Regresión Polinomial

Resumen

El procedimiento **Regresión Polinomial** está diseñado para construir una modelo estadístico que describa el impacto de un solo factor cuantitativo X en una variable dependiente Y. Se ajusta a los datos un modelo polinomial que involucra a X y potencias de X. Se realizan pruebas para determinar el orden apropiado del polinomio. Se puede graficar el modelo ajustado con intervalos de confianza y/o predicción. También se pueden grafican residuos e identificar observaciones influyentes.

StatFolio de Ejemplo: *polynomial reg.sgp*

Datos de Ejemplo:

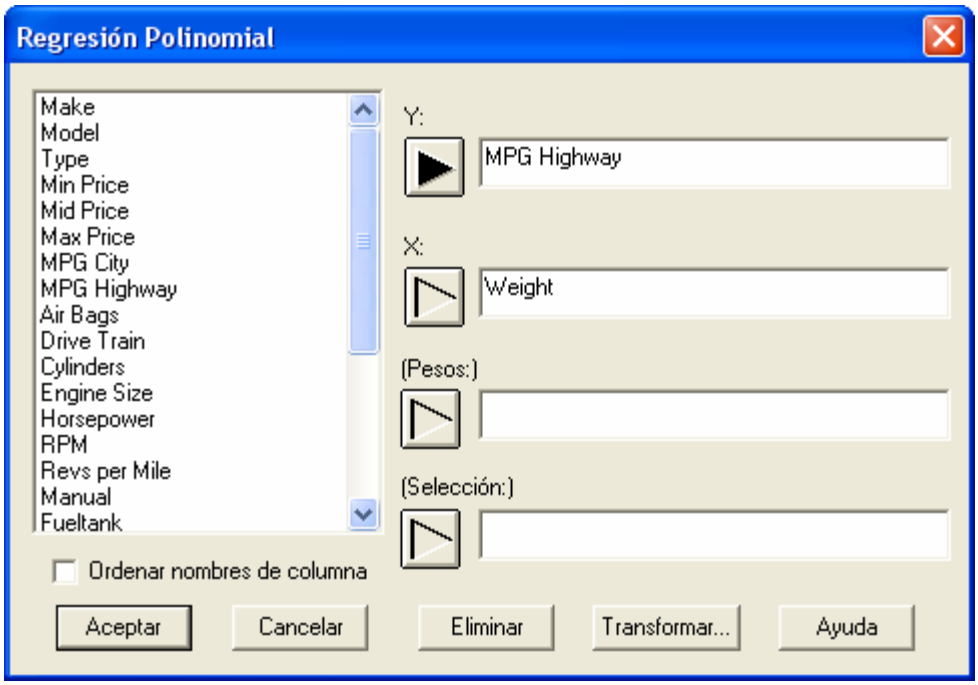
El archivo *93cars.sf3* contiene información de 26 variables para *n* = 93 marcas (*Make*) y modelos (*Model*) de automóviles, tomada de Lock (1993). La tabla a continuación muestra una lista parcial de 4 columnas de ese archivo:

Make	*Model*	*MPG Highway*	*Weight*
Acura	Integra	31	2705
Acura	Legend	25	3560
Audi	90	26	3375
Audi	100	26	3405
BMW	535i	30	3640
Buick	Century	31	2880
Buick	LeSabre	28	3470
Buick	Roadmaster	25	4105
Buick	Riviera	27	3495
Cadillac	DeVille	25	3620
Cadillac	Seville	25	3935
Chevrolet	Cavalier	36	2490

Se desea un modelo que relacione millas por galón en carretera (*MPG Highway*) con el peso de los vehículos (*Weight*).

Ingreso de Datos

La caja de diálogo del ingreso de datos solicita el nombre de las columnas que contienen la variable dependiente Y y la variable independiente X:



- **Y:** columna numérica que contiene las *n* observaciones para la variable dependiente Y.
- **X:** columna numérica que contiene las *n* observaciones para la variable independiente X.
- **Pesos:** una columna numérica opcional que contiene los pesos o ponderadores que se aplicarán al cuadrado de los residuos cuando de realice un ajuste por mínimos cuadrados ponderados.
- **Selección:** selección de un subgrupo de datos.

Resumen de Análisis

El *Resumen del Análisis* muestra información sobre el modelo estimado.

