Índice general

Capítulo 4

Regresión Simple

El objetivo del análisis de la regresión simple es estudiar un modelo o función (f) que pretende explicar el comportamiento de una variable (variable endógena, explicada o dependiente), que denotaremos por Y, utilizando la información proporcionada por los valores tomados por otra variable (variable exógena, explicativa o independiente), que denotaremos por X.

Si representamos la nube de puntos de las dos variables bajo estudio (X, Y), la regresión simple consiste en encontrar la curva (f) que, aunque no pase por todos los puntos de la nube, al menos esté lo más próxima posible a ellos.

Utilizando la función considerada, podríamos predecir un valor de la variable Y conocido un valor de la variable X: $\widehat{y}_i = f(x_i)$. Dicho valor no tiene porque coincidir con el que toma la variable Y, y_j, y por tanto se estaría cometiendo un error en dicha predicción:

$$e_{ij} = y_j - \widehat{y}_i$$

el cual se denomina residuo.

Los residuos nos dan una medida del error cometido en el ajuste del modelo a los datos de que se dispone. En particular, el criterio de mínimos cuadrados para el ajuste de una función a unos datos se basa en minimizar la suma de los residuos al cuadrado.

El paquete estadístico Statgraphics analiza ampliamente este tema aportando procedimientos para la regresión simple tanto lineal como no lineal, incluyendo así el posible análisis de una amplia variedad de modelos de regresión con una sola variable independiente. Para el análisis de los modelos se basa en el ajuste por mínimos cuadrados.

Los modelos que se pueden estudiar son:

- Lineal: $Y = a + bX$
- Exponencial: $Y = e^{a+bX}$ o $\ln Y = a + bX$
- Recíproco$-X$: $Y = a + b/X$
- Logarítmico$-X$: $Y = a + b \log X$
- Raíz cuadrada$-Y$: $Y = (a + bX)^2$
- Logístico: $Y = \exp(a + bX)/(1 + \exp(a + bX))$
- Multiplicativo: $Y = aX^b$ o $\ln Y = \ln a + b \ln X$
- Recíproco$-Y$: $1/Y = a + bX$
- Doble Recíproco: $Y = 1/(a + bX)$
- Raíz cuadrada$-X$: $Y = a + b\sqrt{X}$
- S-curva: $Y = \exp(a + b/X)$
- Log Probit: $Y = normal(a + b \log X)$

El sistema nos permite representar el modelo ajustado y los residuos para todos los modelos, generar y representar las predicciones para valores dados de X o Y, así como guardar las predicciones y los residuos para posteriores análisis.

Para poder empezar a estudiar como realizar un análisis de regresión simple sobre dos variables con el paquete Statgraphics, lo primero es cargar en memoria el fichero de datos que contiene las dos variables

que queremos analizar. Para ello, recordemos que debemos, tras abrir el programa, seleccionar **Archivo / Abrir / Abrir Datos**.

A lo largo de todo este tema trabajaremos con el archivo de datos *miempresa.sf3*, los cuales incluyen un estudio sobre los empleados de una empresa. En particular vamos a analizar una regresión simple de las variables X =*Salari06* e Y =*Salari07*, es decir vamos a estudiar si existe una función o modelo que me permita predecir, de forma más o menos correcta, el salario de un empleado en el año 2007 conocido su salario en el año 2006. También realizaremos el estudio cambiando la variable Y a la variable *Seguro07* para poder profundizar en varias opciones.

	Id	Sexo	FechNac	Educ	CatLab	Salari06
1	13	h	12.06.1984	1	A	1535
2	26	h	21.03.1984	1	A	1340
3	34	m	13.11.1981	1	A	1420

Figura 4.1: Ventana de datos

Para llevar a cabo el análisis de regresión simple vamos a seguir una serie de pasos marcados por las siguientes secciones.

4.1. Nube de puntos. Covarianza y coeficiente de correlaciones

4.1.1. Nube de puntos

Una vez cargado el fichero de datos (figura 4.1), para poder comenzar con el análisis de regresión, debemos primero comprobar si gráficamente se observa que exista relación entre las dos variables que queremos estudiar y, si es posible, determinar que modelo parece más adecuado para el ajuste. Para ello debemos comenzar representando la nube de puntos o gráfico de dispersión de las dos variables. Dicho gráfico se obtiene seleccionando **Gráficos / Gráficos de Dispersión / Gráficos X-Y** (figura 4.2) o haciendo clic en el icono  .

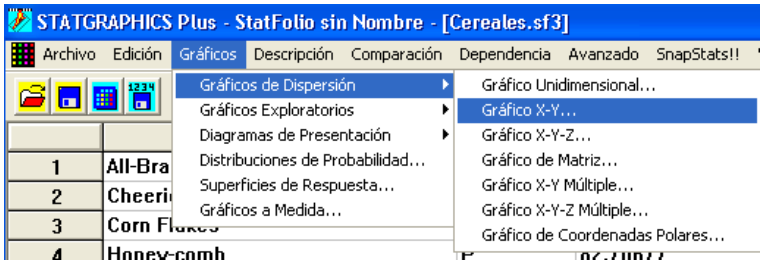

Figura 4.2: Gráficos X-Y en el menú Gráficos

Al hacerlo aparece un cuadro de diálogo (figura 4.3) donde deberemos seleccionar las variables que queremos representar en el gráfico: X =*Salari06* e Y =*Salari07*.

