Regresión Múltiple

Resumen

El procedimiento de **Regresión Múltiple** está diseñado para construir un modelo estadístico describiendo el impacto de dos o más factores cuantitativos X sobre una variable dependiente Y. El procedimiento incluye una opción para realizar *regresión por pasos*, en la cual se selecciona una de las variables X antes establecidas. El modelo colocado puede ser usado para hacer predicciones, incluyendo límites de confianza y límites de predicción. Los residuos pueden también ser graficados observando la manera en que influyen.

El procedimiento contiene opciones adicionales para transformar los datos usando una transformación Box-Cox o Cochrane-Orcutt. La primera opción es útil para establecer la variabilidad de los datos, mientras que la segunda es útil para manejar datos de series de tiempo, en los que los residuos exhiben correlación serial.

Muestra StatFolio: *multiple reg.sgp*

Datos de muestra:

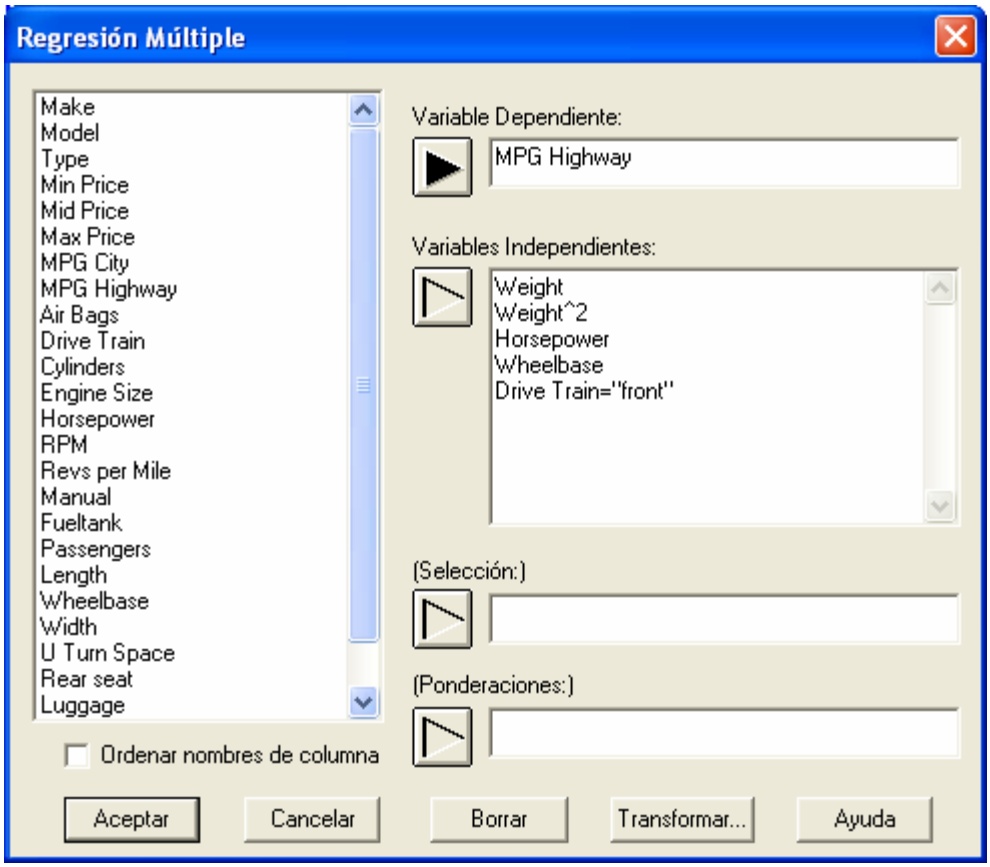
El archivo *93cars.sf3* contiene información sobre 26 variables por $n = 93$ marcas y modelos de automóviles, tomadas de Lock (1993). La tabla a continuación muestra una lista parcial de 4 columnas de ese archivo:

Mark (Marca)	*Model (Modelo)*	*MPG Highway (MPG en Autopista)*	*Weight (Peso)*	*Horsepower (Caballos de fuerza)*	*Wheelbase (Distancia entre ejes)*	*Drivetrain*
Acura	Integra	31	2705	140	102	frontal
Acura	Legend	25	3560	200	115	frontal
Audi	90	26	3375	172	102	frontal
Audi	100	26	3405	172	106	frontal
BMW	535i	30	3640	208	109	trasera
Buick	Century	31	2880	110	105	frontal
Buick	LeSabre	28	3470	170	111	frontal
Buick	Roadmaster	25	4105	180	116	trasera
Buick	Riviera	27	3495	170	108	frontal
Cadillac	DeVille	25	3620	200	114	frontal
Cadillac	Seville	25	3935	295	111	frontal
Chevrolet	Cavalier	36	2490	110	101	frontal

Se desea que un modelo pueda predecir *MPG Carretera* a partir de *Peso*, *Horsepower*, *Wheelbase* y *Drivetrain*.

Datos de entrada

La ventana de dialogo de datos de entrada necesita los nombres de las columnas que contienen la variable dependiente Y y las variables independientes X:



- **Variable Dependiente:** columna numérica que contiene las *n* observaciones para la variable dependiente Y.

- **Variables Independientes:** columnas numéricas que contienen los *n* valores para las variables independientes X. Pueden ser ingresados los nombres de las columnas o expresiones STATGRAPHICS.

- **Selección:** Subconjunto a seleccionar.

- **Ponderaciones:** una columna numérica opcional que contiene pesos para ser aplicados a los cuadrados de los residuos cuando se realice un ajuste de mínimos cuadrados ponderados.

En el ejemplo, note el uso de la expresión *Weight^2* para añadir un término de segundo orden que involucra el peso del vehículo. Esto fue añadido después de examinar la gráfica X-Y que mostró una curvatura significativa con respecto a *Peso*. El factor categórico *Drivetrain* ha sido también introducido en el modelo a través de la expresión booleana *Drivetrain="front"*, la cual establece una variable indicadora que toma el valor de 1 si es verdadero y 0 si es falso. El modelo para ser ajustado toma la forma:

$$MPG\ Carretera = \beta_0 + \beta_1 Peso + \beta_2 Peso^2 + \beta_3 Caballos\ de\ fuerza + \beta_4 Distancia\ entre\ ejes + \beta_5 X_5 \quad (1)$$

donde

$$X_5 = \begin{cases} 1 & \text{si } Drivetrain = front \\ 0 & \text{si } Drivetrain = rear \end{cases} \quad (2)$$

Resumen del análisis

El *Resumen del análisis* muestra la información acerca del modelo ajustado.

Regresión Múltiple - MPG Highway

Variable dependiente: MPG Highway (miles per gallon in highway driving)
Variables independientes:
- Weight (pounds)
- Weinght^2
- Horsepower (maximum)
- Wheelbase (inches)
- Drive Train="front"

		Error	*Estadístico*	
Parámetro	*Estimación*	*Estándar*	*T*	*Valor-P*
CONSTANTE	49.8458	10.5262	4.73539	0.0000
Weigght	-0.0273685	0.00530942	-5.1547	0.0000

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
Modelo	1902.18	5	380.435	46.41	0.0000
Residuo	713.136	87	8.19696		
Total (Corr.)	2615.31	92			

R-cuadrado (ajustado para g.l.) = 71.1652 porciento
Error Estándar Est. = 2.86303
Error Absoluto medio = 2.13575
Estadístico Durbin-Watson = 1.685 (P=0.0601)
Autocorrelación de residuos en Retraso 1 = 0.156111

Están incluidas en la salida:

- **Variables:** identificación de la variable dependiente. La forma general del modelo es

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \quad (3)$$

  Donde *k* es el número de variables independientes.

- **Coeficientes:** los coeficientes estimados, errores estándar, estadístico t y P-valores. Las estimaciones de los coeficientes del modelo pueden ser usadas para escribir la ecuación ajustada, que en el ejemplo es

$$\begin{aligned} MPG\ Carretera = {} & 49.8458 - 0.0273685 * Peso + 0.00000261405 * Peso^2 \\ & + 0.0145764 * Horsepower + 0.338687 * Wheelbase \\ & + 0.632343 * Drive\ Train="front" \end{aligned} \quad (4)$$

  El estadístico t evalúa la hipótesis nula que corresponde al parámetro del modelo igual a 0, basado en la suma de cuadrados Tipo 3 (la suma extra de cuadrados atribuible a cada variable si se ingresa en el modelo). P-Valores grandes (mayores que o iguales a 0.05 si operan a un

nivel de significancia de 5%) indican que un termino puede ser descartado sin degradar significativamente el modelo *provisto de todas los demás variables en el modelo*. En este caso, ambos *Caballos de fuerza* y *Tracción* no son significativos. Por lo tanto, cualquier variable (pero no necesariamente ambas) puede ser descartada del modelo sin dañar su poder predictivo significativamente.