Regresión Polinomial - MPG Highway vs. Weight
Variable dependiente: MPG Highway (miles per gallon in highway driving)
Variable independiente: Weight (pounds)
Orden del polinomio = 2

		Error	*Estadístico*	
Parámetro	*Estimado*	*Estándar*	*T*	*Valor-P*
CONSTANTE	73.8491	7.82234	9.4408	0.0000
Weight	-0.0225792	0.00526637	-4.28744	0.0000
Weight^2	0.00000251567	8.6416E-7	2.91111	0.0045

Análisis de Varianza

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
Modelo	1795.86	2	897.928	98.62	0.0000
Residual	819.455	90	9.10506		
Total (Corr.)	2615.31	92			

R-cuadrada = 68.667 porciento
R-cuadrada (ajustada por g.l.) = 67.9707 porciento
Error estándar del est. = 3.01746
Error absoluto medio = 2.28849
Estadístico Durbin-Watson = 1.71378 (P=0.0789)
Autocorrelación de residuos lag 1 = 0.142564

En la salida se incluyen:

- **Variables y modelo:** identificación de las variables de entrada y el modelo que se ajustó. Por omisión, se ajusta un modelo cuadrático de la forma

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 \qquad (1)$$

  aunque se puede seleccionar un polinomio de diferente orden usando las *Opciones de Análisis*.

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

$$MPG\ Highway = 73.8491 - 0.0225792 * Weight + 0.00000251567 * Weight^2 \qquad (2)$$

  El estadístico t prueba la hipótesis nula de que el parámetro del modelo correspondiente es igual a 0, contra la hipótesis alterna de que no es igual a 0. Valores de P pequeños (menor de 0.05 si se trabaja con un nivel de significancia del 5%) indican que un coeficiente del modelo es significativamente diferente de 0. De particular interés cuando se ajusta un polinomio es el valor de P para el término de mayor orden. Si este término no es significativo, entonces el modelo puede simplificarse razonablemente disminuyendo el orden del polinomio. En los datos de muestra, el valor de P para $Weight^2$ (peso^2) es pequeño, así que se necesita un modelo de al menos orden 2 para describir adecuadamente la relación entre Y y X.

- **Análisis de Varianza:** descomposición de la variabilidad de la variable dependiente Y en una suma de cuadrados del modelo y una suma de cuadrados residual o del error. De particular interés es la prueba de F y su valor de P asociado, que prueban la significancia estadística del modelo ajustado. Un Valor de P pequeño (menor de 0.05 si se trabaja con un nivel de significancia del 5%) indica que existe una relación significativa de la forma especificada entre Y y X. En los datos de muestra, el modelo es altamente significativo.

- **Estadísticas:** estadísticas de resumen para el modelo ajustado, incluyendo:

  R-cuadrada - representa el porcentaje de la variabilidad en Y que ha sido explicado por el modelo de regresión ajustado, que va de 0% a 100%. Para los datos del ejemplo, la regresión ha dado cuenta de alrededor del 68.5% de la variabilidad en las millas por galón. El restante 31.5% es atribuible a la desviación alrededor de la línea, que puede deberse a otros factores, a errores de medición, o a una falla del modelo polinomial actual para ajustar los datos adecuadamente.

  R-cuadrada ajustada – el estadístico R-cuadrada, ajustado para el número de coeficientes en el modelo. Este valor se usa frecuentemente para comparar modelos con diferente número de coeficientes.

  Error Estándar de Est. – La desviación estándar estimada de los residuos (las desviaciones alrededor del modelo). Este valor se usa para crear límites de predicción para nuevas observaciones.

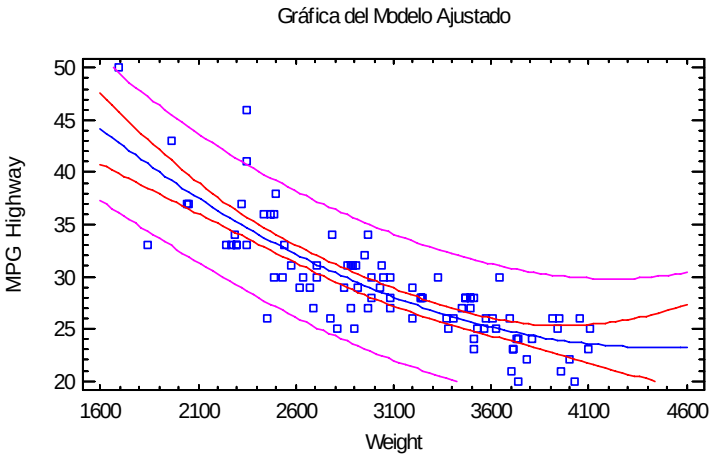
  Error Absoluto Medio – el valor absoluto promedio de los residuos.