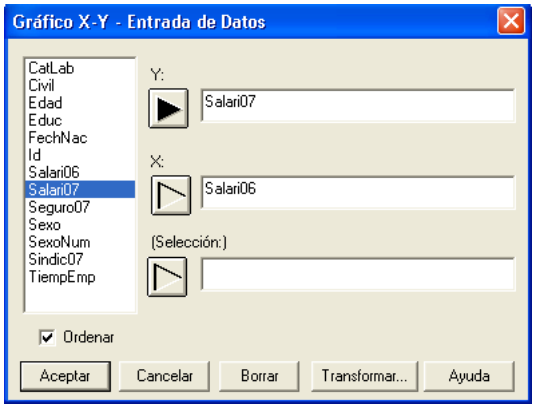

Figura 4.3: Entrada de datos de Gráficos X-Y

Una vez seleccionadas las variables, haciendo clic en **Aceptar** aparece la ventana de resultados (figura 4.4), la cual muestra a la izquierda el resumen del procedimiento y a la derecha el gráfico deseado.

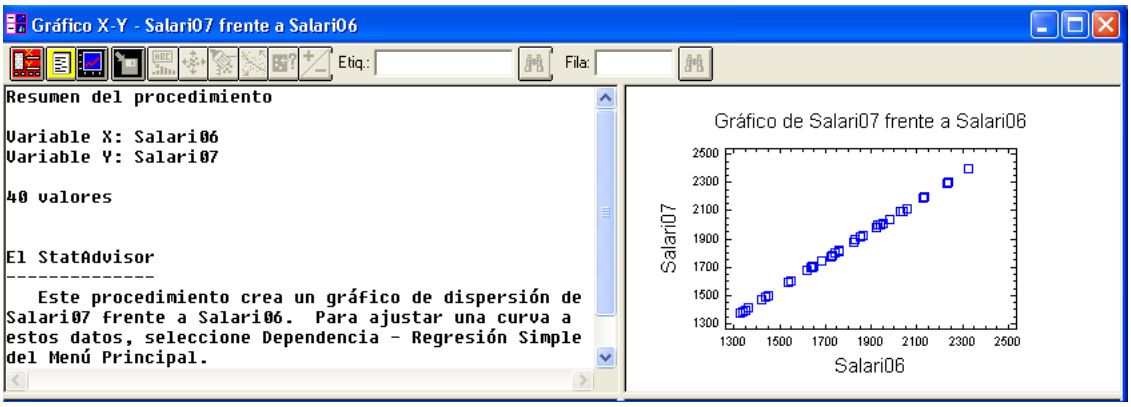

Figura 4.4: Ventana de resultados de Gráficos X-Y

Si queremos ver mejor el gráfico, basta con hacer clic dos veces consecutivas sobre el mismo. El gráfico de dispersión, al igual que los gráficos ya estudiados, se puede editar haciendo clic con el botón derecho sobre él y seleccionando las **Opciones Gráficas**. Un ejemplo podría ser modificación del gráfico que aparece en la figura (4.5).

En este caso, el gráfico resultante muestra una nube de puntos concentrada y lineal lo que nos hace pensar que existe una fuerte relación lineal entre las dos variables y que, por tanto, el ajuste lineal es el ideal para estas dos variables. Antes de pasar a realizar dicho ajuste vamos a estudiar en detalle las posibilidades que presenta el paquete Statgraphics en el gráfico de dispersión o nube de puntos. Cualquiera de las siguientes aplicaciones que permite realizar este gráfico sólo están activas si el gráfico ha sido seleccionado o editado, es decir, si se ha ampliado de la ventana de resultados haciendo doble clic sobre él.

La primera opción interesante es la de poder identificar cualquiera de los puntos del gráfico. Si deseamos saber que empleado es el representado por uno de los puntos y que valores toma para las variables

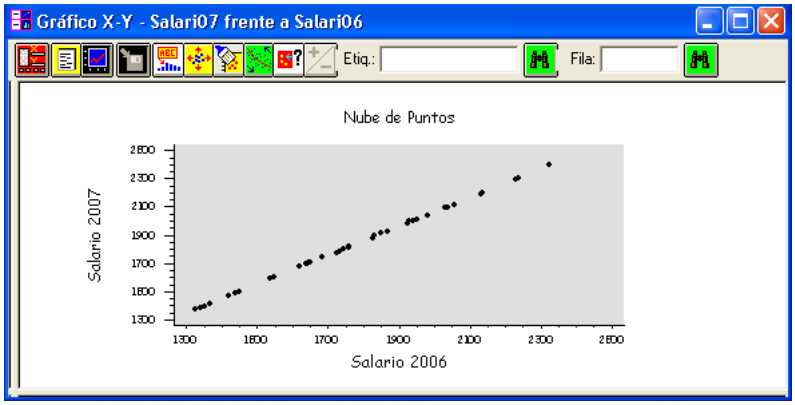

Figura 4.5: Ejemplo de gráfico X-Y tras modificarlo

bajo estudio, basta hacer clic sobre él dejando pulsado el ratón. De esta forma aparecerá a la derecha una ventana pequeña mostrando la información (figura 4.6).

