Regresión Binomial Negativa

Resumen

El procedimiento **Regresión Binomial Negativa** está diseñado para ajustar un modelo de regresión en el cual la variable dependiente *Y* consiste de conteos. El modelo de regresión ajustado relaciona *Y* con una o más variables predictoras *X*, que pueden ser cuantitativas o categóricas. El procedimiento ajusta un modelo usando máxima verosimilitud o mínimos cuadrados ponderados. La selección de variables por pasos es una opción. Se realizan pruebas de razón de verosimilitud para probar la significancia de los coeficientes del modelo. El modelo ajustado puede graficarse y generarse predicciones a partir del mismo. Se identifican y grafican residuos atípicos.

Este procedimiento es similar al procedimiento *Regresión Poisson*, excepto se permite que la varianza condicional de *Y* sea mayor que la media. Así que es útil para conteos que están "sobredispersos" comparados con los de un proceso Poisson.

StatFolio de Ejemplo: *Negbin reg.sgp*

Datos de Ejemplo

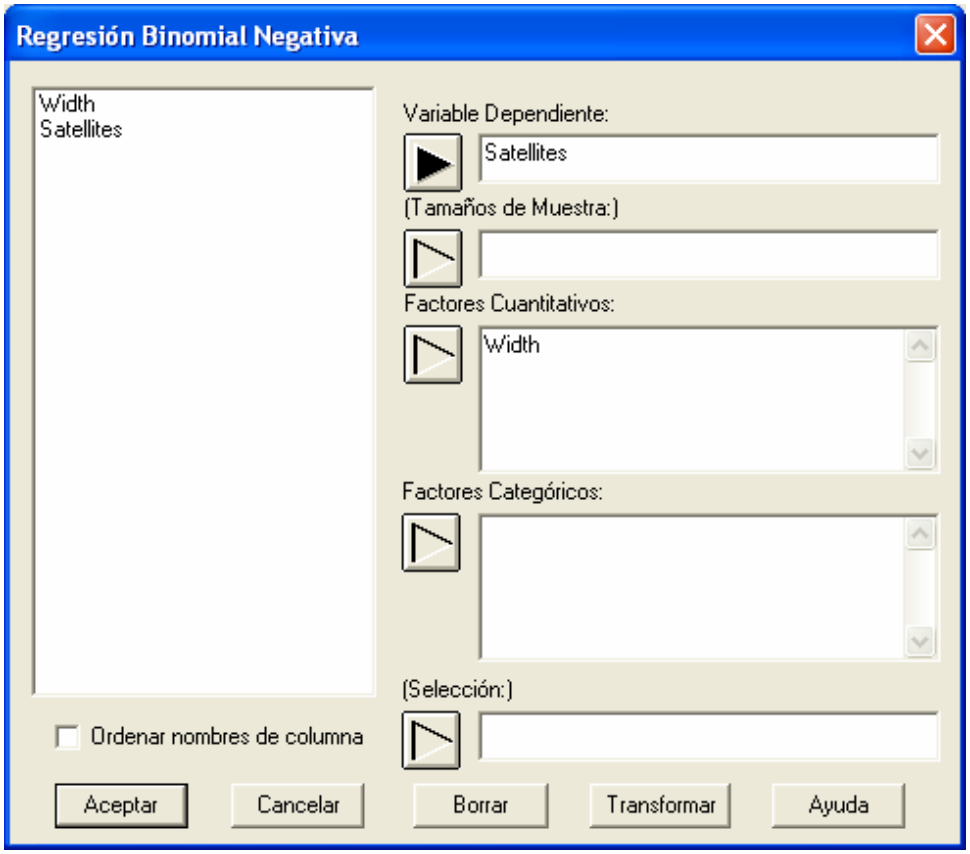
El archivo *crabs.sf6* contiene un conjunto de datos de un estudio de cangrejos herradura, presentado por Agresti (2002). Los datos consisten de información sobre $n = 173$ cangrejos herradura hembra. A continuación se muestra una parte de los datos:

Satellites	*Width*
8	28.3
0	22.5
9	26
0	24.8
4	26
0	23.8
0	26.5
0	24.7
0	23.7
0	25.6
…	…

Se desea relacionar el número de *Satellites* (satélites - cangrejos machos que residen cerca) con el *Width* (ancho) del caparazón de los cangrejos hembra.

