de la máquina con los productos de bebida y el mantenimiento de rutina. El ingeniero industrial responsable del estudio sugiere que las dos variables más importantes que afectan el tiempo de reparto son el número de máquinas expendedoras con producto que revisa (número de casos) y la distancia recorrida por el conductor en la ruta. El ingeniero recolectó 25 observaciones del tiempo de reparto, como se muestra en la tabla de datos.

x_1 Número de casos	x_2 Distancia (Pies)	y Tiempo de reparto
7	560	16.68
3	220	11.50
3	340	12.03
4	80	14.88
6	150	13.75
7	330	18.11
2	110	8.00
7	210	17.83
30	1460	79.24
5	605	21.50
16	688	40.33
10	215	21.00
4	255	13.50
6	462	19.75
9	448	24.00
10	776	29.00
6	200	15.35
7	132	19.00
3	36	9.50
17	770	35.10
10	140	17.90
26	810	52.32
9	450	18.75
8	635	19.83
4	150	10.75

Solución:

1. Crear un archivo con tres columnas, en la hoja de cálculo de STATGRAPHICS, una para la variable dependiente y otras dos para las variables independientes.

2. Seguir la secuencia

Relate → Multiple Regression

3. En el diálogo que aparece colocar en el sitio correspondiente la variable respuesta, Tiempo (*Y)*, así como las independientes, Casos y Distancia (X_1 y X_2).

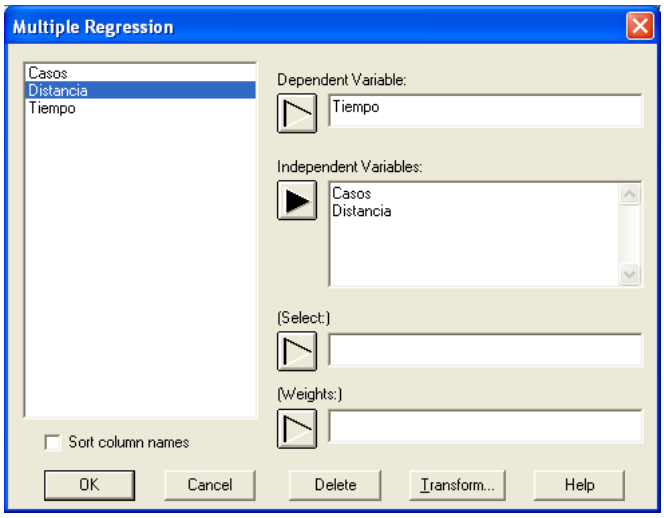


Fig. 30 Regresión Múltiple

4. Dar OK

RESULTADOS

Multiple Regression - Tiempo

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: Tiempo
-----------------------------------------------------------------------------
                                     Standard          T
Parameter               Estimate        Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                 2.34123       1.09673        2.13474         0.0442
Casos                    1.61591      0.170735        9.46442         0.0000
Distancia              0.0143848    0.00361309        3.98131         0.0006
-----------------------------------------------------------------------------

                          Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                    5550.81      2      2775.41     261.24       0.0000
Residual                 233.732     22      10.6242
-----------------------------------------------------------------------------
Total (Corr.)            5784.54     24

R-squared = 95.9594 percent
R-squared (adjusted for d.f.) = 95.592 percent
Standard Error of Est. = 3.25947
Mean absolute error = 2.28396
Durbin-Watson statistic = 1.16957 (P=0.0119)
Lag 1 residual autocorrelation = 0.361037

The StatAdvisor
---------------
   The output shows the results of fitting a multiple linear regression model to describe the
relationship between Tiempo and 2 independent variables. The equation of the fitted model is