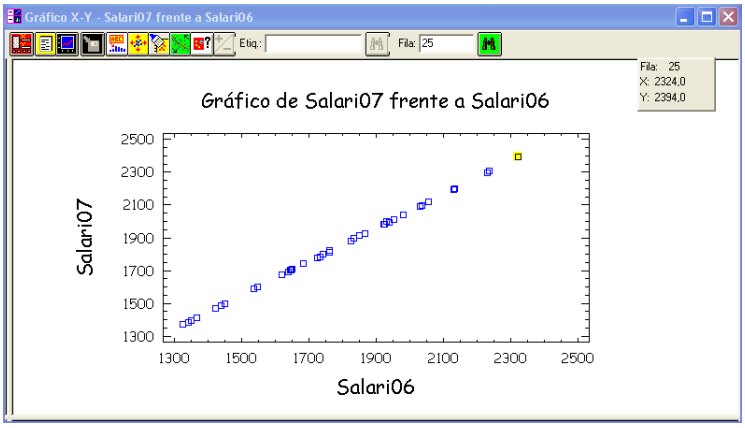

Figura 4.6: Ejemplo de identificación de dato en el gráfico X-Y

La segunda aplicación que puede se útil es la de resaltar puntos, es decir, el programa permite distinguir aquellos puntos que están por encima y por debajo del valor medio con respecto a una variable externa de tipo numérico. Por ejemplo, si queremos saber cuales de los empleados representados en la nube de puntos tiene un nivel educativo superior a la media y cuales inferior, para ver si ello afecta a los salarios, debemos: hacer clic sobre el icono  , introducir la variable *Educ* (figura 4.7) y hacer clic en **Aceptar**. Esta opción colorea de rojo los puntos correspondientes a empleados cuyo nivel educativo es inferior a la media de dicha variable, es decir inferior a 1.5, y de azul al resto (figura 4.8).

Además de distinguir entre los puntos aparece una barra en la parte superior de la ventana (figura 4.9) que nos permite modificar el valor que queramos usar de referencia en la división, y que originalmente esta en 1.5 como ya hemos visto.

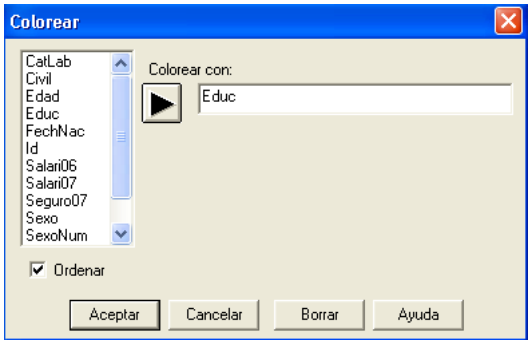

Figura 4.7: Opción de Colorear

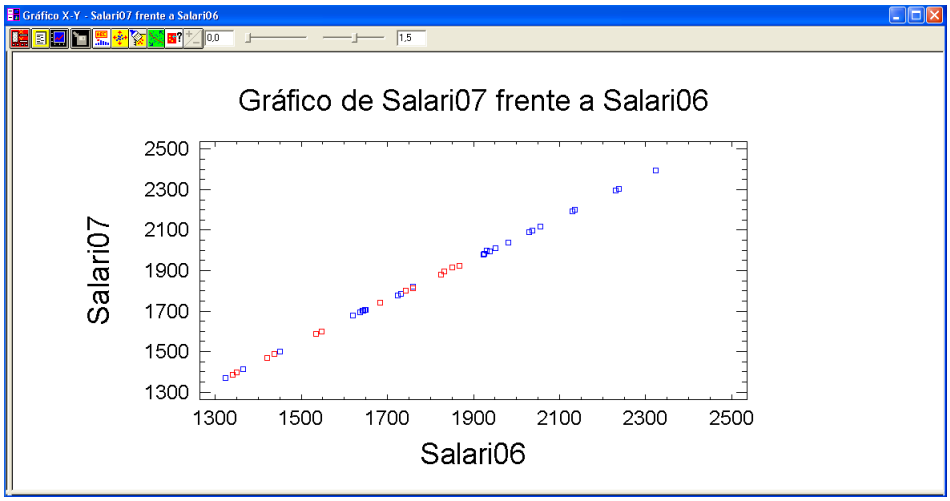

Figura 4.8: Ejemplo de la opción colorear en el gráfico X-Y

Figura 4.9: Barra de modificación de la opción Colorear en el gráfico X-Y

Otra opción viene dada por el icono que permite identificar los puntos según una variable externa. Por ejemplo si queremos ver en el gráfico de dispersión claramente los puntos que representan a los empleados con categoría laboral "T" debemos hacer clic en dicho icono, seleccionar la variable *CatLab* (figura 4.10), hacer clic en **Aceptar**, escribir **T** en la ventana **Etiq** (figura 4.11) y hacer clic sobre el icono de la derecha de la ventana obteniendo así los empleados con categoría laboral "T" identificados en color rojo (figura 4.12).

De forma análoga se puede identificar en el gráfico el punto deseado según su posición, o fila, en el archivo de datos. Para ello se usa la ventana **Fila**, se escribe el número de la misma y se hace clic en el icono de la derecha.

Figura 4.10: Opción de Identificación de Puntos

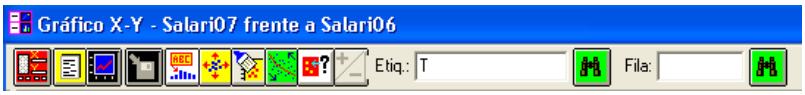

Figura 4.11: Barra de selección de la opción Identificación de Puntos en el gráfico X-Y