Ingreso de Datos

La caja de diálogo del ingreso de datos solicita información sobre las variables de entrada:



- **Variable Dependiente:** variable numérica que contiene los *n* valores de la variable dependiente y_i. *Y* debe consistir de conteos enteros no negativos.

- **(Tamaños de Muestra)**: tamaño de muestra opcional t_i correspondiente a cada cuenta. Si no se especifica, todos los t_i se igualan a 1.

- **Factores Cuantitativos**: columnas numéricas que contienen los valores de cualesquiera factores cuantitativos a ser incluidos en el modelo.

- **Factores Categóricos**: columnas numéricas o no numéricas que contienen los niveles de cualesquiera factores a ser incluidos en el modelo.

- **Selección**: selección de un subgrupo de datos.

Modelo Estadístico

El modelo estadístico asumido para los datos es que los valores de la variable dependiente *Y* siguen una distribución binomial negativa de la forma

$$p(Y) = \frac{\Gamma(Y+\alpha^{-1})}{\Gamma(Y+1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{Y}, \quad \mu > 0,\ \alpha \geq 0 \qquad (1)$$

donde la media μ es el producto de λ, la tasa a la cual ocurren los eventos, y el periodo de muestreo t de acuerdo con

$$E(Y) = \mu = \lambda t \quad (2)$$

La varianza de Y está dada por

$$Var(Y) = \mu + \alpha \mu^2 \quad (3)$$

Si $\alpha = 0$, distribución binomial negativa se reduce a la distribución Poisson. Se supone además que la tasa se relaciona con las variables predictoras a través de una función de enlace log-lineal de la forma

$$\log(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k \quad (4)$$

Resumen del Análisis

El *Resumen del Análisis* presenta una tabla que muestra el modelo estimado y las pruebas de significancia para los coeficientes del modelo. A continuación se muestra una salida típica:

Regresión Binomial Negativa - Satellites
Variable dependiente: Satellites
Factores:
Width

Modelo Estimado de Regresión (Máxima Verosimilitud)

		Error	*Razón de Momios*
Parámetro	*Estimado*	*Estandar*	*Estimada*
CONSTANTE	-4.75395	0.267509	
Width	0.22032	0.00940901	1.24648
Alpha	0.566406		

Análisis de Desviación

Fuente	*Desviación*	*Gl*	*Valor-P*
Modelo	24.0164	1	0.0000
Residuo	275.024	171	0.0000
Total (corr.)	299.04	172	

Porcentaje de desviación explicado por el modelo = 8.03116
Porcentaje ajustado = 6.69355

Pruebas de Razón de Verosimilitud

Factor	*Chi-Cuadrada*	*Gl*	*Valor-P*
Width	24.0164	1	0.0000
Alpha	195.893	1	0.0000

Análisis de Residuos

	Estimación	*Validación*
n	173	
CME	165.698	
MAE	3.27878	
MAPE		
ME	-1.3307	
MPE		

La salida incluye:

- **Resumen de los Datos:** un resumen de los datos que fueron ingresados.

- **Modelo Estimado de Regresión:** estimaciones de los coeficientes del modelo de regresión, con errores estándar y las razones de momios estimadas. Las razones de momios se calculan a partir de los coeficientes del modelo $\hat{\beta}_j$ por medio de

$$\text{razón de momios} = \exp\left(\hat{\beta}_j\right) \tag{5}$$

  La razón de momios representa el incremento porcentual en el momio de los eventos por unidad de incremento en X.

- **Análisis de Desviación:** descomposición de la desviación de los datos en un componente explicado (*Modelo*) y un componente no explicado (*Residuo*). La *Desviación* compara la función de verosimilitud de un modelo con el valor más grande que puede alcanzar la función de verosimilitud, de tal forma que un modelo perfecto tendría una desviación igual a 0. Hay tres renglones en la tabla:

  1. **Total (corr.)** – la desviación de un modelo que contiene únicamente un término constante, $\delta(\beta_0)$.

  2. **Residuo** – la desviación que queda después de haber ajustado el modelo.

  3. **Modelo** – la reducción en la desviación debida a las variables predictoras, $\delta(\beta_1,\beta_2,\ldots,\beta_k|\beta_0)$, igual a la diferencia entre los otros dos componentes.