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

  Autocorrelación de Residuos Lag 1 – la correlación estimada entre residuos consecutivos, en una escala de –1 a 1. Valores alejados del 0 indican que en el modelo queda estructura significativa sin explicar.

Gráfica del Modelo Ajustado

Esta ventana grafica el modelo ajustado, junco con los límites de confianza y de predicción si así se desea.



La gráfica incluye:

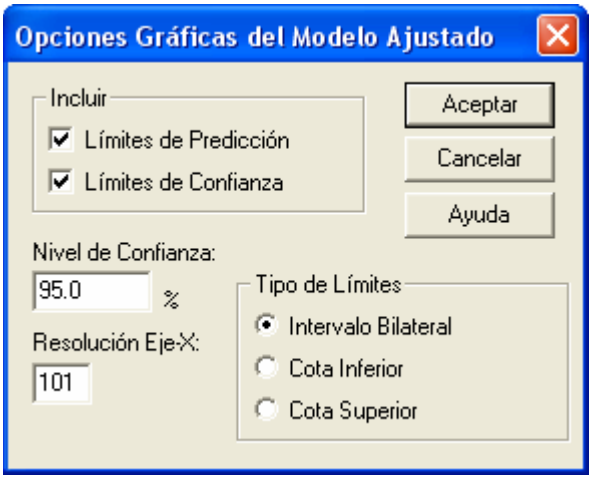
- La línea de mejor ajuste o de **la ecuación de predicción**. Ésta es la ecuación que se usaría para predecir valores de la variable dependiente Y dados valores de la variable independiente X. Advierta que hace un trabajo relativamente bueno recobrando mucha de la relación entre *MPG Highway* (millas por galón en carretera) y *weight* (peso).

- **Intervalos de confianza** para la respuesta media en X. Estos son los límites interiores en el gráfico anterior y describen que tan bien se ha estimado la localización de la línea dada la muestra de datos disponible. Conforme aumenta el tamaño *n* de la muestra, estos límites se harán más estrechos. Debe advertir también que el ancho de los límites varía en función de X, con la línea estimada más precisamente cerca del valor medio $\bar{x}$.

- **Límites de predicción** para nuevas observaciones. Estos son los límites exteriores en el gráfico anterior y describen con que tanta precisión se puede predecir dónde estaría una nueva observación. Sin importar el tamaño de la muestra, las nuevas observaciones variarán alrededor de la verdadera línea con una desviación estándar igual a σ.

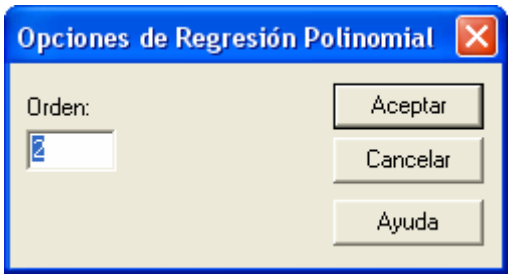
La inclusión de los límites de confianza y de predicción y su nivel de confianza por omisión se determinan por las especificaciones en la pestaña *ANOVA/Regresión* de la caja de diálogo de las *Preferencias*, del menú *Editar*.

Opciones de Ventana



- **Incluir**: los límites a incluir en el gráfico.
- **Nivel de Confianza:** el porcentaje de confianza para los límites.
- **Resolución Eje-X**: el número de valores de X en los cuales se determina la línea al graficar. Mayores resoluciones dan como resultado gráficas más suaves.
- **Tipo de Límites**: si se grafican límites de confianza bilaterales o cotas de confianza unilaterales.

Opciones del Análisis



- **Orden:** el orden del polinomio a ajustar a los datos.