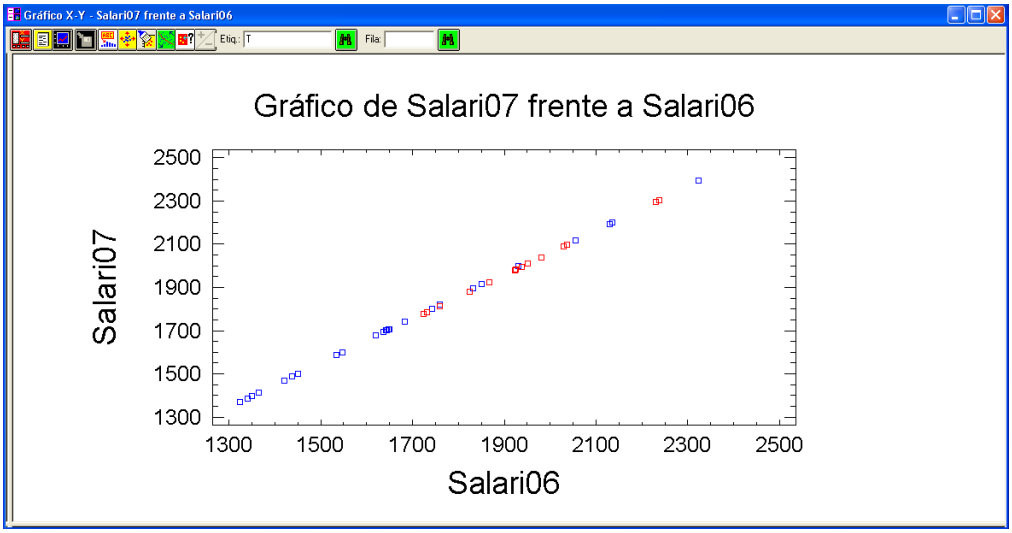

Figura 4.12: Ejemplo de identificación de puntos en el gráfico X-Y

Una última opción, que puede ser de gran utilidad en algunos casos, es la que nos permite etiquetar los puntos, es decir, nos permite identificar para cada punto la modalidad que tiene de otra de las variables disponibles. Para analizar mejor esta opción primero modificamos el gráfico de dispersión cambiando la variable Y a *Seguro07*. Para ello hacemos clic sobre el icono  de la ventana de resultados del gráfico de dispersión y aparecerá, de nuevo, el cuadro de diálogos donde se seleccionaban las variables. Quitamos de Y *Salari07* e introducimos *Seguro07* para obtener el nuevo gráfico de dispersión (figura 4.13).

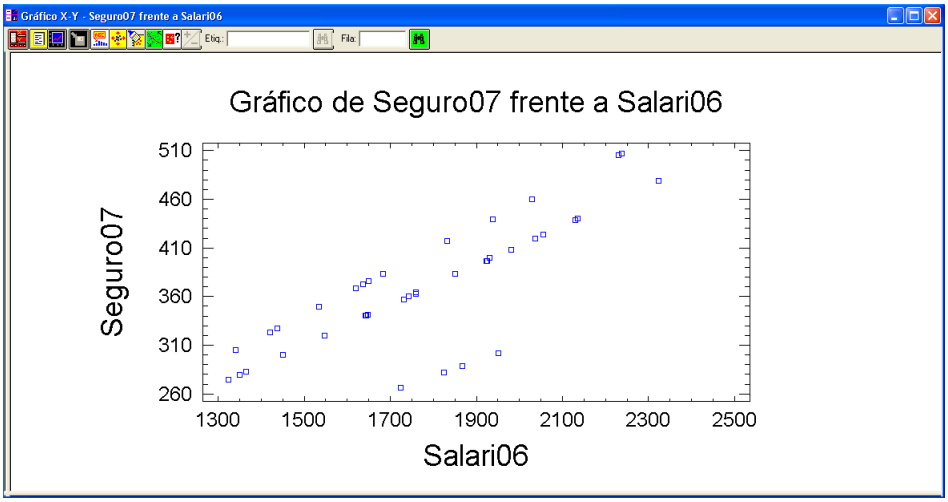

Figura 4.13: Segundo ejemplo de gráfico X-Y

En este gráfico se observa que los puntos se agrupan en tres conjuntos más o menos lineales. Sería interesante saber si dichos grupos vienen propiciados por alguna otra variable del estudio. Para comprobarlo podemos establecer marcas entre los puntos según alguna de las variables, normalmente de tipo cualitativo, para que cada punto se vea representado por un símbolo distinto según el valor o categoría que adopte la mencionada variable. Dicha opción se obtiene haciendo clic sobre el gráfico con el botón derecho y seleccionando **Opciones de Ventana...** (figura 4.14)

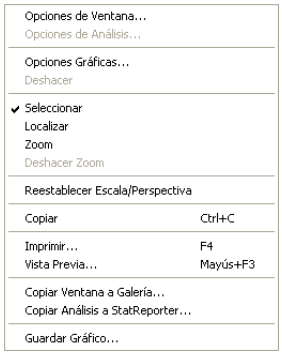

Figura 4.14: Opciones de Ventana

De esta forma aparece el un cuadro de diálogo (figura 4.15) donde introduciremos la variable que queremos usar para distinguir. Por ejemplo podemos probar con la variable *civil*, para ver si el estado civil influye en la relación entre el salario y el coste del seguro.

Haciendo clic en **Aceptar** modificamos el gráfico de dispersión apareciendo marcas distintas para las categorías de la variable *civil*: solteros, casados, separados, divorciados y viudos. Podemos observar ahora en el gráfico (figura 4.16) que, efectivamente, los grupos observados vienen influenciados por el estado civil del empleado. Se puede ver que existe cierta separación entre los solteros; los casados, separados y divorciados; y los viudos.