Tiempo = 2.34123 + 1.61591*Casos + 0.0143848*Distancia

Since the P-value in the ANOVA table is less than 0.01, there is a statistically significant
relationship between the variables at the 99% confidence level.
```

The R-Squared statistic indicates that the model as fitted explains 95.9594% of the variability in Tiempo. The adjusted R-squared statistic, which is more suitable for comparing models with different numbers of independent variables, is 95.592%. The standard error of the estimate shows the standard deviation of the residuals to be 3.25947. This value can be used to construct prediction limits for new observations by selecting the Reports option from the text menu. The mean absolute error (MAE) of 2.28396 is the average value of the residuals. The Durbin-Watson (DW) statistic tests the residuals to determine if there is any significant correlation based on the order in which they occur in your data file. Since the P-value is less than 0.05, there is an indication of possible serial correlation. Plot the residuals versus row order to see if there is any pattern which can be seen.

In determining whether the model can be simplified, notice that the highest P-value on the independent variables is 0.0006, belonging to Distancia. Since the P-value is less than 0.01, the highest order term is statistically significant at the 99% confidence level. Consequently, you probably don't want to remove any variables from the model.

INTERPRETACIÓN

Los resultados parciales muestran que un ajuste que considera a todas las variables, es significativo y que la ecuación de regresión para el tiempo de reparto, en función del número de casos y la distancia recorrida es:

$$Tiempo = 2.34123 + 1.61591\,Casos + 0.0143848\,Distancia$$

El análisis de varianza indica que todos los coeficientes del modelo son diferentes de cero, por lo que muy probablemente no haya que eliminar ninguna de las variables del modelo.

El coeficiente de determinación indica una variación explicada del 95.592%

La estadística de Durbin-Watson tiene un p-value de 0.0119 < 0.05, lo que indica que se rechaza la hipótesis de independencia de los residuos (de no autocorrelación); es decir existe indicación de posible correlación de los residuos, lo cual podría indicar que la significación de las variables explicativas no es tan pequeña como se aprecia.

NOTA: La recomendación es leer con calma el texto del StatAdvisor.

Continuando con el análisis, regresamos a la ventana de resultados en el STATGRAPHICS para

5. Seleccionar las opciones tabulares y gráficas. En las opciones tabulares, considerar todas las opciones excepto **Condicional Sums of Squares** y **reports.** En las opciones gráficas, seleccionar todas las de residuos y la de observed vs. predicted.

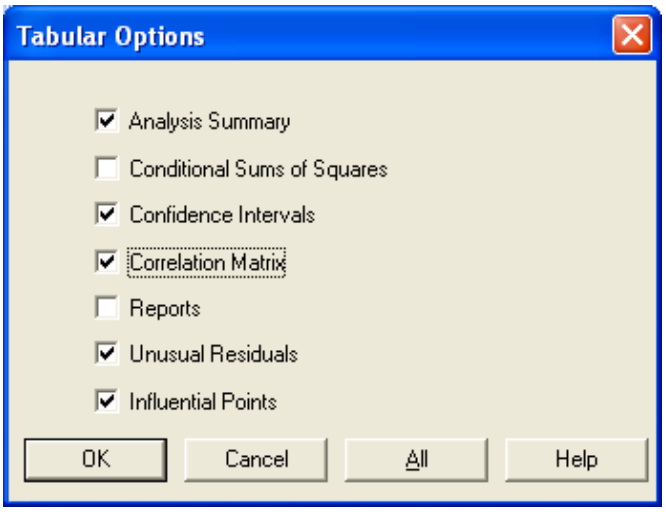


Fig. 31 Opciones Tabulares

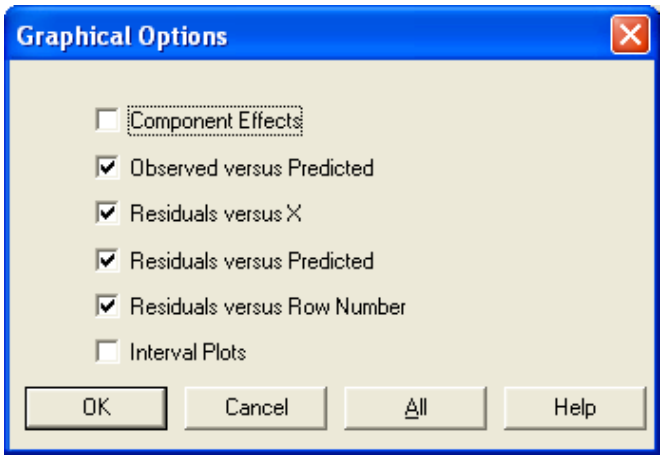


Fig. 32 Opciones Gráficas

RESULTADOS

95.0% confidence intervals for coefficient estimates

Parameter	Estimate	Standard Error	Lower Limit	Upper Limit
CONSTANT	2.34123	1.09673	0.0667472	4.61572
Casos	1.61591	0.170735	1.26182	1.96999
Distancia	0.0143848	0.00361309	0.00689173	0.0218779

The StatAdvisor

This table shows 95.0% confidence intervals for the coefficients in the model. Confidence intervals show how precisely the coefficients can be estimated given the amount of available data and the noise which is present.

INTERPRETACIÓN:

La tabla anterior muestra los intervalos del 95% de confianza para los parámetros de regresión:

$$0.0667472 < \beta_0 < 4.61572$$

$$1.26182 < \beta_1 < 1.96999$$

$$0.00689173 < \beta_2 < 0.0218779$$

Correlation matrix for coefficient estimates

	CONSTANT	Casos	Distancia
CONSTANT	1.0000	-0.2524	-0.2243
Casos	-0.2524	1.0000	-0.8242
Distancia	-0.2243	-0.8242	1.0000

The StatAdvisor

This table shows estimated correlations between the coefficients in the fitted model. These correlations can be used to detect the presence of serious multicollinearity, i.e., correlation amongst the predictor variables. In this case, there is 1 correlation with absolute value greater than 0.5 (not including the constant term).

Unusual Residuals

Row	Y	Predicted Y	Residual	Studentized Residual
9	79.24	71.8203	7.41971	4.31

The StatAdvisor

The table of unusual residuals lists all observations which have Studentized residuals greater than 2.0 in absolute value. Studentized residuals measure how many standard deviations each observed value of Tiempo deviates from a model fitted using all of the data except that observation. In this case, there is one Studentized residual greater than 3.0. You should take a careful look at the observations greater than 3.0 to determine whether they are outliers which should be removed from the model and handled separately.

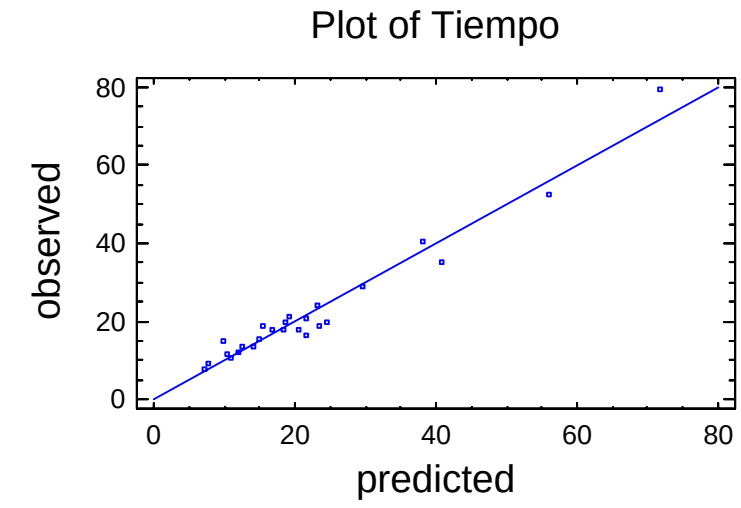
INTERPRETACIÓN:

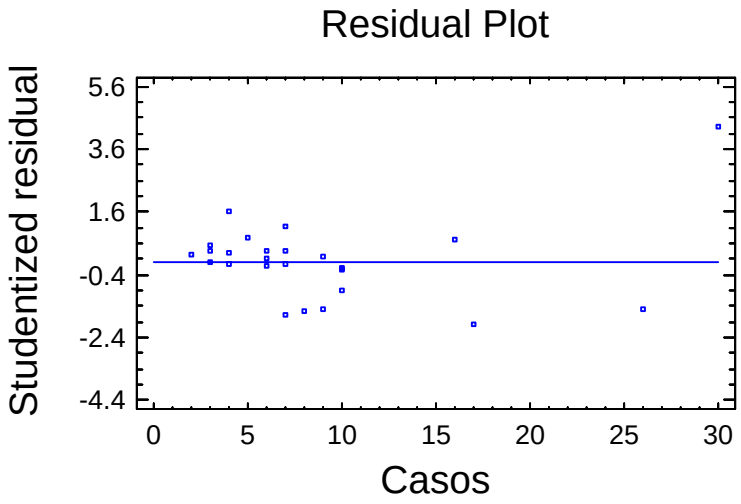
La matriz de correlación muestra una correlación de –0.8242 entre las variables explicativas, lo que indica multicolinealidad, por lo que se podría eliminar una de ellas, ya que ambas explican lo mismo. La pregunta es cuál eliminar, vemos que ambas son altamente significativas, pero la menos significativa es la "Distancia", por lo que se podría eliminar ésta.

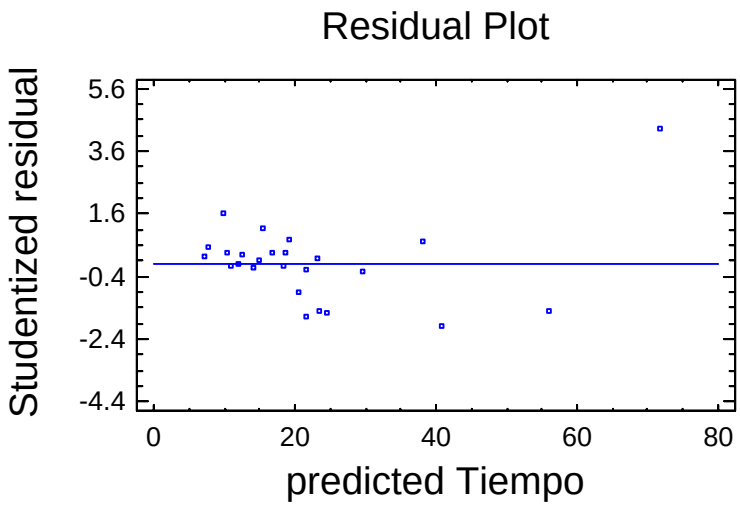
Cabe hacer notar que, como se explicó en la teoría, éste no sería el criterio idóneo, si contáramos con las herramientas del Análisis Multivariado, se podría realizar un "Análisis Factorial" para realizar una combinación de ambas variables en un único factor que considerara una combinación de las propiedades de ambas variables.

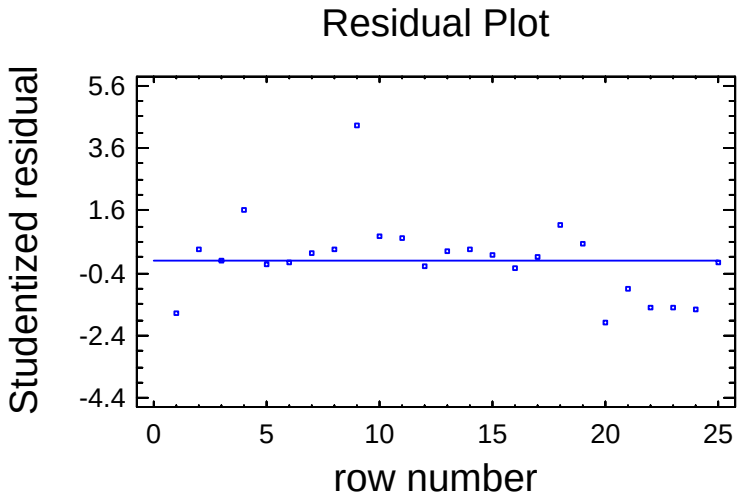
La tabla de residuos indica que existe un residuo studentizado mayor que 3.0, es necesario revisar si hay "outliers", en cuyo caso este punto se debe eliminar.

```
Influential Points
-------------------------------------------------
                       Mahalanobis
Row         Leverage      Distance         DFITS
-------------------------------------------------
     9      0.498292       21.8851       4.29608
    22      0.391575       13.8442      -1.19504
-------------------------------------------------
Average leverage of single data point = 0.12