Ejemplo – Ajustando un Polinomio de Tercer Orden

Si se ajusta un polinomio de tercer orden a los datos, los resultados son como los mostrados a continuación:

Regresión Polinomial - MPG Highway vs. Weight

Variable dependiente: MPG Highway (miles per gallon in highway driving)
Variable independiente: Weight (pounds)
Orden del polinomio = 3

		Error	*Estadístico*	
Parámetro	*Estimado*	*Estándar*	*T*	*Valor-P*
CONSTANTE	114.476	31.385	3.64748	0.0004
Weight	-0.0660918	0.0329821	-2.00387	0.0481
Weight^2	0.0000175809	0.0000113068	1.55489	0.1235
Weight^3	-1.69022E-9	1.26487E-9	-1.33628	0.1849

Análisis de Varianza

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
Modelo	1811.97	3	603.991	66.91	0.0000
Residual	803.337	89	9.02626		
Total (Corr.)	2615.31	92			

R-cuadrada = 69.2833 porciento
R-cuadrada (ajustada por g.l.) = 68.2479 porciento
Error estándar del est. = 3.00437
Error absoluto medio = 2.25416
Estadístico Durbin-Watson = 1.68521 (P=0.0615)
Autocorrelación de residuos lag 1 = 0.157148

El modelo ajustado incluye ahora X, X^2, y X^3. Advierta que el Valor de P parar *Weight*^{3} (peso^3) está bastante por arriba de 0.05, indicando que el término de tercer orden *no* es estadísticamente significativo. Esto indica que el modelo de segundo orden probablemente era adecuado para estos datos. Nota: aunque el valor de P para el término de segundo orden no es significativo, no debe asumirse que es innecesario un modelo de segundo orden, ya que el valor de P para *Weight*^{2} (peso^2) cambiará si *Weight*^{3} (peso^3) es removido del modelo. Para seleccionar un orden razonable para el polinomio, vea la ventana *Sumas de Cuadrados Condicionales* descrita a continuación.

Sumas de Cuadrados Condicionales

La ventana *Sumas de Cuadrados Condicionales* presenta una tabla que muestra la significancia estadística de cada coeficiente en el modelo conforme se agrega al ajuste:

ANOVA para las Variables según Orden de Ajuste

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
Weight	1718.7	1	1718.7	188.22	0.0000
Weight^2	77.1615	1	77.1615	8.45	0.0046
Weight^3	16.1176	1	16.1176	1.77	0.1875
Weight^4	7.94288	1	7.94288	0.87	0.3536
Weight^5	0.969712	1	0.969712	0.11	0.7453
Modelo	1820.89	5			

La tabla descompone la suma de cuadrados del modelo SCR en contribuciones debidas a cada coeficiente mostrando el incremento en la SCR conforme se agrega cada término al modelo. Estas sumas de cuadrados frecuentemente son llamadas *Sumas de cuadrados tipo I*. Las razones F comparan el cuadrado medio para cada término con el CME del modelo de mayor orden, en este caso un polinomio de quinto orden. En la tabla anterior, todos los términos más allá del segundo tienen valores de P bastante mayores de 0.05, sugiriendo que un modelo de segundo orden es suficiente para estos datos.

Prueba de Carencia de Ajuste

Cuando se ha registrado más de una observación en el mismo valor de X, se puede realizar una prueba de falta de ajuste para determinar si el modelo elegido describe adecuadamente la relación entre Y y X. La ventana *Carencia de Ajuste* presenta la siguiente tabla:

Análisis de Varianza con Carencia de Ajuste

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
Modelo	1795.86	2	897.928	98.62	0.0000
Residual	819.455	90	9.10506		
Carencia de Ajuste	739.455	78	9.48019	1.42	0.2563
Error Puro	80.0	12	6.66667		
Total (Corr.)	2615.31	92			

La prueba de carencia de ajuste descompone la suma de cuadrados del error en dos componentes:

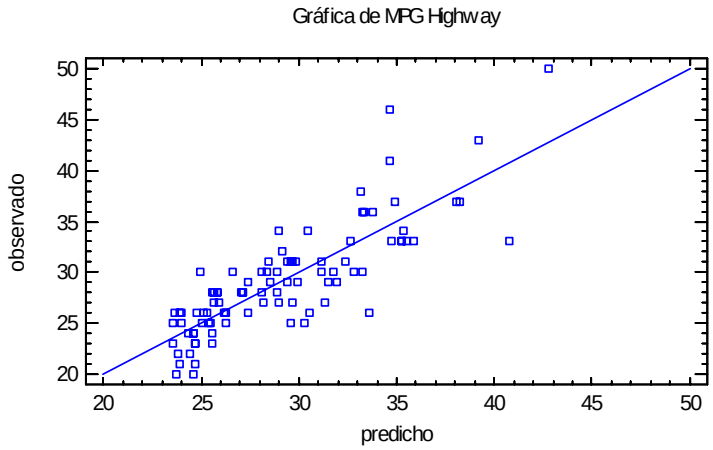
1. *Error puro:* variabilidad de los valores de Y en el mismo valor de X.
2. *Carencia de ajuste:* variabilidad de los valores promedio Y alrededor del modelo ajustado.