  El Valor de P para el *Modelo* prueba si el añadir las variables predictoras reduce significativamente la desviación comparada con un modelo que contiene sólo un término constante. Un Valor de P pequeño (menor de 0.05 si se trabaja con un nivel de significancia del 5%) indica que el modelo ha reducido significativamente la desviación y es así útil para predecir a *Y*. El Valor de P para el término *Residuo* prueba si hay una falta de ajuste significativa, i.e., si puede haber un modelo mejor. Un Valor de P pequeño indica que una desviación significativa queda aún en los residuos, así que puede haber un mejor modelo.

- **Porcentaje de Desviación** – el porcentaje de desviación explicada por el modelo, calculada por medio de

$$R^2 = \frac{\delta(\beta_1,\beta_2,...,\beta_k \mid \beta_0)}{\delta(\beta_0)} \tag{6}$$

  Es similar a una estadística R cuadrada en regresión múltiple, en que va de 0% a 100%. También se calcula una desviación ajustada con

$$R^2_{adj} = \frac{\delta(\beta_1,\beta_2,...,\beta_k \mid \beta_0) - 2p}{\delta(\beta_0)} \tag{7}$$

donde *p* es igual al número de coeficientes en el modelo ajustado, incluyendo al término constante. Es semejante a la estadística R-cuadrada ajustada en que compensa el número de variables en el modelo.

- **Pruebas de Razón de Verosimilitud** – una prueba de significancia para cada efecto en el modelo ajustado, y para el parámetro binomial negativo α. Las pruebas para los efectos comparan la función de verosimilitud del modelo completo con la del modelo en el cual sólo el efecto indicado ha sido removido. Valores de P pequeños indican que el modelo ha mejorado significativamente por el efecto correspondiente. La prueba para α compara el modelo ajustado contra un modelo de regresión Poisson, correspondiente a $\alpha = 0$. Un Valor de P pequeño para esa prueba indica que los datos están significativamente sobredispersos (la varianza es mayor que la media).

- **Análisis de Residuos** – si un subgrupo de filas en la hoja de datos han sido excluidas del análisis usando el campo *Seleccionar* en la caja de diálogo de ingreso de datos, el modelo ajustado se usa para hacer predicciones de los valores de *Y* para estas filas. Esta tabla muestra estadísticas sobre los errores de predicción, definidos por

$$e_i = y_i - \hat{\mu}_i \tag{8}$$

  Se incluyen el cuadrado medio del error (CME), el error absoluto medio (EAM), el error porcentual absoluto medio (EPAM), el error medio (EM), y el error porcentual medio (EPM). Estas estadísticas de validación pueden ser comparadas con las estadísticas del modelo ajustado para determinar qué tan bien el modelo predice las observaciones fuera de los datos usados para ajustarlo.

El modelo ajustado para los datos del ejemplo es

$$\hat{\lambda} = \exp(-4.75395 + 0.22032 Width) \tag{9}$$

La regresión explica un poco más del 8% de la desviación de un modelo con sólo una constante. El Valor de P para *Width* es muy pequeño, indicando que es un predictor significativo de *Satellites*.

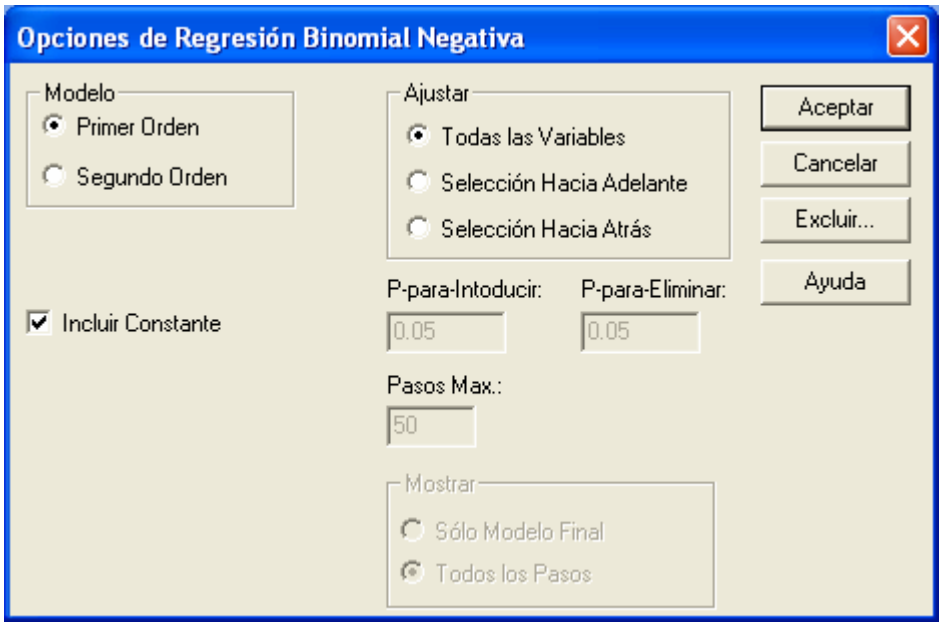
La media y varianza de *Satellites* están dadas por:

$$E(Satellites) = \hat{\mu} = \exp(-4.75395 + 0.22032 Width) \tag{10}$$

$$Var(Satellites) = \hat{\mu}(1 + 0.566406\hat{\mu}) \tag{11}$$

Además, el Valor de P para *Alpha* es muy pequeño, indicando que hay sobredispersión significativa, así que un modelo Poisson no sería apropiado para estos datos.

Opciones de Análisis



- **Modelo**: orden del modelo a ser ajustado. Los modelos de primer orden incluyen solo efectos principales. Los modelos de segundo orden incluyen efectos cuadráticos para los factores cuantitativos e interacciones de dos factores entre todas las variables.

- **Incluir Constante**: Si no se marca esta opción, el término constante β_0 será omitido del modelo.

- **Ajustar**: especifica si todas las variables independientes especificadas en caja de diálogo del ingreso de datos deben ser incluidas en el modelo final, o si se debe aplicar una selección por pasos de las variables. La selección por pasos intenta encontrar un modelo parsimonioso que contenga sólo variables significativas estadísticamente. Un ajuste *por Pasos hacia Adelante* comienza sin variables en el modelo. Un ajuste *por Pasos hacia Atrás* comienza con todas las variables en el modelo.

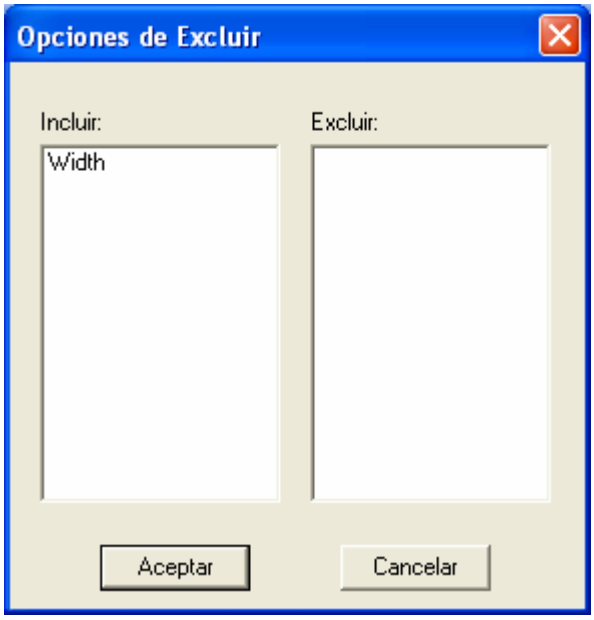
- **P-para-Introducir** - En un ajuste por pasos, las variables entrarán al modelo en un paso dado si sus valores de P son menores o iguales al valor especificado de *P-para-Introducir*.

- **P-para-Eliminar** - En un ajuste por pasos, las variables serán removidas del modelo en un paso dado si sus valores de P son mayores que el valor especificado de *P-para-Eliminar*.

- **Pasos Max:** máximo número de pasos permitidos cuando se lleva a cabo un ajuste por pasos.

- **Mostrar**: si se muestran los resultados en cada paso cuando se lleva a cabo un ajuste por pasos.