- **Análisis de varianza:** descomposición de la variabilidad de la variable dependiente Y en un modelo de suma de cuadrados y un residuo o suma de errores cuadráticos. De interés particular es el F-prueba y su asociado P-valor, el cual evalúa la significancia estadística del modelo ajustado. Un P-valor pequeño (menor a 0.05 si se opera a un nivel de significancia de 5%) indica que una relación significativa de la forma especificada existe entre Y y las variables independientes. En los datos de muestra, el modelo es altamente significativo.

- **Estadísticos:** resumen de estadísticos para el modelo ajustado, incluyendo:

  R-cuadrado - representa el porcentaje de variabilidad de Y que se ha explicado mediante el modelo ajustado de regresión, oscilando de 0% a 100%. Para los datos de muestra, la regresión ha computado alrededor de 72.7% de la variabilidad en las millas por galón. El restante 27.3% es atribuido a las desviaciones del modelo, las cuales pueden aparecer debido a otros factores, errores de medición o fallas del modelo actual para ajustar los datos adecuadamente.

  R-cuadrados ajustados – el estadístico R-cuadrado, ajustado para el número de coeficientes en el modelo. Este valor es regularmente usado para comparar modelos con diferentes números de coeficientes.

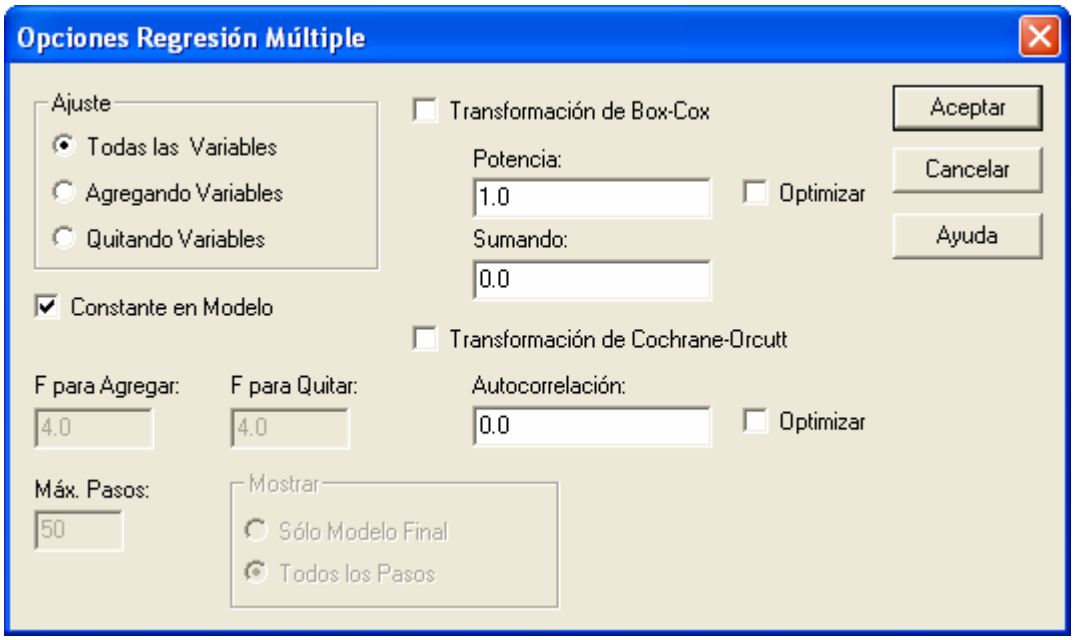
  Error estándar de Est. – La desviación estándar estimada de los residuos (las desviaciones alrededor del modelo). Este valor es usado para crear límites de predicciones para nuevas observaciones.

  Error medio absoluto – el valor absoluto del promedio de los residuos.

  Estadístico Durbin-Watson – una medida de correlación serial en los residuos. Si los residuos varían aleatoriamente, este valor debe ser cercano a 2. Un P-valor pequeño indica un patrón no aleatorio en los residuos. Para datos grabados a través del tiempo, un P-valor pequeño podría indicar que alguna tendencia a través del tiempo no ha sido computada. En el ejemplo, el P-valor es mayor que 0.05, entonces no hay una correlación significativa al nivel de significancia de 5%.

  Autocorrelación residual Lag 1– la correlación estimada entre residuos consecutivos, en la escala de -1 a 1. Valores lejanos de 0 indican que la estructura significativa permanece sin computar por el modelo.

Opciones de análisis



- **Ajuste** – especifica si todas las variables independientes especificadas en el cuadro de diálogo de entrada de datos debe ser incluido en el modelo final, o si una selección de variables por pasos debe aplicarse. La selección por pasos trata de encontrar el mejor modelo que contenga sólo variables estadísticamente significativas. Para la regresión por pasos. Un ejemplo de regresión por pasos es incluida más adelante.

- **Constante en Modelo** – Si esta opción no es marcada, el término β_0 será omitido del modelo. Eliminar el término permite hacer la regresión.

- **F para Agregar** - En una regresión por pasos, las variables serán ingresadas en el modelo a un paso dado si sus F valores son mayores que o iguales al valor especificado de *F para Agregar*.

- **F para Quitar** – En una regresión paso por paso, las variables serán eliminadas del modelo a un paso dado si sus F valores son menores que el valor especificado de *F para Quitar*.

- **Máx. Pasos** – número máximo de pasos permitido cuando se hace una regresión por pasos.

- **Mostrar** – Desplegar o no los resultados a cada paso cuándo se hace una regresión por pasos.

- **Transformación de Box-Cox** – Si se selecciona, una transformación Box-Cox será aplicada a la variable dependiente. Las transformaciones Box-Cox son un método para manejar situaciones en las que las desviaciones del modelo de regresión no tienen una varianza constante. Usted puede especificar los parámetros Box-Cox o pedirle al programa que automáticamente encuentre la potencia óptima. Para ver más detalles, vea la documentación *Transformaciones Box-Cox.*

- **Transformación de Cochrane-Orcutt** – provee un mecanismo para manejar situaciones en las que los residuos del modelo no son independientes. Usted puede especificar la autocorrelación lag 1 usada en la transformación y dejar que el programa la determine

mediante iteraciones. La transformación Cochran-Orcutt es ilustrada más adelante en este documento.

Ejemplo – Regresión por pasos

El modelo se ajusta a los datos de los automóviles mostrando 2 variables insignificantes. Para eliminarlas del modelo se usa *Opciones de análisis* para ejecutar una selección por pasos ascendente o descendente.

- **Selección ascendente** – Comienza con un modelo que involucra un solo término constante e ingresa una variable a un tiempo basado en su significancia estadística si se añade al modelo actual. En cada paso, el algoritmo trae en el modelo la variable que será estadísticamente la más significativa si se ingresa. La selección de variables se basa en una evaluación de *F para Agregar*. La variable más significativa será traída dentro del modelo Mientras que tenga un valor F mayor o igual al especificado en el cuadro de diálogo *Resumen del análisis*. La selección de variables termina cuando ninguna variable tiene un valor F suficientemente grande. Además, las variables traídas al modelo con anterioridad en el procedimiento pueden ser eliminadas más tarde si sus valores F caen debajo del criterio *F para Quitar.*

- **Selección descendente** – Comienza con un modelo que involucra todas las variables especificadas en el cuadro de diálogo de entrada de datos y elimina una variable a la vez basándose en su importancia estadística en el modelo actual. En cada paso, el algoritmo elimina del modelo la variable que es estadísticamente de menor importancia. La eliminación de las variables se basa en una evaluación *F para Quitar.* Si la variable menos significativa tiene un valor F menor que el especificado en el cuadro de diálogo *Resumen del análisis,* será eliminada del modelo. Cuándo todas las variables restantes tienen valores F grandes, el procedimiento se detiene. Además, las variables eliminadas del modelo con anterioridad mediante el procedimiento, pueden ser reingresadas más tarde si sus valores F alcanzan el criterio *F para Agregar.*

En el presente ejemplo, una selección descendente permite lo siguiente:

Regresión por Pasos
Método: Selección Hacia Atrás
F para Introducir: 4.0
F para Eliminar: 4.0

Paso 0:
5 variable(s) en el modelo. 87 g.l. para el error.
R-cuadrado = 72.73% R-cuadrado ajustado = 71.17% CME = 8.19696

Paso 1:
Eliminando variable Drive Train="front" con F para eliminar =0.732595
4 variable(s) en el modelo. 88 g.l. para el error.
R-cuadrado = 72.50% R-cuadrado ajustado = 71.25% CME = 8.17206

Paso 2:
Eliminando variable Horsepower con F para eliminar =2.22011
3 variable(s) en el modelo. 89 g.l. para el error.
R-cuadrado = 71.81% R-cuadrado ajustado = 70.86% CME = 8.28409
Modelo Final seleccionado.