The StatAdvisor
---------------
   The table of influential data points lists all observations which have leverage values greater
than 3 times that of an average data point, or which have an unusually large value of DFITS.
Leverage is a statistic which measures how influential each observation is in determining the
coefficients of the estimated model. DFITS is a statistic which measures how much the estimated
coefficients would change if each observation was removed from the data set. In this case, an
average data point would have a leverage value equal to 0.12. There are 2 data points with more
than 3 times the average leverage, but none with more than 5 times.  There are 2 data points with
unusually large values of DFITS.
```

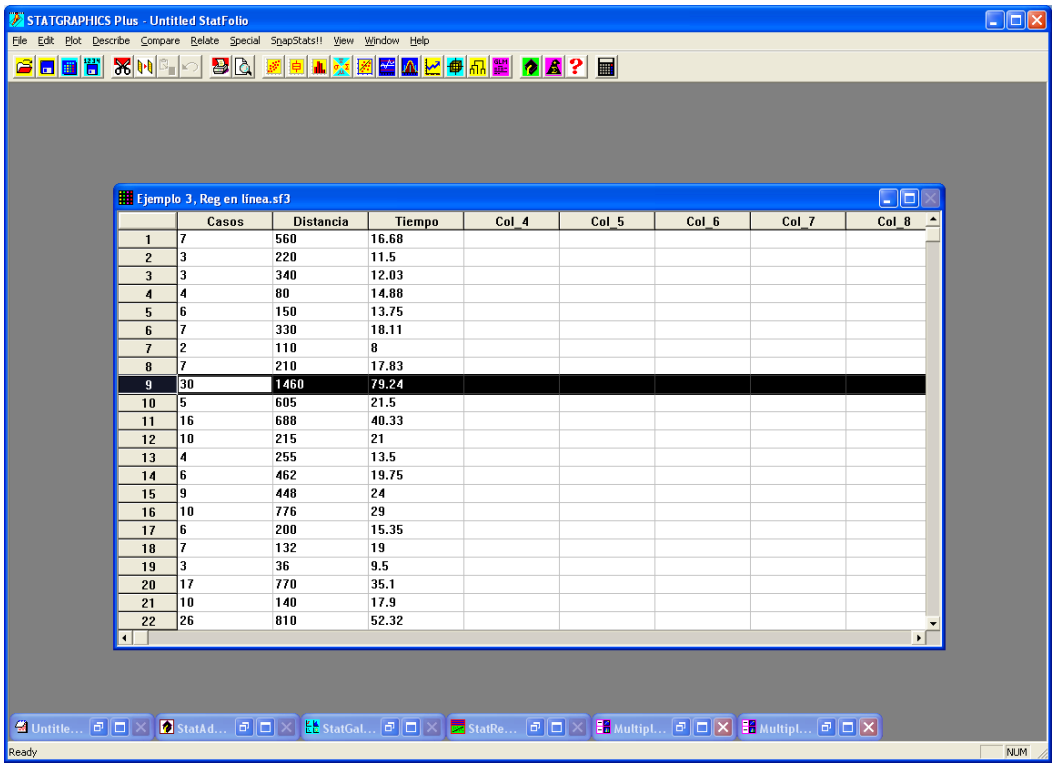








Si se elimina el punto #9, que es un outlier y se "corre" nuevamente el modelo lineal múltiple se obtienen los siguientes resultados



	Casos	Distancia	Tiempo	Col_4	Col_5	Col_6	Col_7	Col_8
1	7	560	16.68					
2	3	220	11.5					
3	3	340	12.03					
4	4	80	14.88					
5	6	150	13.75					
6	7	330	18.11					
7	2	110	8					
8	7	210	17.83					
9	30	1460	79.24					
10	5	605	21.5					
11	16	688	40.33					
12	10	215	21					
13	4	255	13.5					
14	6	462	19.75					
15	9	448	24					
16	10	776	29					
17	6	200	15.35					
18	7	132	19					
19	3	36	9.5					
20	17	770	35.1					
21	10	140	17.9					
22	26	810	52.32					

Fig. 33 Eliminación del punto # 9

RESULTADOS

Multiple Regression - Tiempo

Multiple Regression Analysis

Dependent variable: Tiempo

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	4.44724	0.952469	4.66917	0.0001
Casos	1.49769	0.130207	11.5024	0.0000
Distancia	0.0103241	0.00285359	3.61792	0.0016

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	2293.24	2	1146.62	194.18	0.0000
Residual	124.002	21	5.90488		
Total (Corr.)	2417.25	23			

R-squared = 94.8701 percent
R-squared (adjusted for d.f.) = 94.3815 percent
Standard Error of Est. = 2.43
Mean absolute error = 1.76865
Durbin-Watson statistic = 1.33133 (P=0.0383)
Lag 1 residual autocorrelation = 0.262598

The StatAdvisor

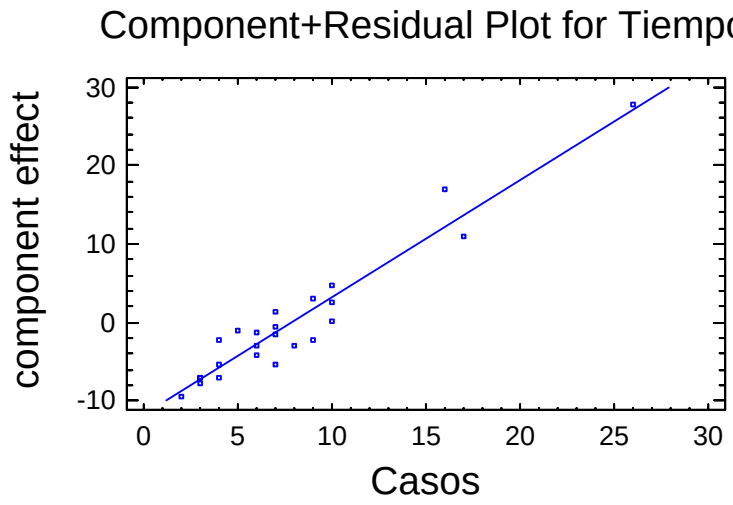
The output shows the results of fitting a multiple linear regression model to describe the relationship between Tiempo and 2 independent variables. The equation of the fitted model is

Tiempo = 4.44724 + 1.49769*Casos + 0.0103241*Distancia

Since the P-value in the ANOVA table is less than 0.01, there is a statistically significant relationship between the variables at the 99% confidence level.

The R-Squared statistic indicates that the model as fitted explains 94.8701% of the variability in Tiempo. The adjusted R-squared statistic, which is more suitable for comparing models with different numbers of independent variables, is 94.3815%. The standard error of the estimate shows the standard deviation of the residuals to be 2.43. This value can be used to construct prediction limits for new observations by selecting the Reports option from the text menu. The mean absolute error (MAE) of 1.76865 is the average value of the residuals. The Durbin-Watson (DW) statistic tests the residuals to determine if there is any significant correlation based on the order in which they occur in your data file. Since the P-value is less than 0.05, there is an indication of possible serial correlation. Plot the residuals versus row order to see if there is any pattern which can be seen.

In determining whether the model can be simplified, notice that the highest P-value on the independent variables is 0.0016, belonging to Distancia. Since the P-value is less than 0.01, the highest order term is statistically significant at the 99% confidence level. Consequently, you probably don't want to remove any variables from the model.

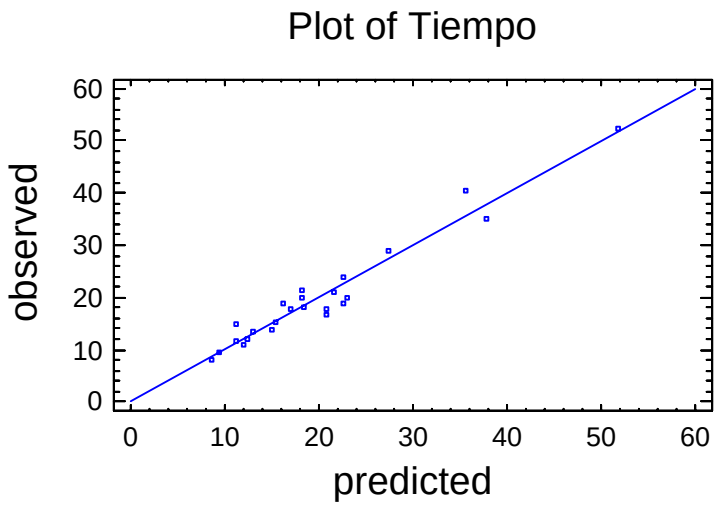


95.0% confidence intervals for coefficient estimates

Parameter	Estimate	Standard Error	Lower Limit	Upper Limit
CONSTANT	4.44724	0.952469	2.46647	6.42801
Casos	1.49769	0.130207	1.22691	1.76847
Distancia	0.0103241	0.00285359	0.00438969	0.0162584

The StatAdvisor

This table shows 95.0% confidence intervals for the coefficients in the model. Confidence intervals show how precisely the coefficients can be estimated given the amount of available data and the noise which is present.

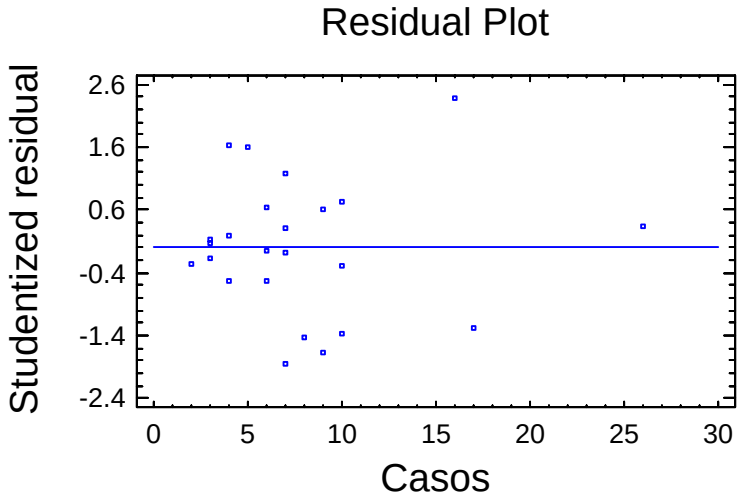


Correlation matrix for coefficient estimates

	CONSTANT	Casos	Distancia
CONSTANT	1.0000	-0.3198	-0.3511
Casos	-0.3198	1.0000	-0.6910
Distancia	-0.3511	-0.6910	1.0000

The StatAdvisor

This table shows estimated correlations between the coefficients in the fitted model. These correlations can be used to detect the presence of serious multicollinearity, i.e., correlation amongst the predictor variables. In this case, there is 1 correlation with absolute value greater than 0.5 (not including the constant term).

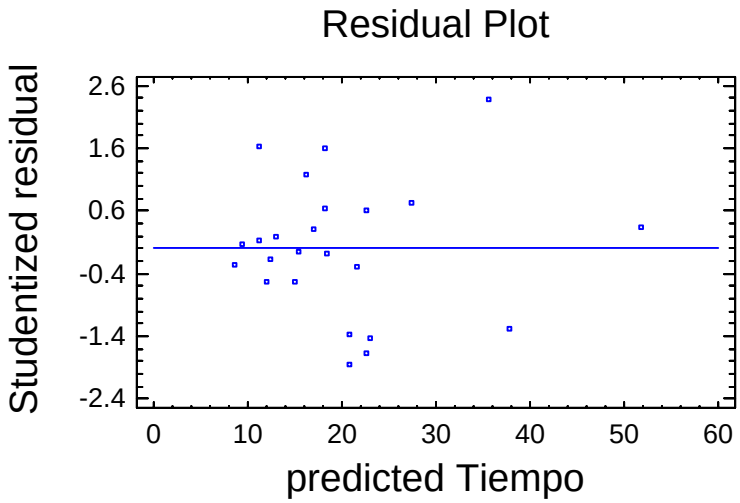


Unusual Residuals

Row	Y	Predicted Y	Residual	Studentized Residual
10	40.33	35.5133	4.81675	2.37

The StatAdvisor

The table of unusual residuals lists all observations which have Studentized residuals greater than 2.0 in absolute value. Studentized residuals measure how many standard deviations each observed value of Tiempo deviates from a model fitted using all of the data except that observation. In this case, there is one Studentized residual greater than 2.0, but none greater than 3.0.



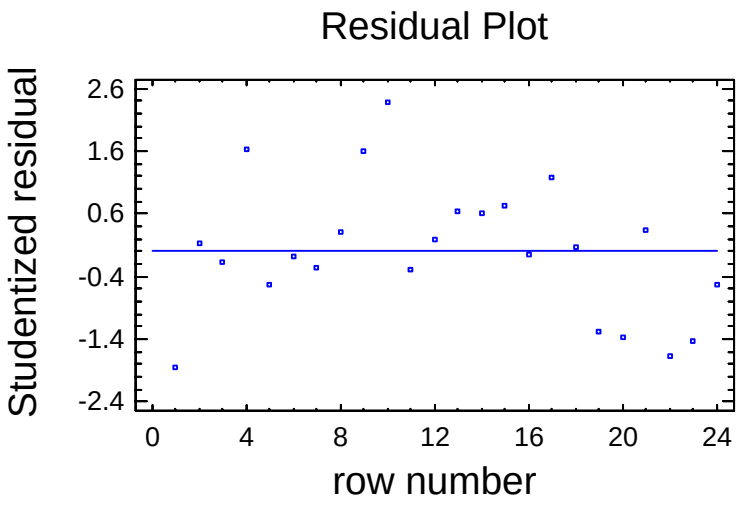
Influential Points

Row	Leverage	Mahalanobis Distance	DFITS
9	0.20438	4.69487	0.803377
10	0.14676	2.82754	0.982968
21	0.556714	26.6729	0.386657

Average leverage of single data point = 0.125

The StatAdvisor

The table of influential data points lists all observations which have leverage values greater than 3 times that of an average data point, or which have an unusually large value of DFITS. Leverage is a statistic which measures how influential each observation is in determining the coefficients of the estimated model. DFITS is a statistic which measures how much the estimated coefficients would change if each observation was removed from the data set. In this case, an average data point would have a leverage value equal to 0.125. There is one data point with more than 3 times the average leverage, but none with more than 5 times. There are 2 data points with unusually large values of DFITS.



INTERPRETACIÓN

Este modelo aparentemente se ajusta muy bien, ya que el p-value del ANOVA es menor que 0.01 y El valor de R^2 es de 94.38%. A pesar de que el problema de la autocorrelación no se resolvió del todo, el p-value aumentó ligeramente, pero las gráficas de los residuos mejoraron significativamente, no hay desviaciones importantes de la normalidad, ni de la homogeneidad de varianzas y la multicolinealidad ya no es problema debido a que la correlación entre las variables explicativas no rebasa 0.75. Tampoco se tienen residuos studentizados mayores que 3.0. Por lo cual nos quedamos con este modelo lineal múltiple.

$$Tiempo = 4.44724 + 1.49769 * Casos + 0.0103241 * Distancia$$

Los intervalos del 95% de confianza para los parámetros de regresión son:

$$2.46647 < \beta_0 < 6.42801$$
$$1.22691 < \beta_1 < 1.76847$$
$$0.00438969 < \beta_2 < 0.0162584$$

3.14 MÉTODOS DE SELECCIÓN DE VARIABLES POR PASOS (PASO A PASO O STEPWISE)

Se han desarrollado métodos alternativos de selección de variables, los cuales identifican bien (aunque no necesariamente los mejores) subconjuntos de modelos, con considerable menos cálculo que el que se requiere para hacer todas las posibles regresiones. Estos métodos son referidos como *métodos de selección por pasos* y se clasifican en tres categorías: (1) Selección Forward (hacia delante), (2) Eliminación Backward (hacia atrás), y (3) Regresión Stepwise (paso a paso), este último es una combinación popular de los procedimientos 1 y 2.

Los subconjuntos de modelos se identifican secuencialmente por adición o eliminación, dependiendo del método, de una variable que tenga mayor impacto en la Suma de Cuadrados residual. Estos métodos por pasos no son garantía para encontrar el mejor subconjunto de variables, y el resultado obtenido por otro método diferente puede no estar acorde con éste.

3.14.1 MÉTODO DE SELECCIÓN HACIA ADELANTE (FORWARD)