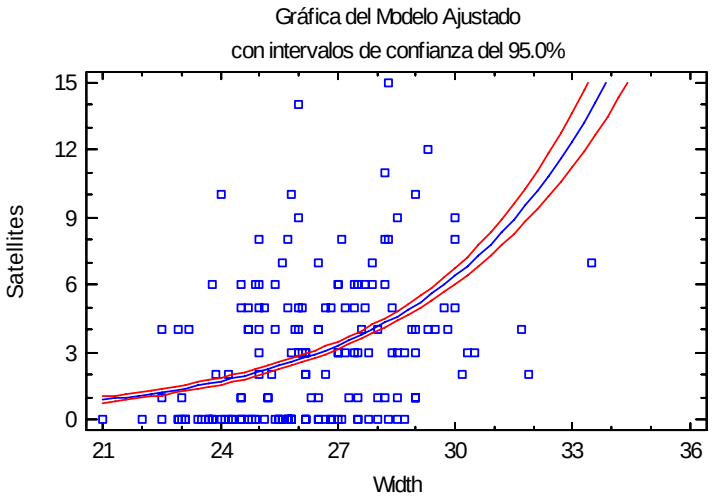
- **Excluir**: Presione este botón para excluir efectos del modelo. Se mostrará una caja de diálogo:



Haga doble clic sobre un efecto para moverlo del campo *Incluir* al campo *Excluir* o para regresarlo.

Gráfica del Modelo Ajustado

El *Gráfica del Modelo Ajustado* muestra la tasa media estimada $\hat{\lambda}(X)$ versus cualquier variable predictora, manteniendo constantes las otras variables.



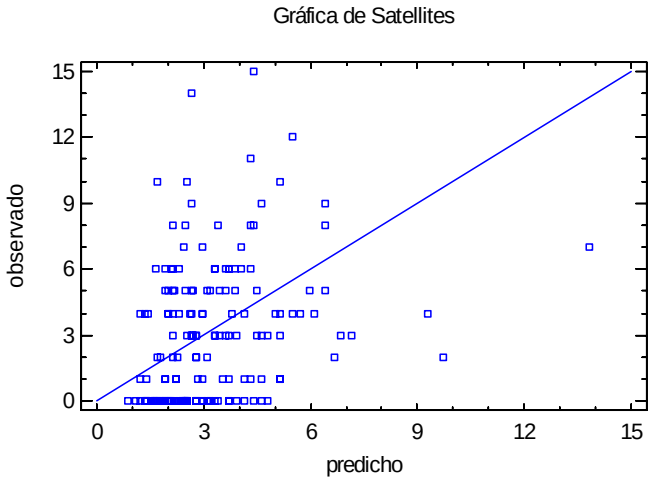
Se incluyen en el gráfico los límites de confianza para $\lambda(X)$. El número medio estimado de satélites (*Satellites*) aumenta de un bajo número de aproximadamente 1 en anchos (*Width*) pequeños a más de una docena en anchos grandes.

Opciones de Ventana

- **Factor:** selecciona el factor a graficar en el eje horizontal.
- **Bajo y Alto:** especifica el rango de valores para el factor seleccionado.
- **Mantener:** valores en los que se mantendrán fijos los factores no seleccionados.
- **Nivel de Confianza:** porcentaje usado para los límites de confianza. Poner en 0 para omitir los límites.

Observados Versus Predichos

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



Si el modelo ajusta bien, los puntos deben estar esparcidos aleatoriamente alrededor de la línea diagonal.

Predicciones

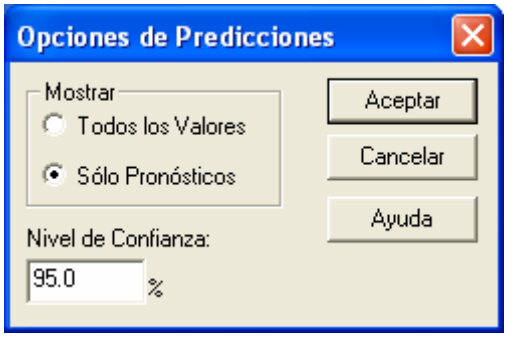
El modelo de regresión ajustado puede usarse para predecir el resultado de nuevas muestras cuyas variables predictoras son dadas. Por ejemplo, suponga que se desea una predicción para un cangrejo con *Width* = 30.3. Se puede agregar una nueva fila a la hoja de datos con 30.3 en la columna *Width*, pero la entrada para *Satellites* se dejaría en blanco. La ventana de *Predicciones* mostraría entonces:

Predicciones para Satellites

	Observado	*Ajustado*	*LC Inferior 95.0%*	*LC Superior 95.0%*
Fila			*para Predicción*	*para Predicción*
174		6.83284	6.44308	7.24619

La tabla muestra el valor ajustado $\hat{\mu}_i$, junto con intervalos de confianza aproximados del 95%.