En el primer paso, es eliminada *Tracción* pues es la menos significativa. En el segundo paso, se elimina *Caballos de fuerza*. El algoritmo entonces se detiene, debido a que todas las variables

restantes tienen valores F para eliminar las mayores a 4, y todas las anteriormente eliminadas tienen valores F para ingresar menores que 4.

El modelo reducido es resumido a continuación:

Regresión Múltiple - MPG Highway

Variable dependiente: MPG Highway (miles per gallon in highway driving)
Variables independientes:
- Weight (pounds)
- Weight^2
- Horsepower (maximum)
- Wheelbase (inches)
- Drive Train="front"

		Error	*Estadístico*	
Parámetro	*Estimación*	*Estándar*	*T*	*Valor-P*
CONSTANTE	51.8628	10.2179	5.07569	0.0000
Weight	-0.0245435	0.00506191	-4.84867	0.0000
Weight^2	0.00000236841	8.25606E-7	2.86869	0.0051
Wheelbase	0.28345	0.0899993	3.14947	0.0022

Análisis de Varianza

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
Modelo	1878.03	3	626.009	75.57	0.0000
Residuo	737.284	89	8.28409		
Total (Corr.)	2615.31	92			

R-cuadrada = 71.809 porciento
R-cuadrado (ajustado para g.l.) = 70.8587 porciento
Error estándar del est. = 2.87821
Error absoluto medio = 2.19976

Estadístico Durbin-Watson = 1.67296 (P=0.0558)
Autocorrelación de residuos en retraso 1 = 0.162386

NOTA: de aquí en adelante en este documento, los resultados se basarán en el modelo reducido sin *Drivetrain* o *Wheelbase*.

Ejemplo – Transformación Box-Cox

Si se sospecha que la variabilidad de Y cambia conforme su nivel varía, es útil considerar el uso de una transformación sobre Y. Las transformaciones Box-Cox son en general de la siguiente forma:

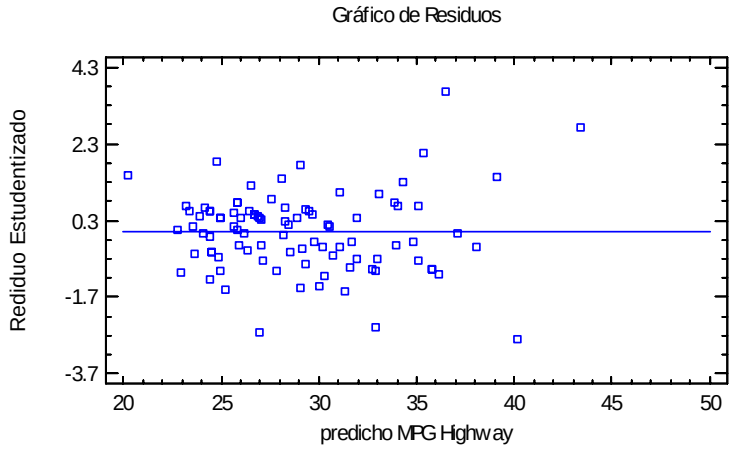
$$Y' = (Y + \lambda_2)^{\lambda_1} \qquad (5)$$

Donde los datos se eleven a la potencia λ_1 después de correrse una cierta cantidad λ_2. Frecuentemente, el parámetro de corrimiento λ_2 se establece como 0. Esta clase incluye raíces cuadradas, logaritmos, recíprocos y otras transformaciones comunes, dependiendo de la potencia. Ejemplos:

Potencia	Transformación	Descripción
$\lambda_1 = 2$	$Y' = Y^2$	Cuadrado
$\lambda_1 = 1$	$Y' = Y$	Dato sin transformar
$\lambda_1 = 0.5$	$Y' = \sqrt{Y}$	Raíz cuadrada
$\lambda_1 = 0.333$	$Y' = \sqrt[3]{Y}$	Raíz cúbica
$\lambda_1 = 0$	$Y' = \ln(Y)$	Logaritmo
$\lambda_1 = -0.5$	$Y' = \frac{1}{\sqrt{Y}}$	Inverso de raíz cuadrada
$\lambda_1 = -1$	$Y' = \frac{1}{Y}$	Recíproco

Usando *Opciones de análisis*, usted puede especificar valores para λ_1 o λ_2, o sólo λ_2 y el programa encontrará el valor óptimo para λ_1 usando los métodos propuestos por Box y Cox (1964).

Para los datos de muestra, una gráfica de residuos contra valores predichos muestra algún cambio en la variabilidad de acuerdo a las fluctuaciones del valor predicho:



Los coches más pequeños tienden a ser un poco más variables que los coches mayores. Cuando se le pide al programa optimizar la transformación de Box- Cox obtenemos lo siguiente:

Regresión Múltiple - MPG Highway

Variable dependiente: MPG Highway (miles per gallon in highway driving)
Variables independientes:
Weight (pounds)
Weight^2
Wheelbase (inches)

Transformación Box-Cox aplicada: potencia = -0.440625 Cambio = 0.0

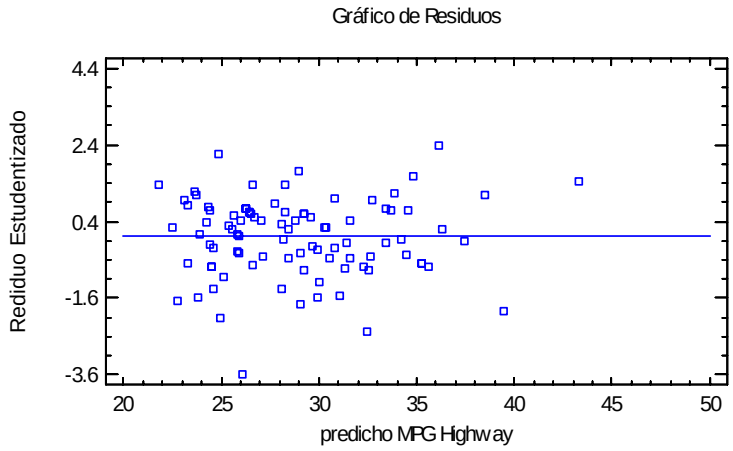
		Error	*Estadístico*	
Parámetro	*Estimación*	*Estándar*	*T*	*Valor-P*
CONSTANTE	230.703	9.37335	24.6126	0.0000
Weight	-0.0129299	0.00464353	-2.78451	0.0065
Weight^2	6.18885E-7	7.57367E-7	0.817153	0.4160
Wheelbase	0.229684	0.0825606	2.782	0.0066

Análisis de Varianza

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
Modelo	1568.28	3	522.761	74.99	0.0000
Residuo	620.444	89	6.97128		
Total (Corr.)	2188.73	92			

R-cuadrada = 71.6528 porciento
R-cuadrado (ajustado para g.l.) = 70.6972 porciento
Error estándar Est. = 2.64032
Error Absoluto medio = 2.08197
Estadístico Durbin-Watson = 1.70034 (P=0.0727)
Autocorrelación de residuos en Retraso 1 = 0.148826

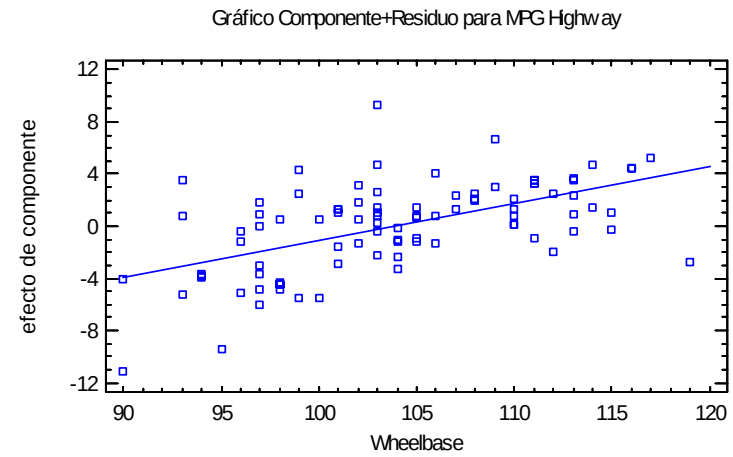
Aparentemente, un inverso de la raíz cuadrada de *MPG Carretera* mejora las propiedades de los residuos, como se ilustra en la nueva gráfica de residuos:



Nota: se necesita tener cuidado aquí, debido a que la trasformación puede verse influenciada fuertemente por una o dos apariciones. Para simplificar la siguiente discusión, el resto de este documento tratará con el modelo no transformado.