La selección Forward inicia con la elección de una variable del subconjunto de variables independientes que explica la cantidad más grande de variación en la variable dependiente. Esta será la variable que tenga la más alta correlación simple con *Y*. Esta variable explicativa o regresor también será el que cause el valor más grande del estadístico *F* en la prueba de significancia de la regresión. Este regresor se incorpora si el estadístico *F* excede un valor *F* preseleccionado, llamado F_{IN} (o *F-to-enter*). El segundo regresor elegido para entrar al modelo será el que ahora tenga la correlación más alta con *Y*, después del ajuste por efecto de la introducción del primer regresor (X_1) en *Y*. Esta correlación se conoce como *correlación parcial*.

Esta correlación es la correlación simple entre los residuales de la regresión $\hat{Y} = b_0 + b_1 X$ y los residuales de la regresión de los otros regresores o variables candidatas en X_1, es decir $\hat{X}_j = \hat{\alpha}_{0j} + \hat{\alpha}_{1j} X_1, j = 2,3,...,k$.

Supongamos que en el paso 2 el regresor con la correlación parcial más alta con *Y* es X_2. Esto implica que el estadístico *F parcial* es

$$F = \frac{SC_{regr.}(X_2 / X_1)}{CM(X_1, X_2)}$$

Si este valor *F* excede F_{IN}, entonces X_2 se adiciona al modelo. En general, en cada paso el regresor que tenga la correlación parcial más alta con *Y* (o equivalentemente la F parcial mas alta dado por los otros regresores ya en el modelo) es adicionado al modelo si esta *F parcial* excede el nivel de introducción F_{IN} preseleccionado. El procedimiento termina cuando el estadístico *F parcial* en un paso particular no excede F_{IN} o cuando el último regresor candidato se adiciona al modelo.

3.14.2 MÉTODO DE ELIMINACIÓN HACIA ATRÁS (BACKWARD)

La eliminación Backward intenta encontrar un buen modelo trabajando en sentido opuesto a la selección Forward. Esto es, inicia con el modelo completo que incluye los *K* regresores candidatos para ir eliminando a cada paso un regresor o variable. Entonces la *F parcial* se calcula para cada regrresor como si fuera la última variable para entrar al modelo . El valor más pequeño de estas *F parciales* es comparado con un valor preseleccionado, F_{OUT} (o *F-to-remove*), por ejemplo, si el valor de F parcial más pequeño es menor que F_{OUT}, el regresor es removido del modelo. Ahora el modelo de regresión con K-1 regresores es ajustado, se calculan los estadísticos *F parciales* para este nuevo modelo y el procedimiento se repite. Una variable que puede haber sido el mejor regresor para entrar en una etapa inicial de un análisis de regresión por

pasos, en una etapa posterior, puede no ser el mejor debido a la relación actual con otras variables en la regresión, hecho por el cual, el valor *F parcial* resulta relevante.

El algoritmo de eliminación Backward termina cuando el valor de *F parcial* más pequeño no es menor que el valor de exclusión F_{OUT} preseleccionado.

La eliminación Backward es con frecuencia un muy buen procedimiento de selección de variables y es particularmente favorecido por analistas que quieren ver los efectos de incluir todas las variables candidatas antes de tener el modelo final.

3.14.3 MÉTODO DE REGRESIÓN PASO A PASO (STEPWISE)

Los dos procedimientos descritos anteriormente sugieren un número de posibles combinaciones. Uno de los más populares es el algoritmo de regresión Stepwise de Efroymson. La regresión Stepwise es una modificación de la selección Forward en la cual en cada paso todos los regresores que entraron en el modelo previamente son reevaluadas vía su estadístico *F parcial*. Un regresor adicionado en un paso temprano puede ahora ser redundante por la relación entre este y los demás regresores que ya se encuentran en la ecuación y si su F parcial de es menor que F_{OUT}, entonces la variable es retirada del modelo. Se continua añadiendo y eliminando variables hasta que el modelo sea estable.

La regresión paso a paso requiere dos cotas, F_{IN} y F_{OUT}. Algunos analistas prefieren elegir $F_{IN} = F_{OUT,}$ aunque esto no es necesario. Frecuentemente se propone elegir $F_{IN}>F_{OUT,}$ para hacer relativamente mas difícil la adición que la eliminación de un regresor.

3.15 COMENTARIOS GENERALES A LOS PROCEDIMIENTOS DE SELECCIÓN DE VARIABLES

Los algoritmos de selección de variables presentados anteriormente han sido criticados por varias razones, la más común es que ninguno de estos procedimientos garantiza generalmente que el mejor subgrupo de variables, de cualquier tamaño, sea identificado para el modelo de regresión. Además todos los tipos de procedimiento stepwise terminan con una ecuación final y la inexperiencia del analista puede conducir a que crean que encontraron un modelo que en algún sentido es óptimo. Parte del problema es que es probable que no sea el mejor modelo, pero será igualmente bueno que otros.

La selección Forward, la eliminación Backward y la regresión Stepwise no necesariamente conducen a la misma elección del modelo final. La intercorrelación entre los regresores afecta el orden de entrada y remoción.

Algunos autores han notado que la selección Forward tiende a concordar con todas la posibles regresiones para pequeños subgrupos (en tamaño) pero no para grandes, mientras que la eliminación Backward tiende a concordar con todas las regresiones de grandes subgrupos pero no con los pequeños. Por estas razones los procedimientos de selección de variables deberán ser

usados con precaución. La recomendación es usar el algoritmo de la regresión Stepwise seguida por la eliminación Backward. El algoritmo de la eliminación Backward es, con frecuencia, menos afectado adversamente por la estructura correlativa de la regresión que la selección Forward.

La elección del valor de cotas F_{IN} y F_{OUT} en los procedimientos tipo stepwise deben considerarse para especificar un "alto a la regla" de este algoritmo. Algunos programas computacionales permiten al analista especificar este número directamente, mientras que otros requieren la elección del Error Tipo I como α para generar F_{IN} y/o $F_{OUT.}$

Algunos usuarios prefieren elegir valores relativamente pequeños de F_{IN} y F_{OUT} más que tener varios regresores adicionales que ordinariamente puedan ser rechazados por valores de F más conservadores. Una propuesta popular es elegir $F_{IN} = F_{OUT} = 4$, que corresponde aproximadamente al 5% superior de la distribución F. También pueden elegirse diferentes valores de F_{IN} y F_{OUT} y observar los efectos de la elección del criterio en los subgrupos obtenidos. Algunos autores recomiendan $\alpha = 0.25$ para la selección Forward, esto corresponde numéricamente a un valor de F_{IN} de entre 1.3 y 2 y $\alpha = 0.10$ para la eliminación Backward.

***Ejemplo* 4 (Regresión Polinómica):** Los siguientes datos acerca de y = concentración de glucosa (g/L) y x = tiempo de fermentación (días), para una mezcla en particular de licor de malta, se obtuvieron de una gráfica de dispersión incluida en el artículo *"Improving Fermentation Productivity with Reverse Osmosis"* (*Food Tech.,* 1984, pp. 92-96). **Tomado de J. L. Devore, (2001), Probabilidad y Estadística para Ingeniería y Ciencias, 5ª. ed., International Thomson Editors, S. A. pp.561.**

x	1	2	3	4	5	6	7	8
y	74	54	52	51	52	53	58	71

Solución:

1. Crear un archivo con dos columnas, en la hoja de cálculo de STATGRAPHICS, una para la variable dependiente y otra para la variable independiente.

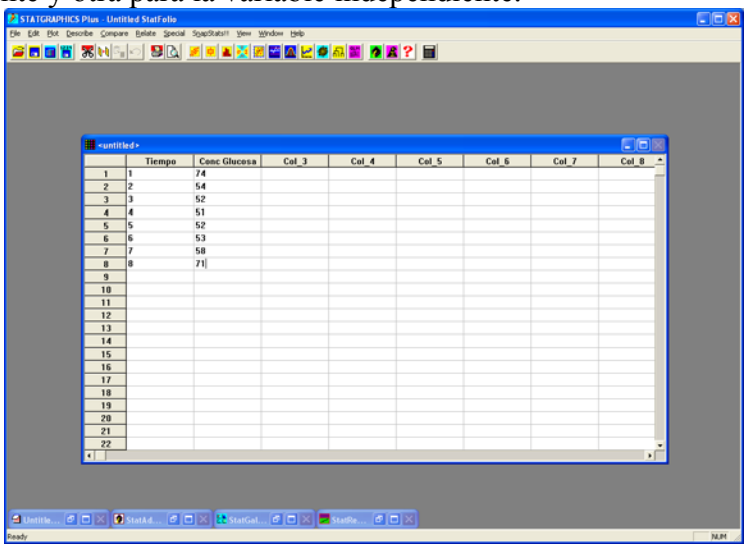


Fig. 34 Datos del ejemplo 4

2. Seguir la secuencia

Relate → Polinomial Regression

3. En el diálogo que aparece colocar en el sitio correspondiente la variable respuesta, (*Y)* Producción, así como la independiente(*X*), Tiempo.

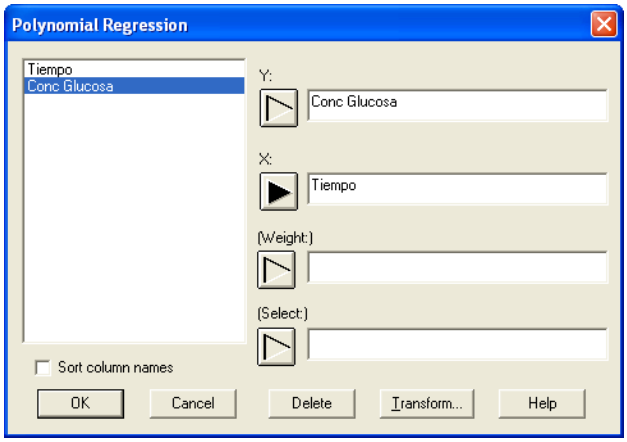


Fig. 35 Regresión Polinómica

4. Dar OK y seleccionar las opciones tabulares y gráficas, de manera semejante a la Regresión lineal Simple

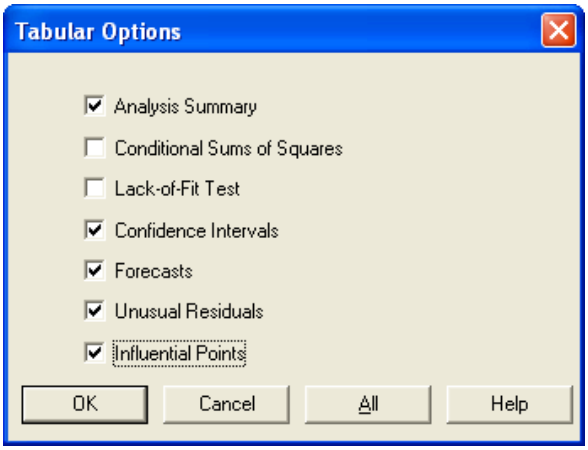


Fig. 36 Opciones Tabulares

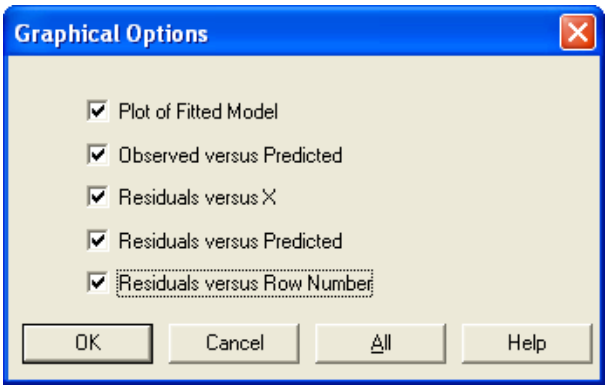


Fig. 36 Opciones Gráficas

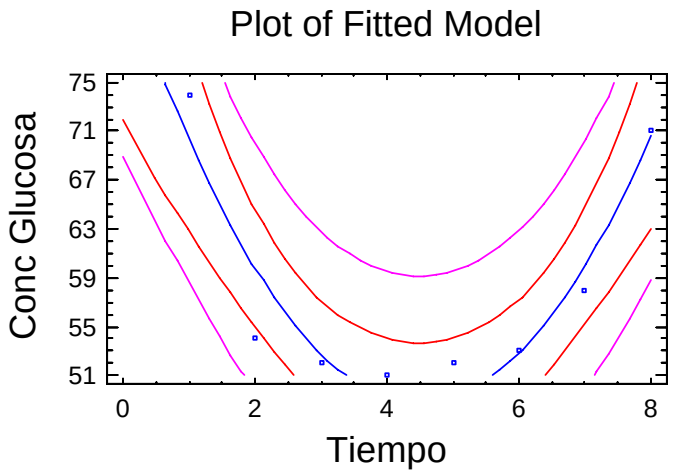
RESULTADOS

Polynomial Regression - Conc Glucosa versus Tiempo