Opciones de Ventana



- **Mostrar**: muestra *Todos los Valores* (predicciones para todas las filas en la hoja de datos), o *Sólo Pronósticos* (predicciones para las filas con valores faltantes para *Y*).

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

Intervalos de Confianza

La ventana *Intervalos de Confianza* muestra el error de estimación potencial asociado con cada coeficiente en el modelo, así como para las razones de tasas.

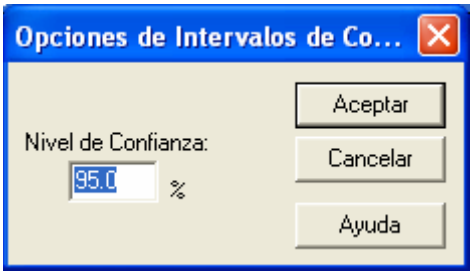
Intervalos de confianza del 95.0% para los estimados de los coeficientes

		Error		
Parámetro	*Estimado*	*Estándar*	*Límite Inferior*	*Límite Superior*
CONSTANTE	-4.75395	0.267509	-5.27826	-4.22965
Width	0.22032	0.00940901	0.201879	0.238761

Intervalos de confianza del 95.0% para la razón de tasas

Parámetro	*Estimado*	*Límite Inferior*	*Límite Superior*
Width	1.24648	1.2237	1.26968

Opciones de Ventana



- **Nivel de Confianza:** nivel porcentual para los intervalos de confianza.

Matriz de Correlación

La *Matriz de Correlación* muestra estimaciones de la correlación entre los coeficientes estimados.

Matriz de correlación para los coeficientes estimados

	CONSTANTE	Width
CONSTANTE	1.0000	-0.9961
Width	-0.9961	1.0000

Esta tabla puede ser útil para determinar que tan bien se han separado unos de otros los efectos de las variables independientes.

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 grandes atípicos.

Residuos Atípicos para Satellites

		Y		*Residuo*	*Residuo de*
Fila	*Y*	*Predicha*	*Residuo*	*Pearson*	*Desviación*
3	9.0	2.64948	6.35052	2.47	1.68
15	14.0	2.64948	11.3505	4.41	2.55
34	8.0	2.48003	5.51997	2.26	1.57
56	15.0	4.39778	10.6022	2.71	1.80
121	6.0	1.63176	4.36824	2.47	1.67
131	6.0	1.90386	4.09614	2.06	1.46
134	10.0	2.53527	7.46473	3.00	1.94
146	8.0	2.12558	5.87442	2.71	1.80
149	10.0	1.70527	8.29473	4.53	2.58

La tabla muestra:

- **Fila** – el número de fila en la hoja de datos.
- **Y** – el valor observado de Y.
- **Y Predicha** – el valor ajustado $\hat{\mu}_i$.
- **Residuo** – la diferencia entre los valores observado y predico, definido por

$$e_i = y_i - \hat{\mu}_i \qquad (12)$$

- **Residuo de Pearson** – un residuo estandarizado en el cual cada residuo es divido entre una estimación de su error estándar:

$$r_i = \frac{e_i}{\sqrt{\hat{\mu}_i\left(1+\hat{\alpha}_i\hat{\mu}_i\right)}} \qquad (13)$$

- **Residuo de Desviación –** un residuo que mide la contribución de cada observación a la desviación de los residuos:

$$d_i = \text{sgn}(r_i)\sqrt{2\left\{y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - \left(y_i + \hat{\alpha}^{-1}\right)\ln\left[\left(y_i + \hat{\alpha}^{-1}\right)/\left(\hat{\mu}_i + \hat{\alpha}^{-1}\right)\right]\right\}} \qquad (14)$$

  La suma de los residuos de desviación cuadrados es igual a la desviación en el renglón de *Residuos* en la tabla del análisis de desviación.