Gráfica de efectos de componente

Graficar un modelo de regresión múltiple no es tan fácil como graficar un modelo de regresión simple, ya que el espacio de las variables X es multidimensional. Una manera útil para ilustrar los resultados es la *Gráfica de efectos de componente*, que grafica una porción del modelo de regresión ajustado que corresponde a cualquier variable.



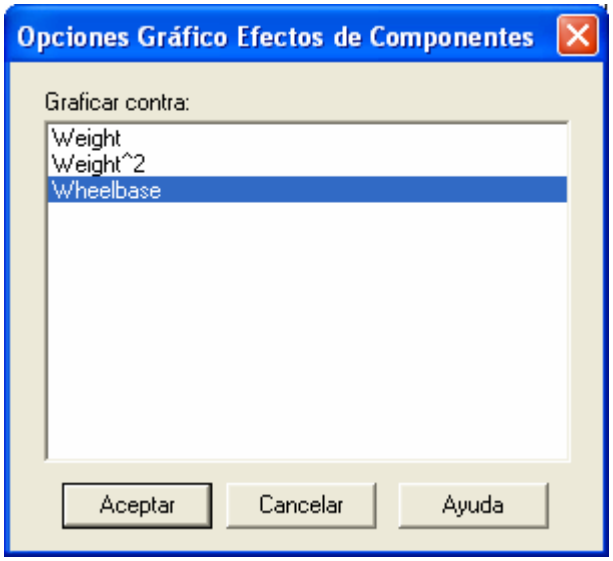
La línea en la gráfica está definida por

$$\hat{\beta}_j\left(x_j - \bar{x}_j\right) \qquad (6)$$

donde $\hat{\beta}_j$ es el coeficiente de regresión estimado para la variable *j*, x_j representa el valor de la variable *j* cómo se graficó en el eje horizontal, y $\bar{x}_j$ es el valor promedio de la variable independiente seleccionada entre las *n* observaciones usadas para ajustar el modelo. Usted puede juzgar la importancia de un factor estudiando qué tanto cambia el efecto de la componente sobre un rango de la variable seleccionada. Por ejemplo, al cambiar *Wheelbase* de 90 a 120, el efecto de la componente cambia alrededor de -4 a +4. Esto implica que las diferencias en *Wheelbase* generan una oscilación de alrededor de 8 millas por galón.

Los puntos en la gráfica anterior representan cada uno de los *n* = 93 automóviles en el conjunto de datos. Las posiciones verticales son iguales al efecto de la componente más el residuo del modelo ajustado. Esto permite evaluar la importancia relativa de un factor comparado con los residuos. En la gráfica anterior, algunos residuos son tan grandes como, si no es que mayores, que el efecto de *Wheelbase*, indicando que otros factores importantes pueden estar faltando en el modelo.

Panel de Opciones



- **Graficar contra**: el factor usado para definir el efecto de la componente.

Suma de cuadrados condicional

El panel de *Sumas de cuadrados condicionales* despliega una tabla mostrando la significancia estadística de cada coeficiente en el modelo como se va añadiendo al ajuste:

ANOVA adicional para Variables en el Orden Ajustado

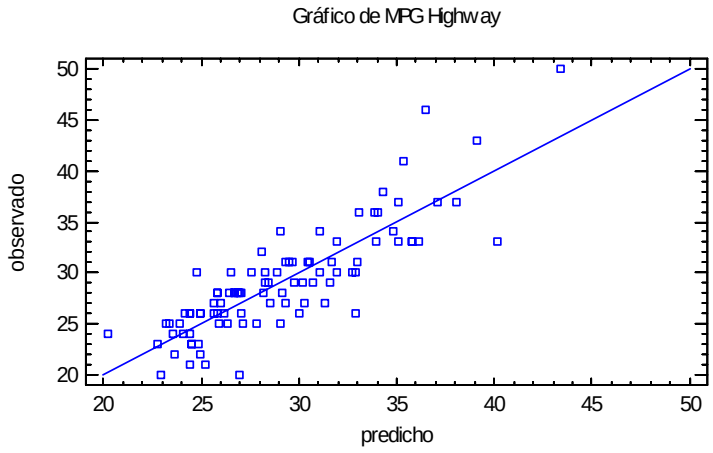
Fuente	*Suma de Cuadrados*	*Gl*	*Media Cuadrática*	*Razón-F*	*Valor-P*
Weight	1718.7	1	1718.7	207.47	0.0000
Weight^2	77.1615	1	77.1615	9.31	0.0030
Wheelbase	82.1713	1	82.1713	9.92	0.0022
Modelo	1878.03	3			

La tabla descompone el modelo de suma de cuadrados SSR en contribuciones debidas a cada coeficiente, mostrando el incremento en SSR cuando cada término se añade al modelo. Estas sumas de cuadrados son regularmente llamadas *Sumas de cuadrados Tipo I*. Los F-radios comparan la media cuadrada para cada término con el MSE del modelo ajustado. Estas sumas de cuadrados son útiles cuando se ajustan modelos polinomiales, como se discute en la documentación de *Regresión Polinomial.*

En la tabla anterior, todas las variables son significativas estadísticamente al nivel de significancia de 1% porque sus P-valores están por debajo de 0.01

Observado contra Predicho

La gráfica *Observado contra Predicho* muestra los valores observados de Y en el eje vertical y los valores predichos $\hat{Y}$ en el eje horizontal.



Si el modelo se ajusta bien, los puntos se deben dispersar aleatoriamente alrededor de la línea diagonal. Cualquier cambio en la variabilidad de los valores bajos de Y o de los valores altos de Y podría indicar la necesidad de transformar la variable dependiente antes de ajustar un modelo a los datos. En la gráfica anterior, la variabilidad incrementa al momento que los valores predichos se hacen mayores.

Gráfica de residuos

Al igual que con todos los modelos estadísticos, es una buena práctica el examinar los residuos. En una regresión, los residuos son definidos por:

$$e_i = y_i - \hat{y}_i \qquad (7)$$

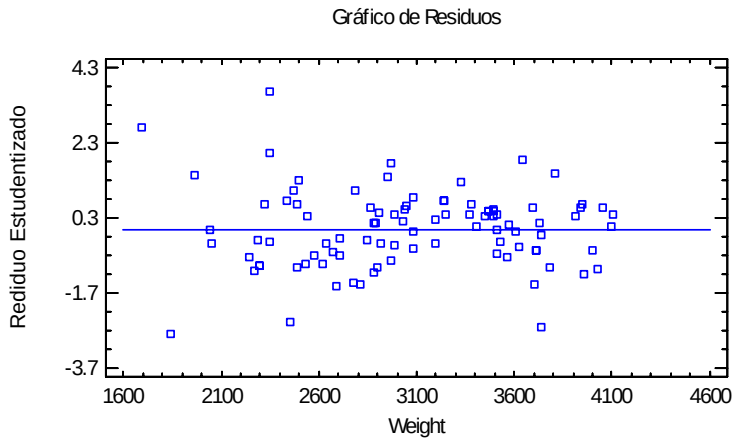
Por ejemplo, los residuos son las diferencias entre los valores de los datos observados y el modelo ajustado.

El procedimiento *Regresión múltiple* crea 3 gráficas de residuos:

1. contra X.
2. contra el valor predicho $\hat{Y}$.
3. contra un número de la lista.