```
Polynomial Regression Analysis
----------------------------------------------------------------------------
Dependent variable: Conc Glucosa
----------------------------------------------------------------------------
                                     Standard          T
Parameter               Estimate       Error       Statistic        P-Value
----------------------------------------------------------------------------
CONSTANT                 84.4821       4.90361       17.2286         0.0000
Tiempo                   -15.875       2.50006      -6.34984         0.0014
Tiempo^2                 1.76786       0.27117       6.51937         0.0013
----------------------------------------------------------------------------

                          Analysis of Variance
----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
----------------------------------------------------------------------------
Model                   525.107      2      262.554      21.25       0.0036
Residual                61.7679      5      12.3536
----------------------------------------------------------------------------
Total (Corr.)           586.875      7

R-squared = 89.4751 percent
R-squared (adjusted for d.f.) = 85.2652 percent
Standard Error of Est. = 3.51476
Mean absolute error = 2.13839
Durbin-Watson statistic = 2.23489 (P=0.0555)
Lag 1 residual autocorrelation = -0.224957

The StatAdvisor
---------------
   The output shows the results of fitting a second order polynomial model to describe the
relationship between Conc Glucosa and Tiempo.

The equation of the fitted model is

Conc Glucosa = 84.4821-15.875*Tiempo + 1.76786*Tiempo^2

Since the P-value in the ANOVA table is less than 0.01, there is a statistically significant
relationship between Conc Glucosa and Tiempo at the 99% confidence level.