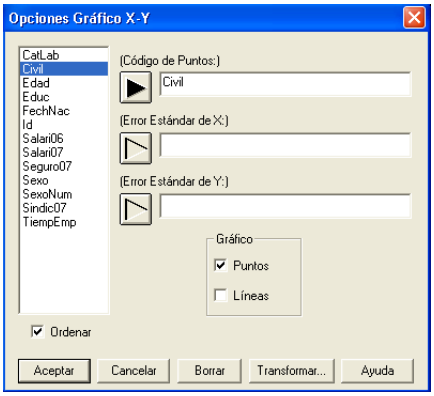

Figura 4.15: Opciones Gráfico X-Y

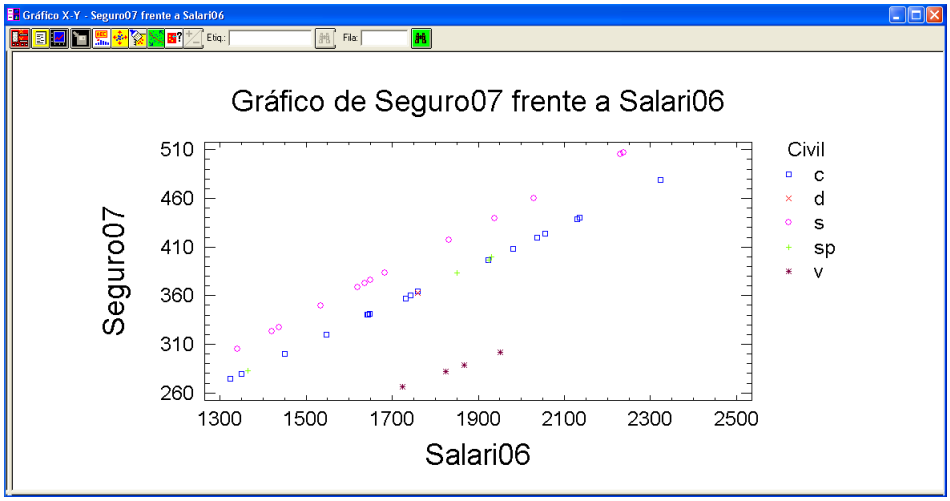

Figura 4.16: Ejemplo de aplicación de código de puntos en el gráfico X-Y

Ejemplo 1. Representar la nube de puntos de las variables $X =$*TiempEmp* e $Y =$*Salari07* y resolver las siguientes cuestiones:

a) Edita el gráfico de modo que le título del mismo sea "Nube de Puntos" y los puntos sean rombos de color rojo.

b) Visualizar en el gráfico la persona que más tiempo lleva en la empresa e indicar cuál es su salario en el año 2007.

(Solución: 2199 euros)

c) Identificar los puntos del gráfico según su categoría laboral y determinar la categoría del empleado que más tiempo lleva en la empresa.

(Solución: D)

d) Identificar los empleados con categoría "D" y comprobar si son los que más salario ganan o los que llevan más tiempo en la empresa. Indicar cuantos empleados aparecen en el gráfico con

dicha categoría.

(Solución: El empleado que más gana y el que más tiempo lleva en la empresa tienen categoría D pero hay dos empleados más con esa categoría que no cumplen ninguna de esas condiciones. Aparecen 4 empleados con cagoría D)

e) Identificar el empleado de la fila 15 e indicar su salario y tiempo en la empresa.
(Solución: 1822 euros y 11.2 años)

4.1.2. Covarianza y coeficiente de correlaciones

A pesar de que los gráficos de dispersión de las variables bajo estudio muestran que el ajuste lineal parece adecuado, sobre todo en el primero estudiado, y que se puede apreciar cierta asociación entre las variables, para corroborar esto y ampliar el estudio debemos obtener la covarianza y el coeficiente de correlación de las variables bajo estudio.

Recordemos que la covarianza, $S_{xy} = \frac{1}{N}\sum_{i=1}^{h}\sum_{j=1}^{k}(x_i - \bar{x})(y_j - \bar{y})n_{ij}$, es una medida que además de dar una idea de la variación conjunta que existe entre las dos variables, principalmente nos indica, por su signo, el sentido de dicha variación. Si la covarianza es positiva, nos indica que ambas variables varían en el mismo sentido alrededor de sus medias, mientras que si es negativa indica que la variación es en sentido contrario. Por otro lado, el coeficiente de correlación nos indica la intensidad con que las dos variables están relacionadas. En particular, el coeficiente de correlación lineal mide el grado de asociación lineal que existe entre ambas variables:

$$r = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^{h}\sum_{j=1}^{k}(x_i - \bar{x})(y_j - \bar{y})n_{ij}}{\sqrt{\sum_{i=1}^{h}(x_i - \bar{x})^2 \sum_{i=1}^{k}(y_i - \bar{y})^2}}$$

- Si $r = 1$ existe asociación perfecta positiva y la relación funcional entre ambas variables es exacta y positiva, variando ambas variables en el mismo sentido (al aumentar una aumenta la otra y al disminuir una disminuye la otra).

- Si $r = -1$ existe asociación perfecta negativa y la relación funcional entre ambas variables es exacta y negativa, variando ambas variables en el sentido opuesto (al aumentar una la otra disminuye y viceversa).

- Si $r = 0$ la asociación es nula, es decir, las variables no están asociadas siendo imposible encontrar una relación funcional entre ellas.

- Si $0 < r < 1$ la asociación es positiva, pero el grado de asociación entre las dos variables será mayor a medida que r se acerque más a 1, y será menor a medida que r se acerque más a 0. Si $-1 < r < 0$, la interpretación es similar a la anterior, pero con asociación negativa.

El paquete Statgraphics permite el cálculo tanto de la covarianza como del coeficiente de correlación lineal. Para ello debemos seleccionar **Descripción / Datos Numéricos / Análisis Multidimensional** (figura 4.17) y aparecerá el un cuadro de diálogos (figura 4.18) donde seleccionaremos las variables bajo estudio. En esta opción se pueden analizar más de dos variables, pero nosotros nos centraremos en el caso bidimensional que es el que estamos estudiando.