De interés primordial es el Valor de P para la carencia de ajuste. Un Valor de P pequeño (menor de 0.05 si se trabaja con un nivel de significancia del 5%) indica que el modelo elegido *no* describe adecuadamente la relación observada.

Para los datos del ejemplo, el valor de P de la carencia de ajuste está muy por arriba de 0.05, indicando que el polinomio de segundo orden explica adecuadamente la relación entre *MPG Highway* (millas por galón en carretera) y *weight* (peso).

Observados Versus Predichos

El gráfico *Observados versus Predichos* muestra los valores observados de Y en el eje vertical y los valores predichos $\hat{Y}$ en el eje horizontal.



Si el modelo ajusta bien, los puntos debieran estar dispersos aleatoriamente alrededor de la línea diagonal. A veces es posible apreciar curvatura en este gráfico, lo que indicaría la necesidad de un polinomio de mayor orden. Cualquier cambio en variabilidad de valores bajos de X a valores altos de X podría también indicar la necesidad de transformar la variable dependiente antes de ajustar un modelo a los datos. En la gráfica anterior, la variabilidad parece aumentar algo conforme los valores predichos son mayores.

Gráficos de Residuos

Al igual que con todos los modelos estadísticos, es una buena práctica examinar los residuos. En una regresión, los residuos se definen por

$$e_i = y_i - \hat{y}_i \qquad (3)$$

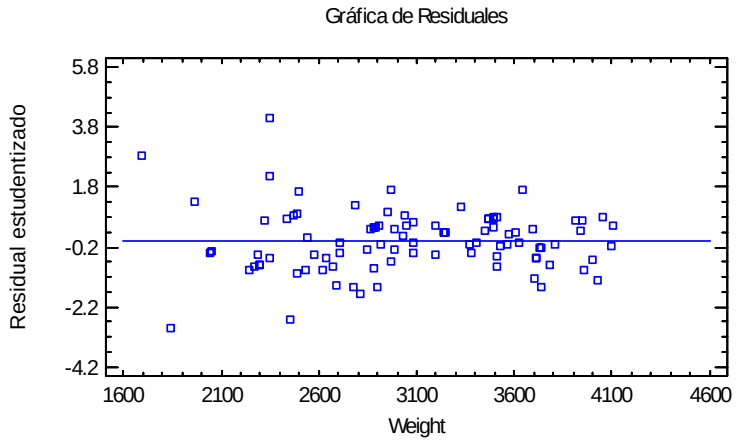
i.e., los residuos son las diferencias entre los valores de los datos observados y el modelo ajustado.

El procedimiento *Regresión Polinomial* crea 3 gráficos de residuos:

1. versus X.
2. versus valor predicho $\hat{Y}$.
3. versus número de fila.

Residuos versus X

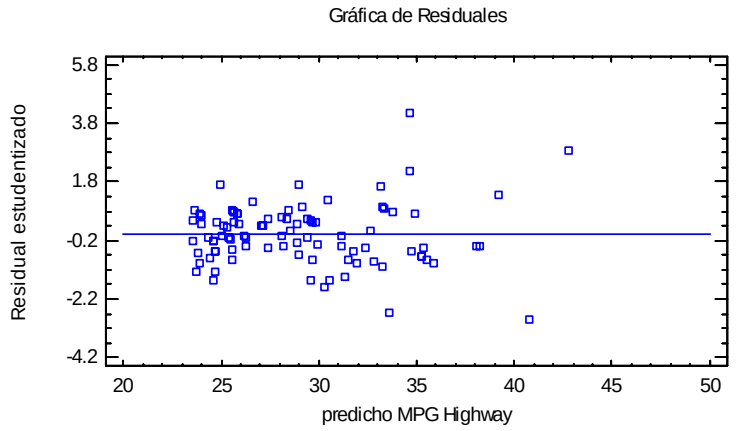
Este gráfico es útil para visualizar cualquier necesidad de un polinomio de mayor orden.



No se detecta curvatura obvia.