   The R-Squared statistic indicates that the model as fitted explains 89.4751% of the
variability in Conc Glucosa. The adjusted R-squared statistic, which is more suitable for
comparing models with different numbers of independent variables, is 85.2652%. The standard error
of the estimate shows the standard deviation of the residuals to be 3.51476. This value can be
```

used to construct prediction limits for new observations by selecting the Forecasts option from the text menu.
The mean absolute error (MAE) of 2.13839 is the average value of the residuals. The Durbin-Watson (DW) statistic tests the residuals to determine if there is any significant correlation based on the order in which they occur in your data file. Since the P-value is greater than 0.05, there is no indication of serial autocorrelation in the residuals.

In determining whether the order of the polynomial is appropriate, note first that the P-value on the highest order term of the polynomial equals 0.00126938. Since the P-value is less than 0.01, the highest order term is statistically significant at the 99% confidence level. Consequently, you probably don't want to consider any model of lower order.



INTERPRETACIÓN

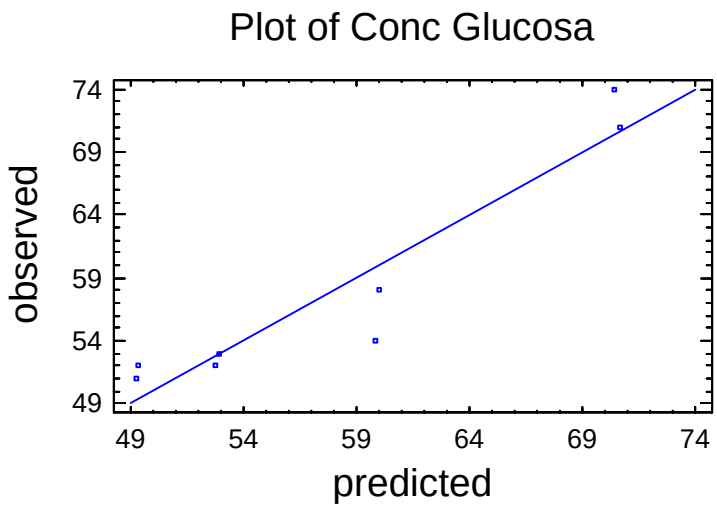
En el gráfico se aprecia que el modelo lineal no es la mejor opción de ajuste, ya que se presenta una curvatura, entonces se propone un modelo cuadrático. Modelo que se ajusta por "default" u omisión, aunque con un clic derecho aparece un menú flotante donde se puede seleccionar Opciones de Análisis y cambiar el orden a 3 = cúbico o uno de mayor orden.

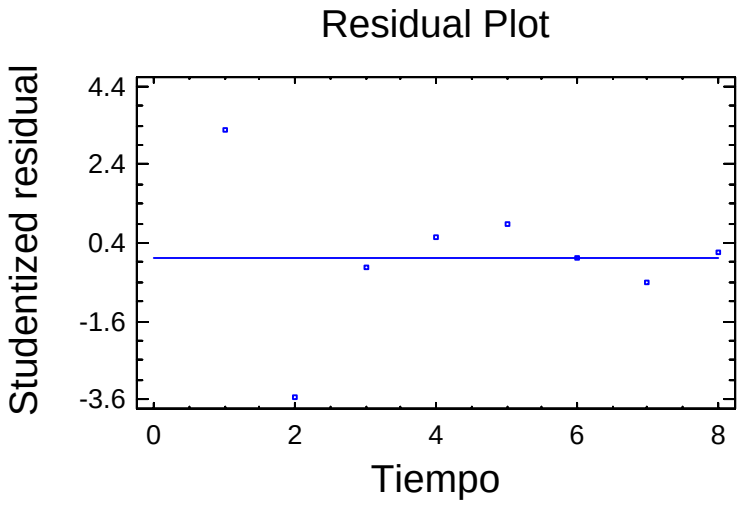
Es importante considerar el valor de R^2 y de R^2 ajustado, este último con un valor del 85.27%

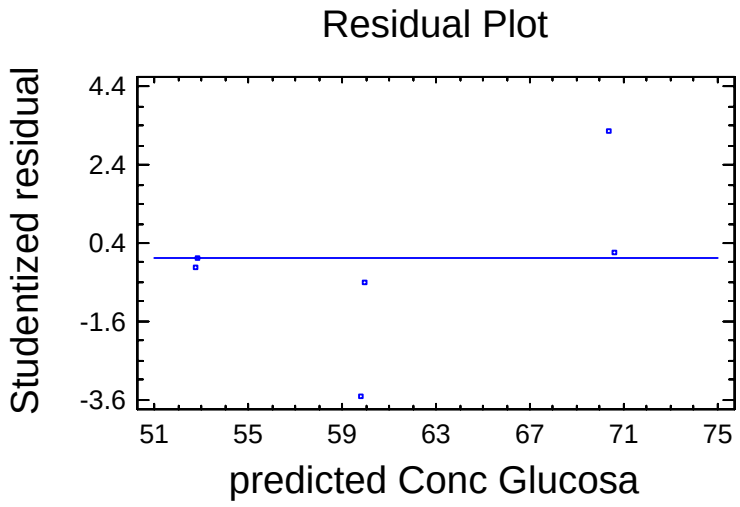
El ANOVA muestra que todos los coeficientes del modelo son diferentes de cero, entonces si hay modelo.

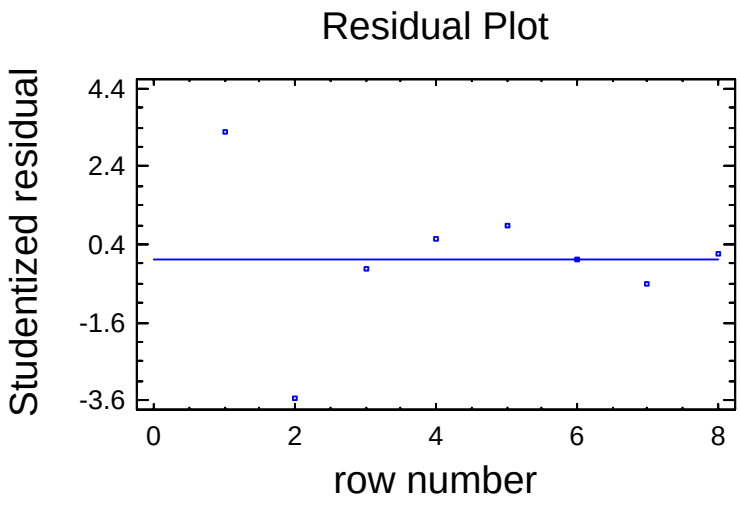
El modelo es $ConcGlu = 84.4821 - 15.875 * Tiempo + 1.76786 * Tiempo^2$

La estadística de Durbin-Watson tiene un p-value de 0.0555 > 0.05, por lo que no hay indicación de no independencia de los residuos.


Plot of Conc Glucosa
observed
predicted
74
69
64
59
54
49
49
54
59
64
69
74


Residual Plot
Studentized residual
Tiempo
4.4
2.4
0.4
-1.6
-3.6
0
2
4
6
8


Residual Plot
Studentized residual
predicted Conc Glucosa
4.4
2.4
0.4
-1.6
-3.6
51
55
59
63
67
71
75



Unusual Residuals

Row	Y	Predicted Y	Residual	Studentized Residual
1	74.0	70.375	3.625	3.28
2	54.0	59.8036	-5.80357	-3.53

The StatAdvisor

The table of unusual residuals lists all observations which have Studentized residuals greater than 2.0 in absolute value. Studentized residuals measure how many standard deviations each observed value of Conc Glucosa deviates from a model fitted using all of the data except that observation. In this case, there are 2 Studentized residuals greater than 3.0. You should take a careful look at the observations greater than 3.0 to determine whether they are outliers which should be removed from the model and handled separately.