Comencemos estudiando las variables X =*Salari06* e Y =*Salari07*. Tras seleccionar las variables hacemos clic en **Aceptar** y aparece la ventana de resultados (figura 4.19). Los resultados incluyen un resumen del procedimiento y el Coeficiente de Correlación Lineal entre ambas variables: $r = 0.9999$, el cual nos indica que existe casi una asociación perfecta positiva entre ambas variables.

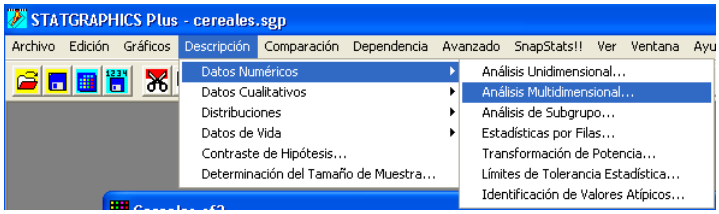

Figura 4.17: Análisis Multidimensional en menú Descripción

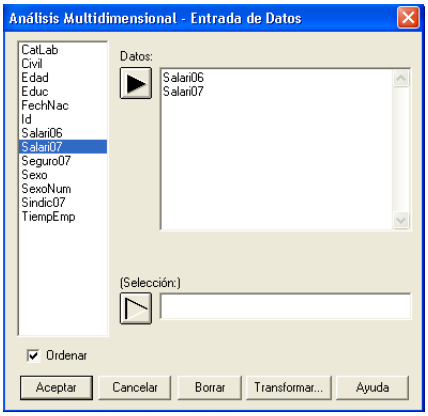

Figura 4.18: Entrada de datos de Análisis Multidimensional

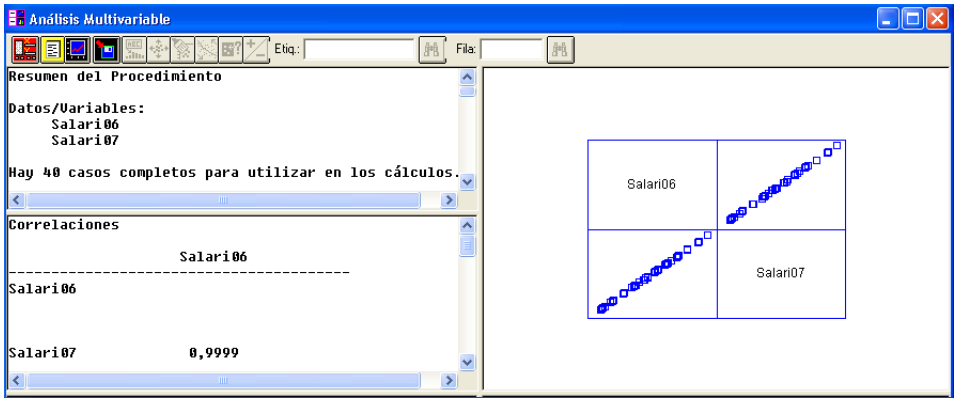

Figura 4.19: Ventana de resultados de Análisis Multidimensional

Para obtener la covarianza es necesario hacer clic sobre el icono  **Opciones tabulares**. De esta forma aparece un nuevo cuadro de diálogo (figura 4.20), donde están marcadas las opciones *Resumen del procedimiento* y *Correlaciones*. Seleccionando además la opción de **Covarianza** y haciendo clic en **Aceptar** se añade a la ventana de resultados una nueva subventana donde aparece la Covarianza: $S_{xy} = 73474.8$ (figura 4.21), la cual nos indica, al ser positiva, que ambas variables varían en el mismo sentido alrededor

de sus medias. Además se muestran las varianzas de las dos variables: S_x^2 =72035.9 y S_y^2 =74951.8 (figura 4.21).

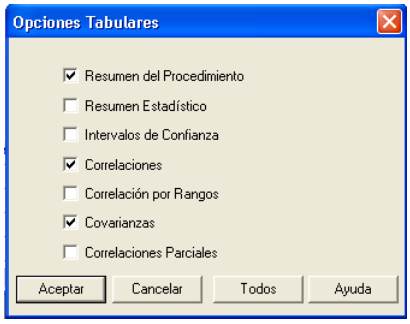

Figura 4.20: Opciones Tabulares del análisis multidimensional

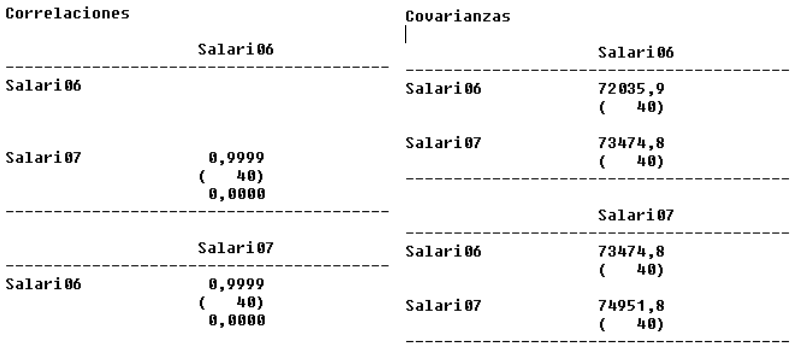

Figura 4.21: Ventanas de resultados de las opciones de Correlaciones y Covarianza

Si repetimos este mismo análisis sobre las variables *Salari06* y *Seguro07* obtenemos un Coeficiente de Correlación Lineal entre ambas variables: $r = 0.8113$, el cual nos indica que existe una considerable asociación positiva entre ambas variables. En este caso la Covarianza resulta ser: $S_{xy} = 13879.4$.