Residuos versus Predichos

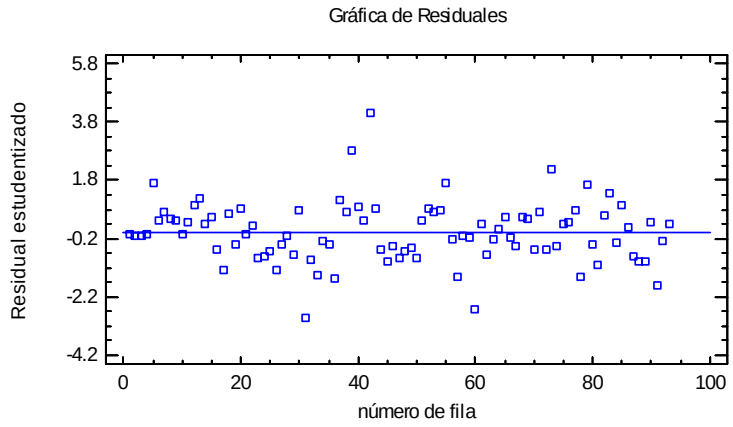
Este gráfico es útil para detectar heteroscedasticidad en los datos.



La heteroscedasticidad ocurre cuando la variabilidad en los datos cambia conforme cambia la media, y puede necesitar transformación de los datos antes de ajustar el modelo de regresión. Generalmente es evidente por un patrón en forma de embudo en el gráfico de los residuos. En el gráfico anterior, se puede ver variabilidad algo mayor en las millas por galón en los valores predichos, que corresponde a los carros pequeños. Para los carros más pequeños, las millas por galón parecen variar más que para carros más grandes.

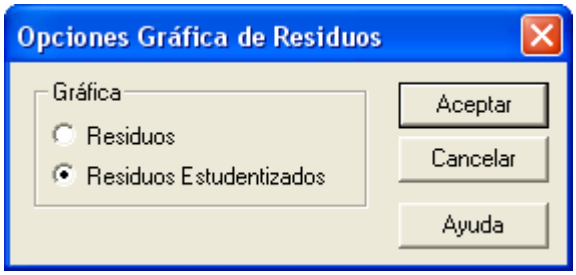
Residuos versus Observación

Este gráfico muestra los residuos versus el número de fila en la hoja de datos:



Si los datos están arreglados en orden cronológico, cualquier patrón en los datos podría indicar una influencia exterior. En el gráfico anterior, no se presenta tendencia obvia, aunque hay un residuo estandarizado superior a 4, indicando que ¡está a más de 4 desviaciones estándar de la curva ajustada!

Opciones de Ventana



Los siguientes residuos pueden graficarse en cada gráfico de residuos:

1. *Residuos* – los residuos del ajuste de mínimos cuadrados.
2. *Residuos Estudentizados* – la diferencia entre los valores observados y_i y los valores predichos $\hat{y}_i$ cuando el modelo es ajustado empleando todas las observaciones excepto la *i-ésima*, dividida entre el error estándar estimado. Estos residuos a veces son llamados *residuos externamente Estudentizados*, ya que miden que tan lejos está cada valor del modelo ajustado cuando tal modelo se ajusta usando todos los datos excepto el punto considerado. Esto es importante, ya que un gran valor atípico podría de otra forma afectar tanto el modelo que no parecería estar inusualmente alejado de la línea.

Residuos Atípicos

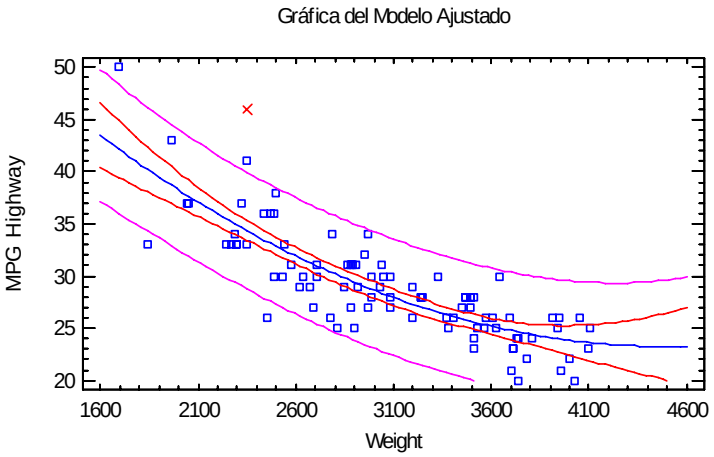
Una vez que el modelo ha sido ajustado, es útil estudiar los residuos para determinar si existe algún valor atípico que debiera ser removido de los datos. La ventana *Residuos Atípicos* lista todas las observaciones que tienen residuos Estudentizados de 2.0 o mayores en valor absoluto.