Residuos contra X

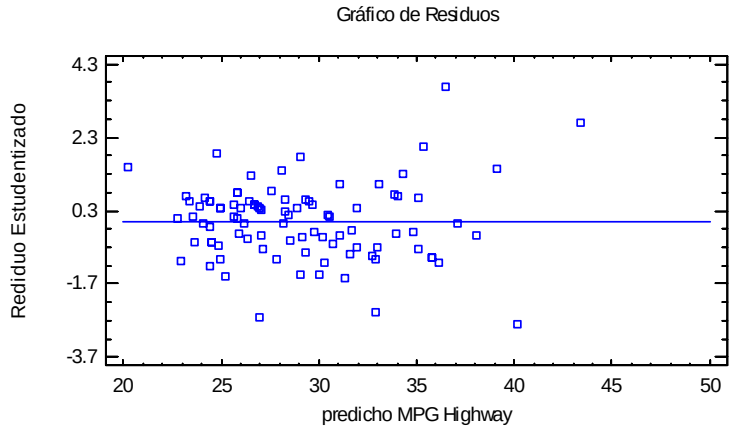
Esta gráfica es de gran ayuda para visualizar cualquier curvatura que se haya perdido con respecto a la variable seleccionada.



No es obvio ver una curvatura en la gráfica anterior.

Residuos contra Predichos

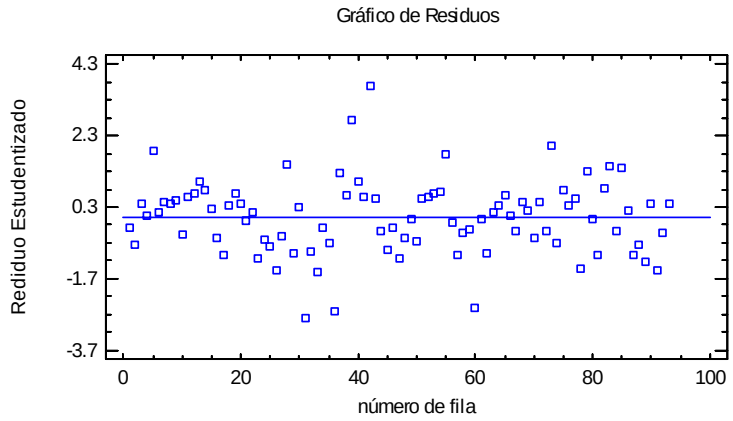
Esta gráfica ayuda a detector cualquier heteroscedasticidad en los datos.



La heteroscedasticidad ocurre cuando la variabilidad de los datos cambia según cambia la media y puede hacerse necesario transformar los datos antes de ajustar el modelo de regresión. Regularmente queda en evidencia mediante un patrón de forma embudo en la gráfica de residuos. En la gráfica anterior, se puede apreciar que la variabilidad incrementa en millas por galón cuando los valores predichos son altos, lo que corresponde a coches pequeños. Para los coches más pequeños, las millas por galón parecen variar más que por los coches mayores.

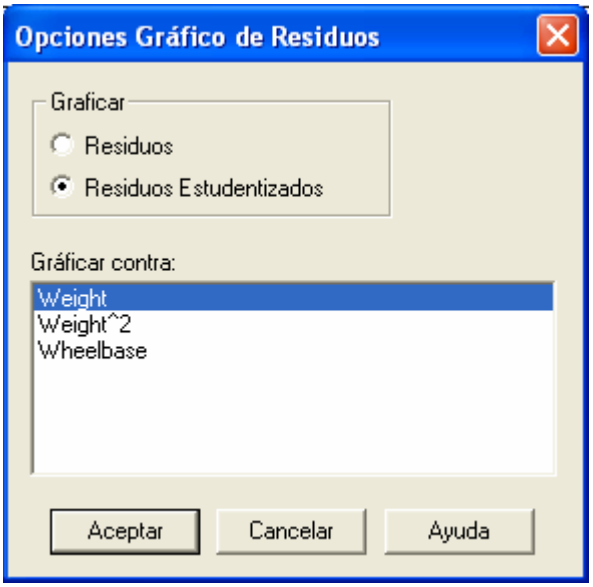
Residuos contra Observación

Esta gráfica muestra los residuos contra números en fila de la hoja de datos:



Si los datos son arreglados en orden cronológico, cualquier patrón en los datos puede indicar una influencia externa. En la gráfica anterior, no hay presente una tendencia obvia, aún cuando hay un residuo estandarizado superior a 3.5, indicando que las desviaciones estándar de la curva ajustada son mayores a 3.5.

Panel de Opciones



- **Graficar**: Los siguientes residuos pueden ser graficados en cada gráfica de residuos:

  1. *Residuos* – los residuos del ajuste por mínimos cuadrados.
  2. *Residuos estudentizados* – la diferencia entre los valores observados y_i y los valores predichos $\hat{y}_i$ cuando el modelo se ajusta usando todas las observaciones excepto la *i-ésima,* dividida entre el error estándar estimado. Estos residuos son algunas veces llamados *residuos borrados externamente*, debido a que miden que tan lejos está cada valor del modelo ajustado cuando ese modelo se ajusta usando todos los datos excepto el punto a ser considerado. Es importante debido a que de otra manera podrían surgir puntos

lejanos que afecten el modelo tanto que no sería extraño que aparecieran muy lejos de la línea.

- **Graficar contra**: La variable independiente a ser graficada en el eje horizontal, en caso de ser relevante.

Residuos Atípicos

Una vez que el modelo ha sido ajustado, es útil estudiar los residuos para determinar si existe algún brote que pudiera ser eliminado de los datos. El panel de *Residuos Atípicos* enlista todas las observaciones que tienen residuos Studentizados de 2.0 o mayores en valor absoluto.

Residuos Atípicos

		Y		*Residuo*
Fila	*Y*	*Predicha*	*Residuo*	*Estudentizado*
31	33.0	40.1526	-7.15265	-2.81
36	20.0	26.9631	-6.96309	-2.62
39	50.0	43.4269	6.5731	2.72
42	46.0	36.4604	9.53958	3.66
60	26.0	32.8753	-6.8753	-2.50
73	41.0	35.3266	5.67338	2.04

I es mayor que 3 corresponden a puntos l modelo ajustado, lo que es un evento muy raro p estra, la fila #42 está a más de 3.5 desviaciones e enlistado en el conjunto de datos como si alcanzara 46 millas por galón, mientras que el modelo predice menos de 37.

Los puntos pueden ser eliminados del ajuste, mientras se examina cualquiera de las gráficas de residuos, pinchando sobre un punto y después presionando el botón *Excluir/Incluir* en la barra de herramientas de análisis.

Puntos Influyentes

Al ajustar un modelo de regresión, no todas las observaciones tienen la misma influencia en el parámetro estimado en el modelo ajustado. Aquellas con valores inusuales de las variables independientes tienden a tener más influencia que las otras. El panel *Puntos Influyentes* despliega cualquier otra observación que tenga una alta influencia en el modelo ajustado:

Puntos Influyentes

		Distancia de	
Fila	*Influencia*	*Mahalanobis*	*DFITS*
19	0.134534	13.1566	0.267983
28	0.237072	27.2882	0.835252
31	0.154695	15.6643	-1.20009
36	0.0936668	8.41545	-0.843505
39	0.244866	28.5193	1.54944
42	0.0667751	5.5222	0.979671
60	0.0298137	1.80729	-0.437436
83	0.10049	9.17701	0.475259

S entes razones:

- ación de la media de las *n* observaciones en el r influencia, mayor el impacto del punto sobre la lista si su influencia es al menos tres veces mayor que un punto promedio.

- **Distancia de Mahalanobis** – mide la distancia de un punto al centro de la colección de puntos en el espacio multivariado de variables independientes. Como la distancia está relacionada con la *influencia*, ésta no se usa para seleccionar puntos para la tabla.

- **DFITS** – mide la diferencia entre los valores predichos $\hat{y}_i$ cuando el modelo es ajustado con y sin el i-ésimo punto. Los puntos se ponen en la lista si el valor absoluto de DFITS excede $2p/\sqrt{n}$, donde *p* es el número de coeficientes en el modelo ajustado.

En los datos muestra, las filas #28 y #39 muestran un valor de influencia cerca de 6 veces el de un punto promedio. Las filas #31 y #39 tienen los mayores valores de DFITS. No se recomienda remover puntos altamente influyentes en una rutina básica. Sin embargo, es importante estar conciente de su impacto en el modelo estimado.