Influential Points

Row	Leverage	Mahalanobis Distance	DFITS
1	0.708333	13.7143	5.11713
2	0.279762	1.47344	-2.20063

Average leverage of single data point = 0.375

The StatAdvisor

The table of influential data points lists all observations which have leverage values greater than 3 times that of an average data point, or which have an unusually large value of DFITS. Leverage is a statistic which measures how influential each observation is in determining the coefficients of the estimated model. DFITS is a statistic which measures how much the estimated coefficients would change if each observation was removed from the data set. In this case, an average data point would have a leverage value equal to 0.375. There are no data points with more than 3 times the average leverage. There are 2 data points with unusually large values of DFITS.

INTERPRETACIÓN

El siguiente paso es verificar los otros supuestos del modelo estadístico, que mediante los gráficos de residuales muestran que hay cierta desviación de la normalidad.

No parece haber desviaciones importantes de la homogeneidad de varianzas,

NOTA ACLARATORIA

Es costumbre realizar el análisis de supuestos con los datos originales, aunque se recomienda que se haga con los residuos como se muestra en este ejemplo.

Se puede probar la normalidad de los residuos primero, guardándolos en la hoja de datos y luego haciendo las pruebas de normalidad correspondientes.

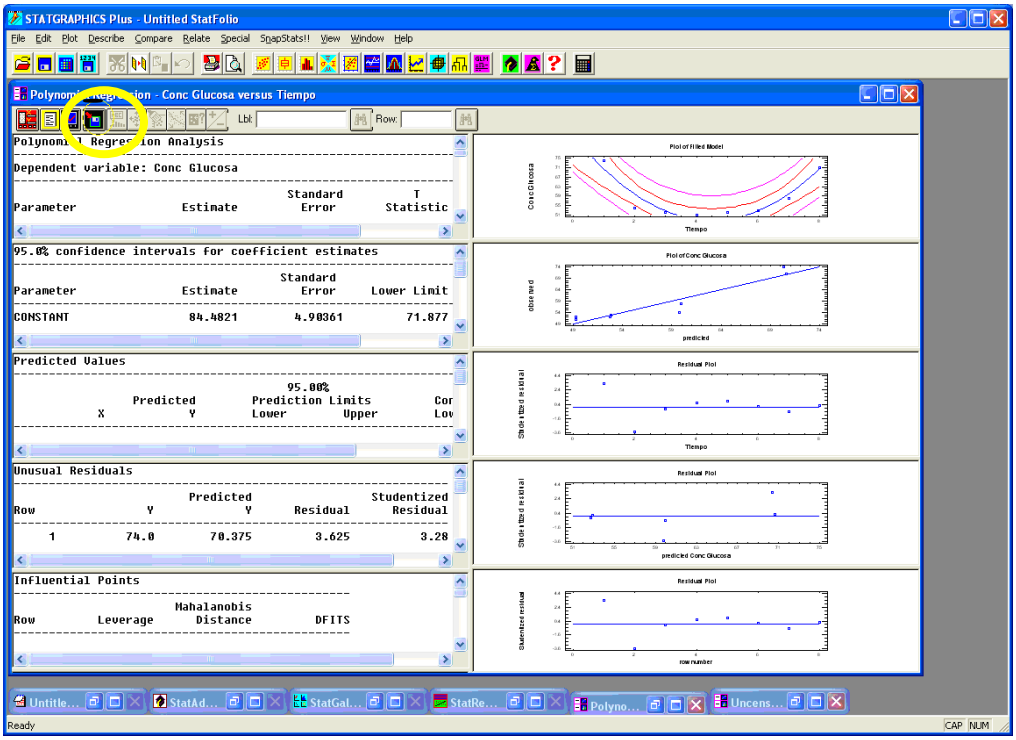


Fig. 38 Selección de la opción SAVE RESULTS

En la VENTANA DE RESULTADOS DE LA REGRESIÓN, dar clic en SAVE RESULTS (4° botón de izquierda a derecha de la esquina superior izquierda de esta ventana).

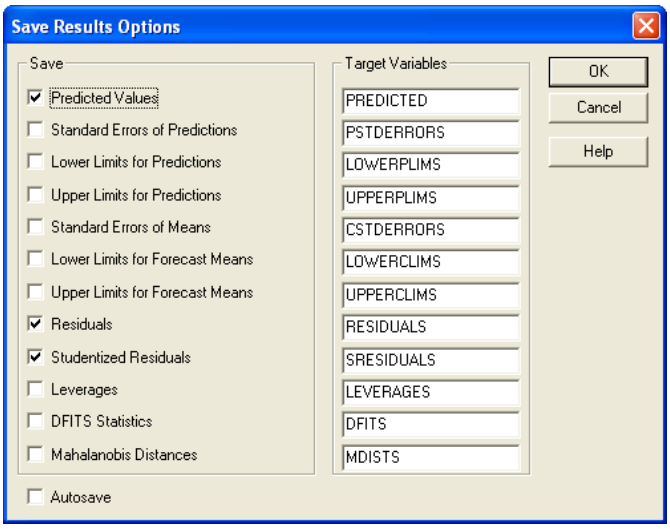
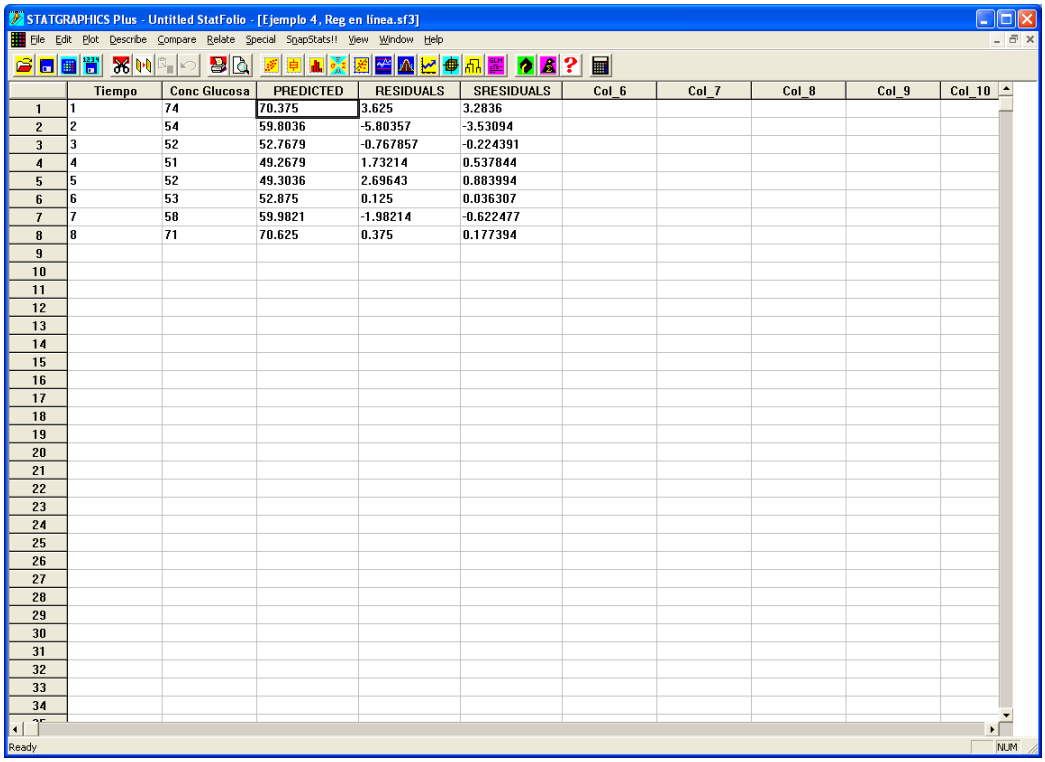


Fig. 39 Guardando los resultados (Save Results Values)

Guardar: Predicted values, Residuals y Studentized Residuals. Dar OK

Su hoja de datos aparece así:



	Tiempo	Conc Glucosa	PREDICTED	RESIDUALS	SRESIDUALS	Col_6	Col_7	Col_8	Col_9	Col_10
1	1	74	70.375	3.625	3.2836					
2	2	54	59.8036	-5.80357	-3.53094					
3	3	52	52.7679	-0.767857	-0.224391					
4	4	51	49.2679	1.73214	0.537844					
5	5	52	49.3036	2.69643	0.883994					
6	6	53	52.875	0.125	0.036307					
7	7	58	59.9821	-1.98214	-0.622477					
8	8	71	70.625	0.375	0.177394					
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										
19										
20										
21										
22										
23										
24										
25										
26										
27										
28										
29										
30										
31										
32										
33										
34										

Fig. 40 Hoja de Datos

El siguiente paso es, del Menú, seguir la secuencia:

Describe → Distributions → Distribution Fitting (Uncensored Data)

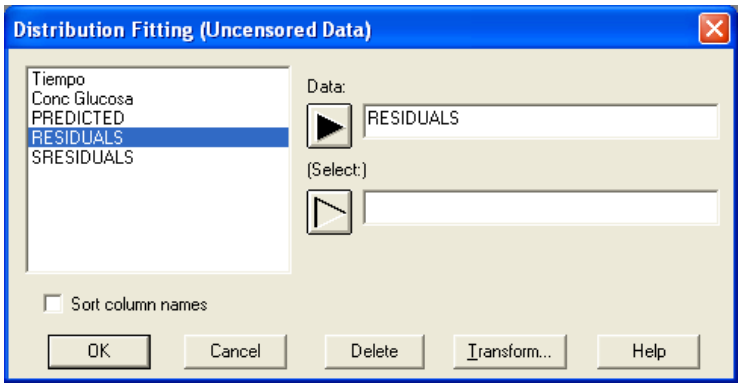


Fig. 41 Bondad de ajuste de los residuos

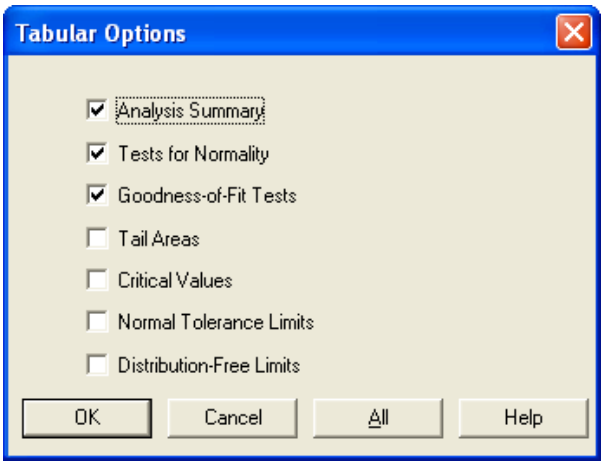


Fig. 42 Opciones tabulares

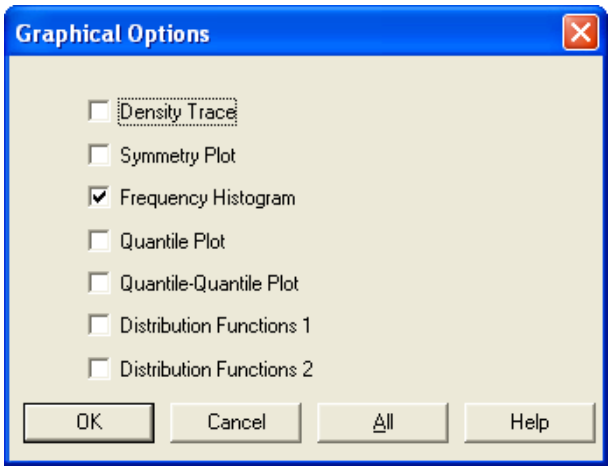
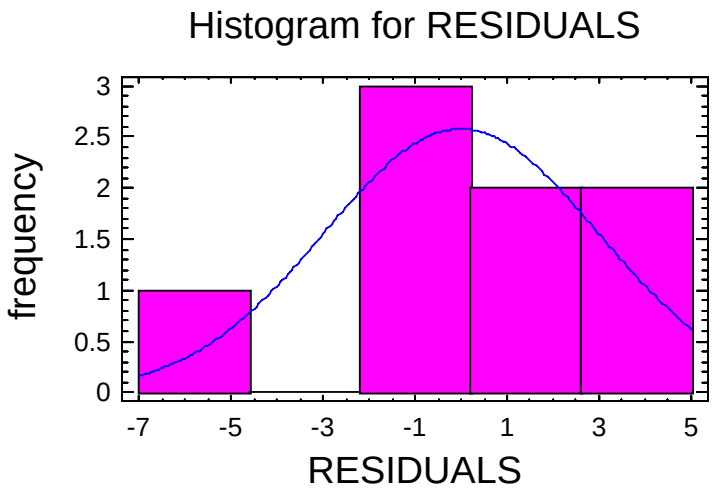


Fig. 43 Opciones Gráficas

RESULTADOS

Uncensored Data - RESIDUALS