Residuos Atípicos

		Y		*Residuos*
Fila	*Y*	*Predicha*	*Residuos*	*Estudentizados*
31	33.0	40.7538	-7.75378	-2.90
39	50.0	42.8049	7.19515	2.84
42	46.0	34.6806	11.3194	4.13
60	26.0	33.6302	-7.63024	-2.64
73	41.0	34.6806	6.31936	2.17

Residuos Estudentizados mayores que 3 en valor absoluto corresponden a puntos a más de 3 desviaciones estándar del modelo ajustado, lo que es un evento raro para una distribución normal. En los datos de muestra, la fila #42 esta a más de 4 desviaciones estándar. La fila #42 es un Honda Civic, que en el conjunto de datos dice alcanzar 46 millas por galón, mientras el modelo predijo menos de 35.

Los puntos pueden ser removidos del ajuste mientras se examina el *Gráfico del Modelo Ajustado* haciendo clic sobre un punto y luego presionando el botón *Excluir/Incluir* en la barra de herramientas del análisis:



Los valores excluidos son marcados con una X. Para los datos del ejemplo, remover la fila #42 tiene poco efecto en el modelo ajustado.

Puntos Influyentes

Cuando se ajusta un modelo de regresión, no todas las observaciones tienen la misma influencia en la estimación de los parámetros del modelo ajustado. En una regresión simple, puntos localizados a valores muy bajos o muy altos de X tienen mayor influencia que aquellos localizados más cerca de la media de X. La ventana *Puntos Influyentes* presenta cualquier observación que tenga gran influencia en el modelo ajustado:

Puntos Influyentes

		Distancia de	
Fila	*Influencia*	*Mahalanobis*	*DFITS*
17	0.0728102	6.0785	-0.39756
31	0.152746	15.2366	-1.24208
39	0.243367	27.959	1.94892
60	0.0235077	1.17761	-0.430029
73	0.0290462	1.70335	0.432007
83	0.102398	9.27809	0.558374

Valor de influencia promedio de un solo punto = 0.0326087

Se colocan puntos en esta lista por una de las siguientes razones:

- **Influencia (leverage)** – mide cuán distante está una observación de la media de las *n* observaciones en el espacio de las variables *independientes*. Entre más grande la influencia, mayor el impacto del punto en los valores ajustados $\hat{y}$. Los puntos son colocados en la lista si la influencia es mayor de tres veces la de un punto promedio.

- **Distancia de Mahalanobis –** mide la distancia de un punto a partir del centro de la colección de los puntos en el espacio multivariado de las variables independientes. Dado que esta distancia está relacionada con *la influencia*, no suele seleccionar puntos para la tabla.

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

En los datos de muestra, la fila #39 muestra un valor de leverage de cerca de 8 veces el de un punto promedio de los datos. La fila #39 es un Geo Metro, el carro más ligero en el conjunto de datos.

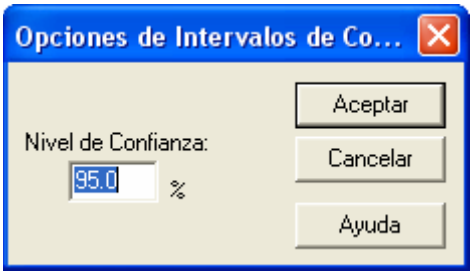
Intervalos de Confianza

La ventana *Intervalos de Confianza* muestra el error de estimación potencial asociado con cada coeficiente en el modelo.

Intervalos de confianza del 95.0% para los coeficientes estimados

		Error		
Parámetro	*Estimado*	*Estándar*	*Límite Inferior*	*Límite Superior*
CONSTANTE	73.8491	7.82234	58.3086	89.3896
Weight	-0.0225792	0.00526637	-0.0330418	-0.0121167
Weight^2	0.00000251567	8.6416E-7	7.98859E-7	0.00000423248

Opciones de Ventana



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

Predicciones

La ventana de *Predicciones* crea predicciones usando modelo ajustado por mínimos cuadrados.