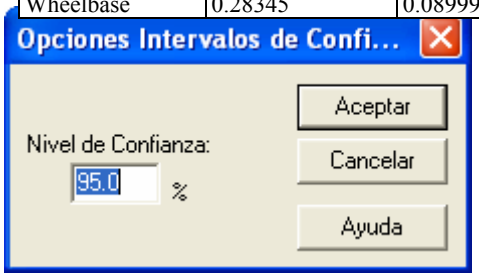
Intervalos de Confianza

El panel *Intervalos de Confianza* muestra error de estimación potencial asociado a cada coeficiente del modelo.

Intervalos de confianza del 95.0% para las estimaciones de los coeficientes

		Error		
Parámetro	*Estimación*	*Estándar*	*Límite Inferior*	*Límite Superior*
CONSTANTE	51.8628	10.2179	31.5601	72.1656
Weight	-0.0245435	0.00506191	-0.0346015	-0.0144856
Weight^2	0.00000236841	8.25606E-7	7.27941E-7	0.00000400887
Wheelbase	0.28345	0.0899993	0.104623	0.462277

P



- **Nivel de Confianza:** porcentaje de nivel de los intervalos de confianza.

Matriz de Correlación

La *Matriz de Correlación* despliega estimadores de la correlación entre los coeficientes estimados.

Matriz de Correlación para las estimaciones de los coeficientes

	CONSTANTE	Weight	Weight^2	Wheelbase
CONSTANTE	1.0000	-0.6355	0.7510	-0.6832
Weight	-0.6355	1.0000	-0.9787	-0.1232
Weight^2	0.7510	-0.9787	1.0000	-0.0566
Wheelbase	-0.6832	-0.1232	-0.0566	1.0000

Esta tabla puede ser de ayuda al determinar que tan bien se han separado los efectos de variables independientes distintas. Note la alta correlación entre los coeficientes de *Weight* y *Weight*^{2}. Esto es normal cada vez que se ajuste un polinomio no centrado y simplemente significa que los coeficientes pudieron cambiar dramáticamente si se hubiese seleccionado un polinomio de orden

distinto. El hecho de que la correlación entre los coeficientes de *Weight* y *Wheelbase* sea pequeña es más interesante, pues implica que hay un poco de confusión entre los efectos estimados de esas variables. Confundir o mezclar los efectos de dos variables es un problema común al intentar interpretar modelos estimados a partir de datos que no fueron recolectados a partir de un experimento diseñado.

Informes

El panel *Informes* crea predicciones usando el modelo ajustado de mínimos cuadrados. Por defecto, la tabla incluye una línea para cada fila de la hoja de datos que contiene información completa de las X variables y un valor faltante para la Y variable. Esto le permite añadir filas en la parte baja de la hoja de datos correspondientes a niveles a los que quiere las predicciones sin afectar el modelo ajustado.

Por ejemplo, suponga que se desea una predicción para un carro con un *Peso* de 3500 y una *Wheelbase* de 105. En la fila #94 de la hoja de datos, esos valores serían añadidos pero la columna *MPG Carretera* se dejaría en blanco. La tabla resultante se muestra a continuación:

Resultados de la Regresión para MPG Highway

	Ajustado	*Error Est.*	*Inferior 95.0%*	*Superior 95.0%*	*Inferior 95.0%*	*Superior 95.0%*
Fila		*LC para Pronóstico*	*LC para Pronóstico*	*LC para Pronóstico*	*LC para la Media*	*LC para la Media*
94	24.7357	2.91778	18.9381	30.5333	23.7842	25.6872

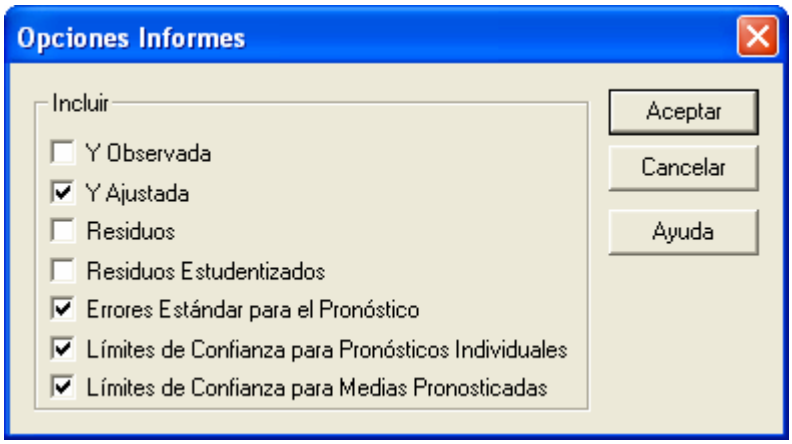
Se incluyen en la tabla:

- **Fila –** el número de fila en la hoja de datos**.**
- **Valor ajustado –** el valor predicho de la variable dependiente usando el modelo ajustado.
- **Error estándar para Pronóstico –** el error estándar estimado de predecir una nueva observación.
- **Límites de Confianza para Pronóstico –** los límites de predicción de nuevas observaciones al nivel seleccionado de confianza.
- **Límites de Confianza para Media –** límites de confianza para el valor medio de Y al nivel de confianza seleccionado.

Para la fila #94, las miles per gallon predichas son 24.7. Puede esperarse que modelos con esas características alcancen entre 18.9 y 30.5 millas por galón manejando en carretera.

Usando el *Panel de Opciones*, se puede incluir información adicional acerca de los valores predichos y de los residuales de los datos usados para ajustar el modelo.

Panel de Opciones

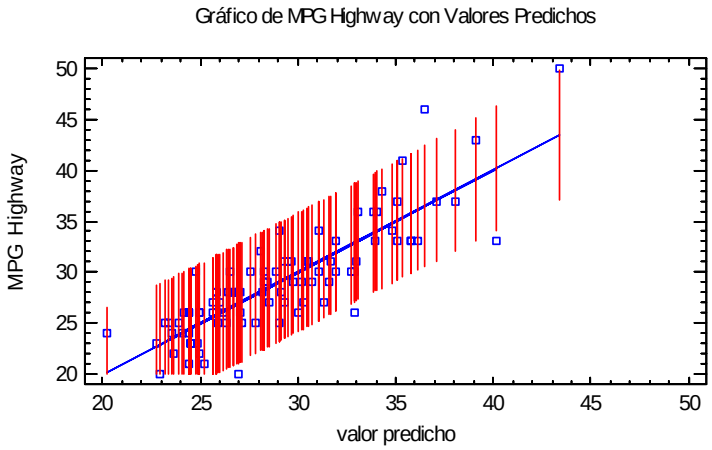


Usted puede incluir:

- *Y Observada* – los valores observados de la variable dependiente.
- *Y Ajustada* – los valores predichos a partir del modelo ajustado.
- *Residuos* – los residuales ordinarios (observados menos predichos).
- *Residuos Estudentizados* – los residuales studentizados borrados como se describió previamente.
- *Errores Estándar para el Pronóstico* – los errores estándar para nuevas observaciones en valores de las variables independientes correspondientes a cada fila de la hoja de datos.
- *Límites de Confianza para Pronósticos Individuales* – intervalos de confianza para nuevas observaciones.
- *Límites de Confianza para Medias Pronosticadas* – intervalos de confianza para el valor medio de Y a valores de las variables independientes correspondientes a cada fila de la hoja de datos.

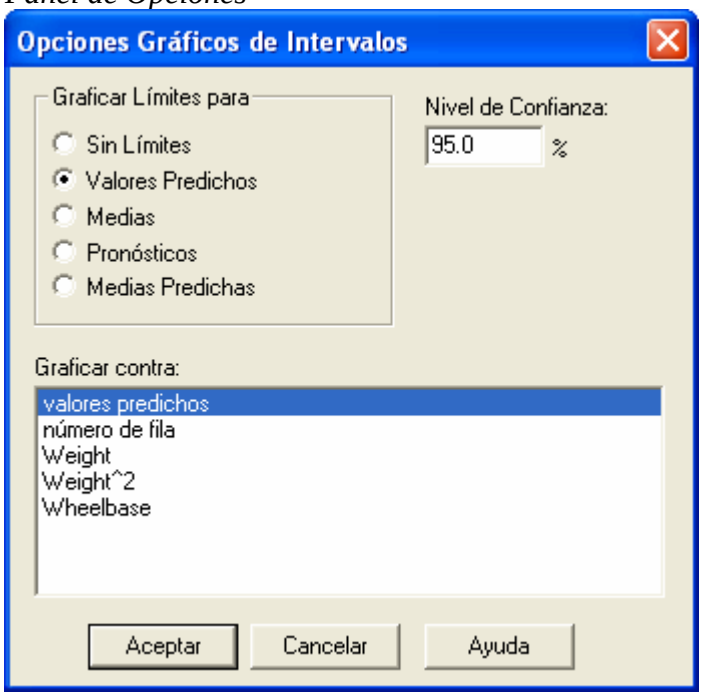
Gráficas de Intervalos

El panel *Gráficas de Intervalos* crea varios tipos interesantes de gráficas. La grafica siguiente muestra cómo las miles per gallon de un automóvil pueden predecirse precisamente.