Tras este análisis visual de como se comportan las variables y del estudio del grado de asociación que existe entre ellas, parece adecuado utilizar el modelo lineal para el ajuste deseado.

Ejemplo 2. Resuelve las siguientes cuestiones para las variables X =*TiempEmp* e Y =*Salari07*:

a) Calcular el Coeficiente de Correlación Lineal e interpretarlo.
(Solución: r =0.2382, prácticamente no existe asociación lineal entre las variables)

b) Calcular la Covarianza e interpretarla.
(Solución: $Cov(X, Y)$ =259.956, al ser positiva nos indica que ambas variables varían en el mismo sentido)

c) Indicar la varianza de la variable *TiempEmp* y de la variable *Salari07*.
(Solución: $Var(X)$ =15.8849 y $Var(Y)$ =74951.8)

d) Indicar si, a la vista de estos valores, sería adecuado el usar una recta como función a ajustar a estos datos.
(Solución: No parece adecuado)

4.2. Expresión de la recta de regresión lineal simple. Varianza residual y coeficiente de determinación

4.2.1. Expresión de la recta de regresión lineal simple

La regresión lineal simple, como ya se ha indicado anteriormente, se basa en el ajuste de una función, en este caso una recta, a los datos observados para las variables X (independiente) e Y (dependiente). Es decir, vamos a buscar entre todas las rectas, cuya ecuación general es $Y = a + bX$, cuál es la que según el método de mínimos cuadrados mejor se ajusta. La recta que se obtiene, y que se denomina recta de regresión de Y/X, es:

$$Y = a + bX \qquad b = \frac{S_{xy}}{S_x^2} \qquad a = \bar{y} - b\bar{x}$$

y sirve para predecir la variable Y conocido el valor de la variable X.

El paquete Statgraphics permite obtener la recta de regresión lineal simple para lo cual debemos seleccionar **Dependencia / Regresión Simple...** (figura 4.22) e introducir las variables bajo estudio: $X =$*Salari06* e $Y =$*Salari07* (figura 4.23).

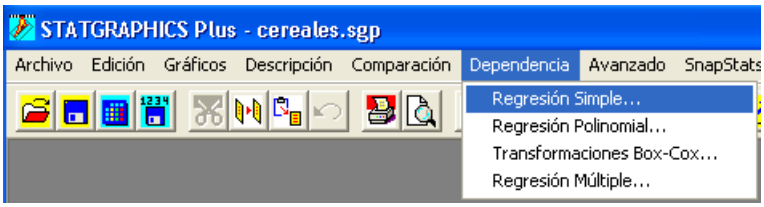

Figura 4.22: Regresión Simple en menú Dependencia

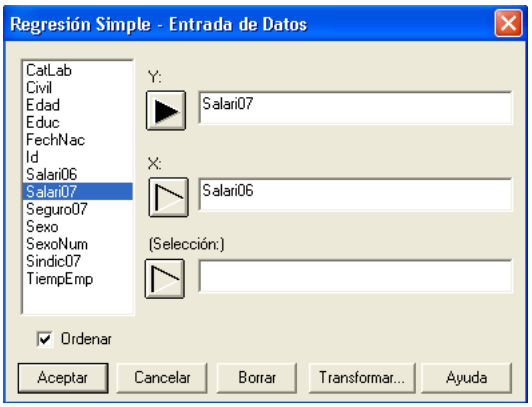

Figura 4.23: Entrada de datos de Regresión Simple

Al hacer clic en **Aceptar** aparece una venta de resultados (ver figura 4.24) que muestra a la izquierda el resumen del procedimiento y a la derecha el gráfico del modelo ajustado. Dentro del resumen del procedimiento se incluye la estimación de los parámetros de la recta de regresión, es decir, el valor estimado para el parámetro a (u ordenada) de la recta es 22.1674 y el del parámetro b (o pendiente) es 1.01998. Por tanto el modelo ajustado es: $Salari07 = 22.1674 + 1.01998 Salari06$. Recordemos que para ver completamente la ventana del resumen del procedimiento (figura 4.25) debemos hacer doble clic sobre la misma.