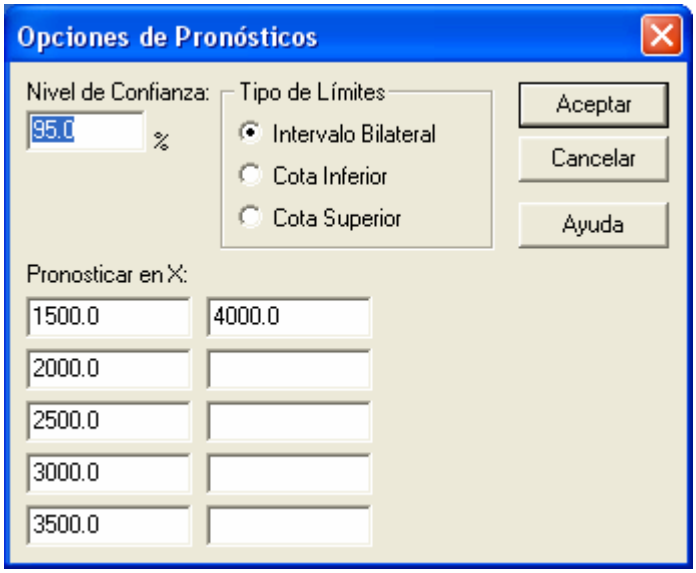
Predicciones

		95.00%		*95.00%*	
	Y	*Inferior*	*Superior*	*Inferior*	*Superior*
X	*Predicha*	*Cota*	*de Predicción*	*Cota*	*de Confianza*
1500.0	44.8754	38.2933	51.4574	41.2947	48.456
2000.0	38.2646	32.4964	44.0328	36.6005	39.9288
2500.0	32.8552	27.2721	38.4382	32.0381	33.6722
3000.0	28.647	23.067	34.2269	27.8511	29.4428
3500.0	25.64	20.0683	31.2118	24.904	26.3761
4000.0	23.8344	18.1311	29.5377	22.4115	25.2573

Se incluyen en la tabla:

- **X** – el valor de la variable independiente en al cual se hará la predicción.
- **Y predicha** - el valor predicho de la variable dependiente usando el modelo ajustado.
- **Cota de Predicción** - límites de predicción para nuevas observaciones al nivel de confianza elegido (corresponde a los límites exteriores en el gráfico del modelo ajustado).
- **Cota de Confianza** - límites de confianza para el valor medio de Y al nivel de confianza elegido (corresponde a los límites interiores en el gráfico del modelo ajustado).

Opciones de Ventana



- **Nivel de Confianza:** porcentaje usado para los intervalos.
- **Tipo de Límites**: si se presentan límites bilaterales o cotas unilaterales.
- **Pronosticar en X**: hasta 10 valores de X en los cuales se harán predicciones.

Salvar Resultados

Se pueden salvar los siguientes resultados en la hoja de datos:

1. *Valores Predichos* – los valores predichos de Y correspondientes a cada una de las *n* observaciones.
2. *Errores Estándar de las Predicciones* – los errores estándar de los *n* valores predichos.
3. *Límites Inferiores para las Predicciones* – los límites inferiores de predicción para cada valor predicho.
4. *Límites Superiores para las Predicciones* – los límites superiores de predicción para cada valor predicho.
5. *Errores Estándar de las Medias* - los errores estándar de los valores medios de Y para cada uno de los *n* valores de X.
6. *Límites Inferiores para el Pronóstico de las Medias* – los límites inferiores de confianza para el valor medio de Y en cada uno de los *n* valores de X.
7. *Límites Superiores para el Pronóstico de las Medias* – los límites superiores de confianza para el valor medio de Y en cada uno de los *n* valores de X.
8. *Residuos* – los *n* residuos.
9. *Residuos Estudentizados* – los *n* residuos Estudentizados.
10. *Leverages* – los valores de las influencias correspondientes a los *n* valores de X.
11. *Estadísticas DFITS* – el valor de las estadísticas DFITS correspondientes a los *n* valores de X.
12. *Distancias de Mahalanobis* – la distancia de Mahalanobis correspondiente a los *n* valores de X.

Note: Si se salvan límites, corresponderán con las especificaciones en la ventana de *Pronósticos*. Si se muestran límites bilaterales en la tabla de Pronósticos, entonces los límites salvados también serán bilaterales. Si se muestran cotas unilaterales en la tabla, entonces los límites salvados también serán unilaterales.

Cálculos

El modelo de regresión polinomial es un caso especial de un modelo de regresión lineal de múltiples variables. Vea la documentación de la *Regresión Múltiple* para detalles con respecto a los cálculos.