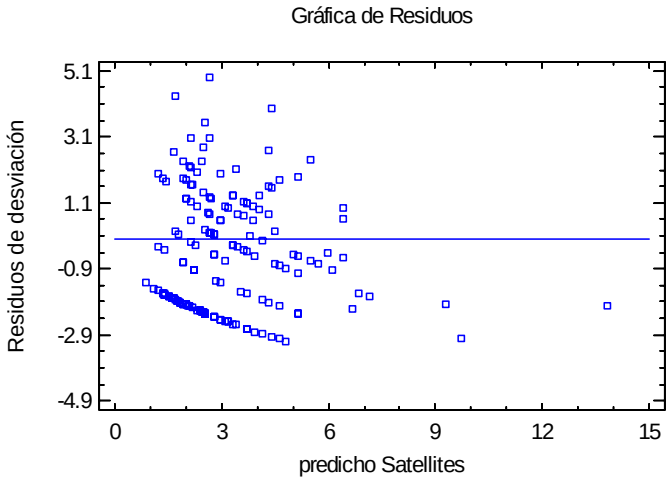
La tabla incluye todas las filas para las cuales el valor absoluto del residual de Pearson es mayor que 2.0. El presente ejemplo muestra 9 residuos que exceden 2.5, 2 de los cuales exceden 3.0.

Gráficas de Residuos

Al igual que con todos los modelos estadísticos, es una buena práctica examinar los residuos. El procedimiento *Regresión Binomial Negativa* incluye varios tipos de gráficas de residuos, dependiendo de las *Opciones de Ventana*.

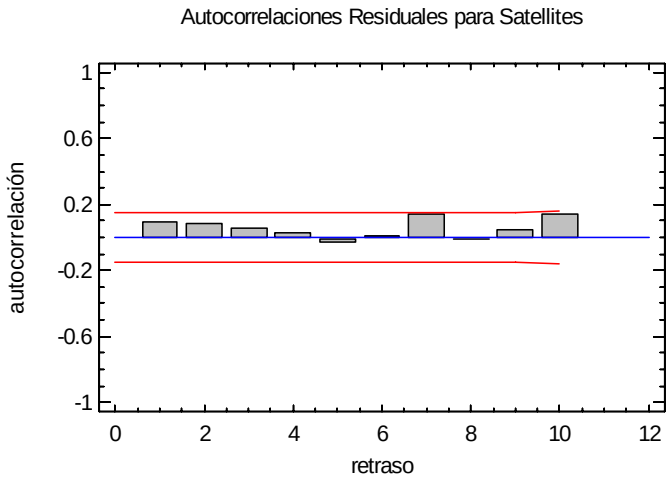
Gráfico de Dispersión versus Valor Predicho

Este gráfico es útil para examinar residuos muy grandes.



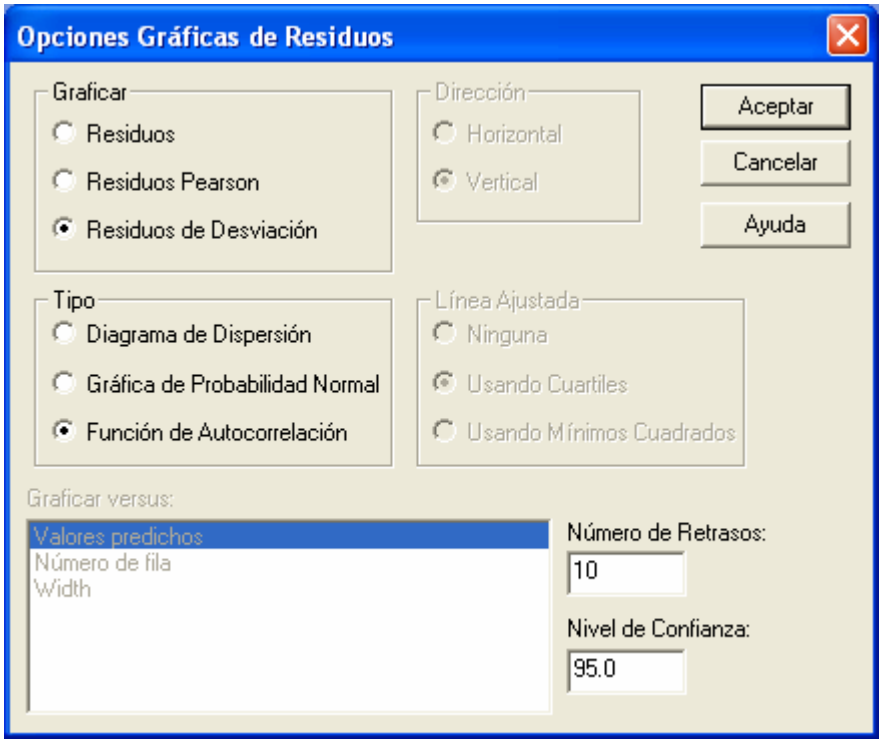
Autocorrelaciones entre Residuos

Este gráfico calcula la autocorrelación entre residuos como una función del número de filas entre ellos en la hoja de datos.