```
Analysis Summary

Data variable: RESIDUALS

8 values ranging from -5.80357 to 3.625

Fitted normal distribution:
     mean = 3.75E-7
     standard deviation = 2.97052


The StatAdvisor
---------------
   This analysis shows the results of fitting a normal distribution to
the data on RESIDUALS.  The estimated parameters of the fitted
distribution are shown above.  You can test whether the normal
distribution fits the data adequately by selecting Goodness-of-Fit
Tests from the list of Tabular Options.  You can also assess visually
how well the normal distribution fits by selecting Frequency Histogram
from the list of Graphical Options.  Other options within the
procedure allow you to compute and display tail areas and critical
values for the distribution.  To select a different distribution,
press the alternate mouse button and select Analysis Options.
```



Tests for Normality for RESIDUALS

Computed Chi-Square goodness-of-fit statistic = 5.5
P-Value = 0.481457

Shapiro-Wilks W statistic = 0.943917
P-Value = 0.653312

Z score for skewness = 0.903239
P-Value = 0.366398

Z score for kurtosis not computed.

The StatAdvisor

This pane shows the results of several tests run to determine whether RESIDUALS can be adequately modeled by a normal distribution. The chi-square test divides the range of RESIDUALS into 9 equally probable classes and compares the number of observations in each class to the number expected. The Shapiro-Wilks test is based upon comparing the quantiles of the fitted normal distribution to the quantiles of the data. The standardized skewness test looks for lack of symmetry in the data. The standardized kurtosis test looks for distributional shape which is either flatter or more peaked than the normal distribution. The standardized kurtosis could not be computed. The lowest P-value amongst the tests performed equals 0.366398. Because the P-value for this test is greater than or equal to 0.10, we can not reject the idea that RESIDUALS comes from a normal distribution with 90% or higher confidence.

Goodness-of-Fit Tests for RESIDUALS

Chi-Square Test

	Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chi-Square
	at or below	-0.752577	3	3.20	0.01
above	-0.752577		5	4.80	0.01

Insufficient data to conduct Chi-Square test.

Estimated Kolmogorov statistic DPLUS = 0.11117
Estimated Kolmogorov statistic DMINUS = 0.148011
Estimated overall statistic DN = 0.148011

Approximate P-Value = 0.994751

EDF Statistic	Value	Modified Form	P-Value
Kolmogorov-Smirnov D	0.148011	0.461639	>=0.10*
Anderson-Darling A^2	0.245289	0.276909	0.6547*

*Indicates that the P-Value has been compared to tables of critical values specially constructed for fitting the currently selected distribution. Other P-values are based on general tables and may be very conservative.

The StatAdvisor

This pane shows the results of tests run to determine whether RESIDUALS can be adequately modeled by a normal distribution. The chi-square test was not run because the number of observations was too small.

Since the smallest P-value amongst the tests performed is greater than or equal to 0.10, we can not reject the idea that RESIDUALS comes from a normal distribution with 90% or higher confidence.

Todas las pruebas de normalidad, tienen p-values > 0.10, lo que indica que no existe ningún problema con la normalidad de los residuos.

Por último damos los intervalos de confianza del 95% de los parámetros del modelo y para la media y el pronóstico, cuando el tiempo es de 2.0 y 9.0:

95.0% confidence intervals for coefficient estimates

Parameter	Estimate	Standard Error	Lower Limit	Upper Limit
CONSTANT	84.4821	4.90361	71.877	97.0873
Tiempo	-15.875	2.50006	-22.3016	-9.44836
Tiempo^2	1.76786	0.27117	1.07079	2.46492

The StatAdvisor

This table shows 95.0% confidence intervals for the coefficients in the model. Confidence intervals show how precisely the coefficients can be estimated given the amount of available data and the noise which is present.

$$71.877 < \beta_0 < 97.0873$$
$$-22.3016 < \beta_1 < -9.44836$$
$$1.07079 < \beta_2 < 2.46492$$

Predicted Values

		95.00% Prediction Limits		95.00% Confidence Limits	
X	Predicted Y	Lower	Upper	Lower	Upper
2.0	59.8036	49.5826	70.0246	55.0247	64.5824
9.0	84.8036	69.2948	100.312	72.1984	97.4087

```
The StatAdvisor
---------------
   This table shows the predicted values for Conc Glucosa using the
fitted model.  In addition to the best predictions, the table shows:
   (1) 95.0% prediction intervals for new observations
   (2) 95.0% confidence intervals for the mean of many observations
The prediction and confidence intervals correspond to the inner and
outer bounds on the graph of the fitted model.
```

$$55.0247 < \mu_{y/x=2} < 64.5824 \qquad 72.1984 < \mu_{y/x=9} < 97.4087$$

$$49.5826 < \hat{Y}_{x=2} < 70.0246 \qquad 69.2948 < \hat{Y}_{x=9} < 100.312$$

REFERENCIAS

Charterjee, S. y Price, B. (1991), *Regression Analysis by Example*, 2ª. ed, John Wiley and Sons, Inc., N. Y., U.S.A

Daniel, C. y F. S. Wood (1980), *Fitting Equations to Data.* 2nd. ed. Wiley, New York, U.S.A.

Daniel, W. W.; *Biestadística, Base para el análisis de las ciencias de la salud*. 3ª ed., Editorial UTEHA, S.A. de C.V., México, 1999.

Devore, J. L. (2001), *Probabilidad y Estadística para Ingeniería y Ciencias,* 5ª. ed., Thomson Learning, México.

Draper, N. R. y Smith, H. (1981), *Applied Regression Analysis*, 2ª. ed., John Wiley and Sons, Inc., N.Y., U.S.A.

Marques, M. J. (2004), *Probabilidad y Estadística para Ciencias Químico Biológicas*, 2ª. Ed., FES Zaragoza, UNAM, México.

Marques de Cantú, M. J. (1991), *Probabilidad y Estadística para Ciencias Químico Biológicas*, McGraw-Hill, México.

Montgomery, D. C., Peck, E. A. y Vining, G. G. (2001), *Introduction to Linear Regression Analysis*, 3rd. ed., John Wiley and Sons, Inc. N.Y., U.S.A.

Rawlings, J. O. (1988), *Applied Regression Analysis, A research Tool*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, California, U.S.A.

Velleman, P. F. Y Hoaglin, D. C. (1981), *Applications, Basics, and Computing of Exploratory Data Analysis.* Duxbury Press, Boston, Massachusets, U.S.A.

Wackerly, D. D., Mendenhall III, W. y Scheaffer, R. (2002), *Estadística Matemática con Aplicaciones,* 6ª. ed., Thomson Learning, México.

Análisis de Regresión

Un Enfoque Práctico

1a. Edición

Se imprimió en el Laboratorio de Aplicaciones

Computacionales de la FES Zaragoza

Con un tiraje inicial de 100 ejemplares

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ESTUDIOS SUPERIORES ZARAGOZA

El presente material se elaboró con la finalidad de dar a conocer las herramientas básicas de la Regresión, no como un libro especializado en la materia, de los cuales existen muchos, sino de mostrar la forma de resolver problemas prácticos con ayudar de un software estadístico, sin perder de vista el rigor de los diferentes modelos. Se pretende que este material no sólo sirva a los estudiantes del DIPLOMADO EN ESTADÍSTICA EN LÍNEA de la FES ZARAGOZA, UNAM, sino a todos aquellos que necesitan de esta herramienta para analizar sus datos derivados de su investigación.

El material consta principalmente de las técnicas de REGRESIÓN LINEAL SIMPLE, CURVILÍNEA, MÚLTIPLE y POLINÓMICA con especial énfasis en el DIAGNÓSTICO de la regresión por medio del ANÁLISIS DE RESIDUOS, AUTOCORRELACIÓN y MULTICOLINEALIDAD; se apoya con el uso del software de análisis estadístico STATGRAPHICS para enfatizar en el análisis e interpretación de los problemas y sus resultados más que en el cálculo.

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ESTUDIOS SUPERIORES ZARAGOZA

PAPIME PE100606

ISBN 978-970-32-4722-6