Se dibuja un intervalo por cada observación de la hoja de datos, mostrando los límites de predicción del 95% para una observación nueva al valor predicho correspondiente.

Panel de Opciones



- **Graficar Límites para:** tipo de límites a incluirse. *Valores Predichos* grafica límites de predicción en el escenario de las variables independientes correspondientes a cada una de las *n* observaciones usadas para ajustar el modelo. *Medias* grafica límites de confianza para el valor medio de Y correspondiente a cada una de las n observaciones. *Pronósticos* grafica límites de predicción para filas de la hoja de datos que tengan valores para Y faltantes. *Medias Predichas* grafica límites de confianza para el valor medio de Y correspondiente a cada fila de la hoja de datos que tengan valores para Y faltantes.

- **Graficar contra:** el valor a graficar en el eje horizontal.

- **Nivel de Confianza:** el porcentaje de confianza usado para los intervalos.

Datos Autocorrelacionados

Cuando se usan modelos de regresión para ajustar datos registrados a lo largo del tiempo, las desviaciones del modelo ajustado son frecuentemente dependientes. Esto puede llevar a estimaciones ineficientes de los coeficientes del modelo de regresión y P-valores que exageren la significancia estadística del modelo ajustado.

Como ilustración, considere los datos siguientes de Neter et al. (1996), contenidos en el archivo *company.sf3*:

Quatre (Año y trimestre)	company sales (Ventas de la Compañía, $ millones)	industry sales (Ventas de la Industria, $ millones)

1983: Q1	20.96	127.3
1983: Q2	21.40	130.0
1983: Q3	21.96	132.7
1983: Q4	21.52	129.4
1984: Q1	22.39	135.0
1984: Q2	22.76	137.1
1984: Q3	23.48	141.2
1984: Q4	23.66	142.8
1985: Q1	24.10	145.5
1985: Q2	24.01	145.3
1985: Q3	24.54	148.3
1985: Q4	24.30	146.4
1986: Q1	25.00	150.2
1986: Q2	25.64	153.1
1986: Q3	26.36	157.3
1986: Q4	26.98	160.7
1987: Q1	27.52	164.2
1987: Q2	27.78	165.6
1987: Q3	28.24	168.7
1987: Q4	28.78	171.7

Hacer regresión de las ventas de las compañía contra las de la industria resulta en un muy buen ajuste lineal, con una R-Cuadrada muy alta:

Regresión Múltiple - company sales

Variable dependiente: company sales
Variables independientes:
industry sales

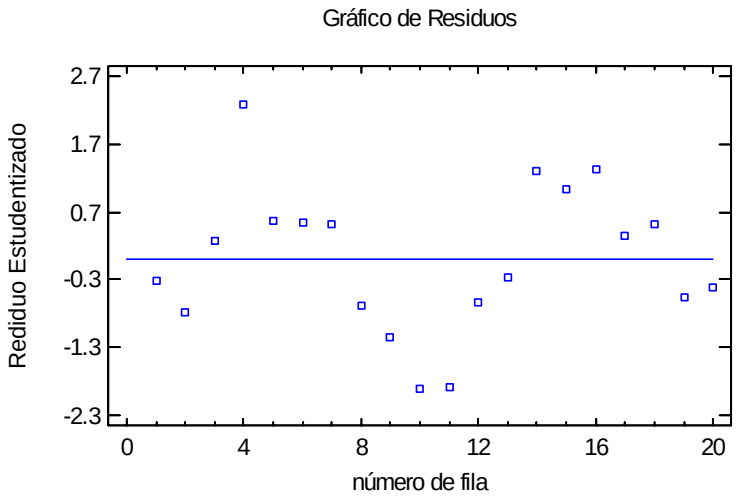
		Error	*Estadístico*	
Parámetro	*Estimación*	*Estándar*	*T*	*Valor-P*
CONSTANTE	-1.45475	0.214146	-6.79326	0.0000
industry sales	0.176283	0.00144474	122.017	0.0000

Análisis de Varianza

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrada Media*	*Razón-F*	*Valor-P*
Modelo	110.257	1	110.257	14888.14	0.0000
Residuo	0.133302	18	0.00740568		
Total (Corr.)	110.39	19			

R-cuadrada = 99.8792 porciento
R-cuadrado (ajustado para g.l.) = 99.8725 porciento
Error estándar del est. = 0.0860563
Error absoluto medio = 0.0691186
Estadístico Durbin-Watson = 0.734726 (P=0.0002)
Autocorrelación de residuos en retraso 1 = 0.626005

Sin embargo, el estadístico de Durbin-Watson es muy importante, y la autocorrelación lag 1 estimada es igual a 0.626. Una gráfica de residuales contra número de fila muestra cambios marcados alrededor de cero.



Claramente, los residuales no están distribuidos aleatoriamente alrededor de la línea de regresión.

Para contar las autocorrelaciones de las desviaciones desde la línea de regresión, puede asumirse una estructura de error más complicada. Una extensión lógica del modelo de error aleatorio es dejar que los errores tengan una estructura autorregresiva de primer orden, en los que la desviación al tiempo *t* es dependiente de la desviación al tiempo *t-1* de la siguiente manera:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \tag{8}$$

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t \tag{9}$$

donde $|\rho| < 1$ y u_t son muestras independientes de una distribución normal con media 0 y desviación estándar σ. En tal caso, transformar tanto la variable dependiente como la independiente de acuerdo a

$$y'_t = y_t - \rho y_{t-1} \tag{10}$$

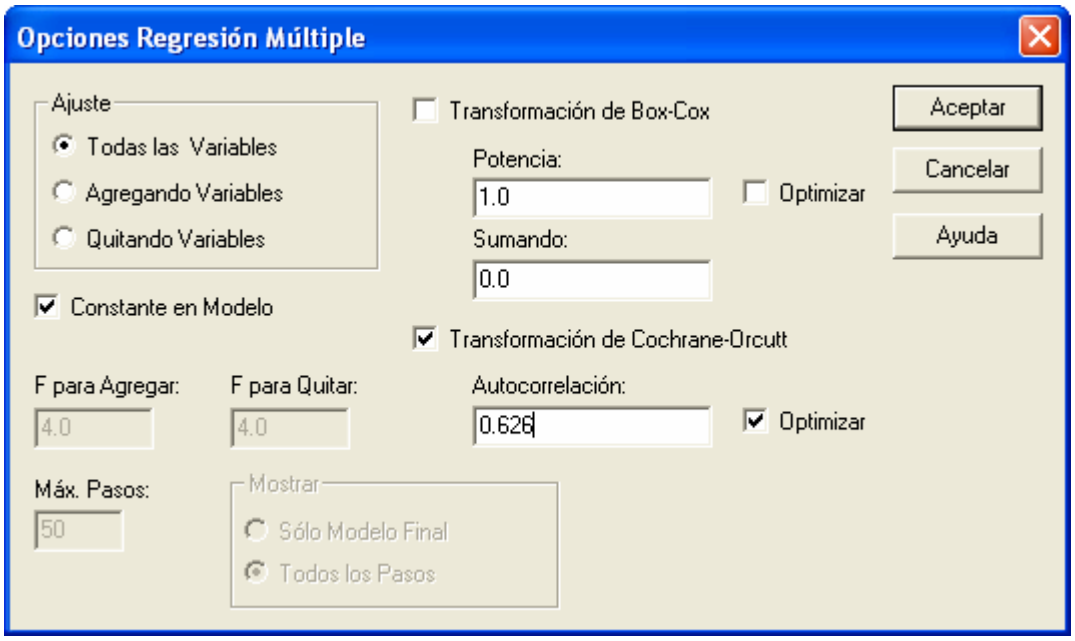
$$x'_t = x_t - \rho x_{t-1} \tag{11}$$

lleva al modelo

$$y'_t = \beta_0 (1-\rho) + \beta_1 x'_t + u_t \tag{12}$$

que es una regresión lineal con términos de error aleatorios

El cuadro de diálogo *Opciones de Análisis* le permite ajustar un modelo de la forma anterior usando el procedimiento de Cochrane-Orcutt:



Usted puede o especificar el valor de ρ en el campo *Autocorrelación*, o seleccionar *Optimizar* y dejar que el valor de ρ sea determinado iterativamente usando el valor especificado como un punto inicial. En el último caso, se usa el siguiente procedimiento:

Paso 1: El modelo es ajustado usando valores transformados de variables basados en el valor inicial de ρ.