Esto sólo es relevante si los datos se colectaron secuencialmente. Cualquier barra extendiéndose más allá de los límites de probabilidad indicaría dependencia significativa entre residuos separados por el retraso indicado.

Opciones de Ventana



- **Gráfico:** el tipo de residuos a graficar:

    1. *Residuos* – los valores observados menos los ajustados.
    2. *Residuos Pearson* – los residuos divididos entre sus errores estándar estimados.
    3. *Residuos de Desviación* – residuos escalados de tal forma que la suma de sus cuadrados es igual a la desviación de los residuos.

- **Tipo:** el tipo del gráfico a crear. Se usa un *Diagrama de Dispersión* para probar curvatura. Se emplea una *Gráfica de Probabilidad Normal* para determinar si los residuos del modelo provienen de una distribución normal (no se espera normalidad en este procedimiento). Se usa una *Función de Autocorrelación* para probar dependencia entre residuos consecutivos.

- **Graficar Versus**: para un *Diagrama de Dispersión*, la cantidad a graficar en el eje horizontal.

- **Número de Retrasos**: para una *Función de Autocorrelación*, el máximo número de retrasos. Para grupos pequeños de datos, el número de retrasos graficados puede ser menor que este valor.

- **Nivel de Confianza:** para una *Función de Autocorrelación*, el nivel usado para crear los límites de probabilidad.

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. Aquellos con valores atípicos de las variables independientes tienden a tener mayor influencia que los otros. La ventana *Puntos Influyentes* presenta cualquier observación que tenga gran influencia en el modelo ajustado:

Puntos Influyentes para Satellites

Fila	*Leverage*
115	*0.113515*
141	0.393344
142	0.0362742
147	0.0966221

Leverage promedio de un solo punto = 0.0115607

La tabla muestra todos los puntos con *punto leverage* alto. El punto leverage es una estadística que mide cuán distante está una observación de la media de las *n* observaciones en el espacio de las variables *independientes*. Entre más grande el punto leverage, mayor el impacto del punto en los valore ajustados $\hat{y}$. Los puntos son colocados en la lista si su leverage es mayor de tres veces el de un punto promedio

La observación con el punto leverage más alto en los datos de ejemplo es la de la fila #141. Es más de 30 veces el leverage promedio, lo que significa que tiene una mayor influencia en el ajuste.

Salvar Resultados

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

1. *Valores Predichos* – los valores ajustados $\hat{\lambda}_i t_i$ correspondientes a cada fila de la hoja de datos.
2. *Límites Inferiores* – los límites inferiores de confianza para $\hat{\lambda}_i t_i$.
3. *Límites Superiores* – los límites superiores de confianza para $\hat{\lambda}_i t_i$.
4. *Residuos* – los residuos ordinarios.
5. *Residuos Pearson* – los residuos estandarizados de Pearson.
6. *Residuos de Desviación* – los residuos de desviación.
7. *Leverages* – los puntos leverage para cada fila.

Cálculos

Sea μ_i = la media estimada para los valores de las variables predictoras en la fila *i*.

Función de Verosimilitud

$$L = \prod_{i=1}^{n} \frac{\Gamma\left(y_i + \alpha^{-1}\right)}{\Gamma\left(y_i + 1\right)\Gamma\left(\alpha^{-1}\right)} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\right)^{\alpha^{-1}} \left(\frac{\mu_i}{\alpha^{-1} + \mu_i}\right)^{y_i} \quad \text{para } \mu > 0,\ \alpha \geq 0 \tag{15}$$

Desviación

$$\delta(\hat{\beta} \mid \alpha) = \sum_{i=1}^{n} \left\{ y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - \left(y_i + \alpha^{-1}\right) \ln\left[\frac{y_i + \alpha^{-1}}{\hat{\mu}_i + \alpha^{-1}}\right] \right\} \tag{16}$$

Punto Leverage

$$h_i = diag\left\{X_i'\left(X'WX\right)^{-1} X_i\right\} w_i \tag{17}$$

$$\overline{h} = \frac{p}{n} \tag{18}$$