Paso 2: El valor de ρ es re-estimado usando los valores de ε_t obtenidos a partir del ajuste del Paso 1.

Paso 3: Los pasos 1 y 2 son repetidos entre 4 y 25 veces hasta que el cambio en el valor derivado de ρ comparado con el paso previo sea menor que 0.01.

Los resultados se resumen a continuación usando los datos de muestra:

Regresión Múltiple - company sales
Variable dependiente: company sales
Variables independientes:
industry sales

Transformación Cochrane-Orcutt aplicada: autocorrelación = 0.91878

		Error	*Estadístico*	
Parámetro	*Estimación*	*Estándar*	*T*	*Valor-P*
CONSTANTE	0.832561	1.11118	0.749258	0.4639
industry sales	0.163206	0.0062571	26.0834	0.0000

Análisis de Varianza

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrada Media*	*Razón-F*	*Valor-P*
Modelo	2.89395	1	2.89395	680.34	0.0000
Residuo	0.0723123	17	0.00425367		
Total (Corr.)	2.96627	18			

R-cuadrada = 97.5622 porciento
R-cuadrado (ajustado para g.l.) = 97.4188 porciento
Error estándar del est. = 0.0652201
Error absoluto medio = 0.0511754
Estadístico Durbin-Watson = 1.69906
Autocorrelación de residuos en retraso 1 = 0.119621

El resultado anterior muestra que, al valor final de $\rho = 0.919$, el estadístico de Durbin-Watson y la autocorrelación lag 1 residual, calculados usando los residuales de la regresión involucrando las variables transformadas, son mucho más en línea que o esperado si los errores fuesen aleatorios. El modelo también cambia de algún modo.

Guardar Resultados

Los siguientes resultados pueden salvarse en la hoja de datos:

1. *Predicciones* – el valor predicho de Y correspondiente a cada una de las *n* observaciones.
2. *Error Estándar de las Predicciones* – los errores estándar para los *n* valores predichos.
3. *Límite Inferior para las Predicciones* – los límites inferiores de predicción para cada valor predicho.
4. L*ímite Superior par alas Predicciones* – los límites superiores de predicción para cada valor predicho.
5. *Error Estándar de las Medias* – los errores estándar para el valor medio de Y en cada uno de los *n* valores de X.
6. *Límite Inferior para las Medias Pronosticadas* – los límites inferiores de confianza para el valor medio de Y en cada uno de los *n* valores de X.
7. *Límite Superior para las Medias Pronosticadas* – los límites superiores de confianza para el valor medio de Y en cada uno de los *n* valores de X.
8. *Residuos* – los *n* residuales.
9. R*esiduos Estudentizados* – los *n* residuales Studentizados.
10. *Influencias* – los valores de influencia correspondientes a los *n* valores de X.
11. *Estadístico DFITS* – el valor del estadístico DFITS correspondiente a los *n* valores de X.
12. *Distancias de Mahalanobis* – la distancia Mahalanobis correspondiente a los *n* valores de X.

Cálculos

Modelo de Regresión

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k \tag{13}$$

Suma de Errores Cuadráticos

No ponderada: $$SSE = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 x_2 - ... - \hat{\beta}_k x_k \right)^2 \tag{14}$$

Ponderada: $$SSE = \sum_{i=1}^{n} w_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 x_2 - ... - \hat{\beta}_k x_k \right)^2 \tag{15}$$

Coeficientes Estimados

$$\hat{\beta} = \left(X'WX \right)^{-1} \left(X'WY \right) \tag{16}$$

$$s^2 \left\{ \hat{\beta} \right\} = MSE \left(X'WX \right)^{-1} \tag{17}$$

$$MSE = \frac{SSE}{n - p} \tag{18}$$

donde $\hat{\beta}$ es un vector columna que contiene los coeficientes estimados de regresión, *X* es una matriz *(n, p)* que contiene un 1 en la primera columna (si el modelo contiene un término constante) y las configuraciones de las *k* variables predictoras en las otras columnas, *Y* es un

vector columna con los valores de la variable dependiente, y W es una matriz diagonal (n, n) que contiene los pesos w_i en la diagonal para una regresión ponderada o 1's en la diagonal si no se especifican los pesos. Se usa un algoritmo de barrida modificado para resolver las ecuaciones luego de centrar y reescalar las variables independientes.

Análisis de Varianza

Con término constante

Fuente	Suma de Cuadrados	Df	Media Cuadrática	F-Radio
Modelo	$SSR = b'X'WY - \frac{\left(\sum_{i=1}^{n} w_i y_i\right)^2}{\sum_{i=1}^{n} w_i}$	k	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$
Residual	$SSE = Y'WY - b'X'WY$	n-k-1	$MSE = \frac{SSE}{n-k-1}$	
Total (corr.)	$SSTO = \sum_{i=1}^{n} w_i (y_i - \bar{y})^2$	n-1		

Sin término constante:

Fuente	Suma de Cuadrados	Df	Media Cuadrática	F-Radio
Modelo	$SSR = b'X'WY$	k	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$
Residual	$SSE = Y'WY - b'X'WY$	n-k	$MSE = \frac{SSE}{n-k}$	
Total	$SSTO = Y'WY$	n		

R-Cuadrada

$$R^2 = 100\left(\frac{SSR}{SSR + SSE}\right)\% \tag{19}$$

R-Cuadrada Ajustada

$$R^2_{adj} = 100\left[1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SSR + SSE}\right]\% \tag{20}$$

Error Estándar de Est.

$$\hat{\sigma} = \sqrt{MSE} \tag{21}$$

Residuales

$$e_i = y_i - \hat{\beta}_o - \hat{\beta}_1 x_1 - ... - \hat{\beta}_k x_k \tag{22}$$

Error Absoluto Medio

$$MAE = \frac{\sum_{i=1}^{n} w_i |e_i|}{\sum_{i=1}^{n} w_i} \tag{23}$$

Estadístico de Durbin-Watson

$$D = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \tag{24}$$

Si $n > 500$, entonces

$$D^* = \frac{|D-2|}{\sqrt{4/n}} \tag{25}$$

se compara con una distribución normal estándar. Para $100 < n \leq 500$, *D/4* se compara con distribución beta con parámetros

$$\alpha = \beta = \frac{n-1}{2} \tag{26}$$

Para muestras más pequeñas, *D/4* se compara con distribución beta con parámetros que se basan en la traza de ciertas matrices relacionadas con las matriz X, como lo describen Durbin y Watson (1951) en la sección 4 de su clásico paper.

Autocorrelación Residual Lag 1

$$r_1 = \frac{\sum_{i=2}^{n} e_i e_{i-1}}{\sum_{i=1}^{n} e_i^2} \tag{27}$$

Influencia

$$h_i = diag\left\{X_i'\left(X'WX\right)^{-1}X_i\right\}w_i \tag{28}$$

$$\overline{h} = \frac{p}{n} \tag{29}$$

Residuales Studentizados

$$d_i = \frac{e_i\sqrt{w_i}}{\sqrt{MSE_i\left(1-h_i\right)}} \tag{30}$$

Distancia Mahalanobis

$$MD_i = \left(\frac{h_i - w_i / \sum_{i=1}^{n} w_i}{1-h_i}\right)\frac{n(n-2)}{n-1} \tag{31}$$

DFITS

$$DFITS_i = \frac{d_i}{\sqrt{w_i}}\sqrt{\left(\frac{h_i}{1-h_i}\right)} \tag{32}$$

Error Estándar para Pronósticos

$$s\left\{Y_{h(new)}\right\} = \sqrt{MSE\left(1 + X_h'\left(X'WX\right)^{-1}X_h\right)} \tag{33}$$

Límite de Confianza para Pronósticos

$$\hat{Y}_h \pm t_{\alpha/2,n-p}\, s\left\{Y_{h(new)}\right\} \tag{34}$$

Límite de Confianza para Medias

$$\hat{Y}_h \pm t_{\alpha/2,n-p}\sqrt{MSE\left(X_h'\left(X'WX\right)^{-1}X_h\right)} \tag{35}$$