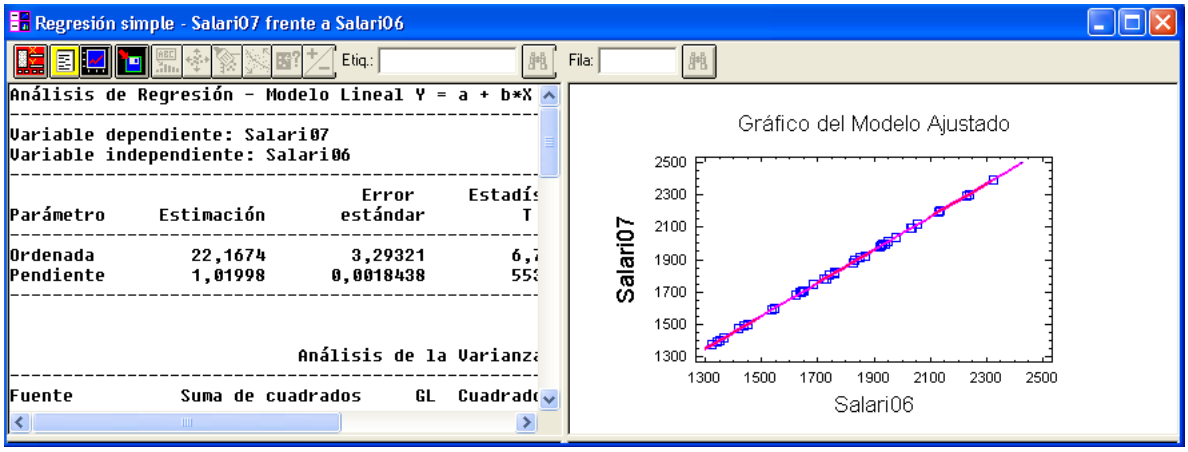

Figura 4.24: Ventana de resultados de Regresión Simple

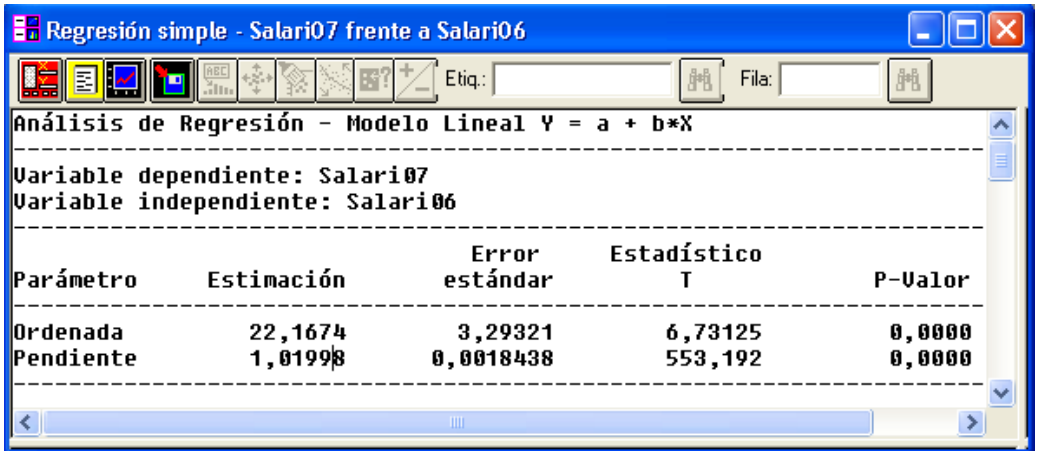

Figura 4.25: Resumen del procedimiento Regresión Simple

Si, una vez comprobado el modelo ajustado, deseamos verlo gráficamente, deberemos primero volver a la pantalla original de resultados, haciendo de nuevo doble clic sobre el resumen del procedimiento. Como ya se ha comentado, en la ventana de resultados aparece el gráfico del modelo ajustado. Si lo ampliamos, haciendo doble clic sobre él (figura 4.26), podemos apreciar la recta de regresión la cual, como era de esperar ya que el Coeficiente de Correlación Lineal era casi 1, pasa por todos los puntos.

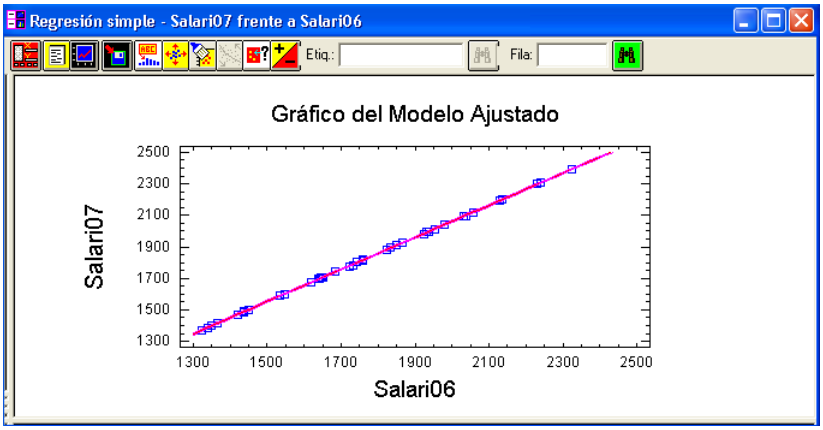

Figura 4.26: Gráfico del modelo ajustado

4.2.2. Varianza residual y coeficiente de determinación

Para determinar la bondad del ajuste realizado se recurre a la varianza residual, en general, y al coeficiente de determinación, en el ajuste lineal. La varianza residual es la suma de los residuos al cuadrado dividida por el número de datos utilizados en el ajuste y se usa como medida de la bondad del ajuste,

$$S_r^2 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{p} e_{ij}^2 n_{ij} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{p}[y_j - f(x_i)]^2 n_{ij}.$$

Esta medida tiene el inconveniente de que no está acotada superiormente, lo cual conduce a que no se pueda determinar si dicho valor es suficientemente grande o pequeño como para admitir un buen o mal ajuste. Por ello se estudia el coeficiente de determinación. Dicho coeficiente se basa en la descomposición de la varianza de la variable dependiente,

$$S_y^2 = S_e^2 + S_r^2$$

donde S_e^2 es la varianza explicada por la regresión.

$$S_e^2 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{p}[\hat{y}_i - \bar{y}]^2 n_{ij}$$

Esta descomposición sólo es válida para funciones lineales en los parámetros, y es sólo en estas funciones donde tiene pleno sentido definir el coeficiente de determinación:

$$R^2 = \frac{S_e^2}{S_y^2} = 1 - \frac{S_r^2}{S_y^2}$$

El coeficiente de determinación toma valores entre $0 \leq R^2 \leq 1$, tomando el valor 0 cuando el modelo no explica nada de Y a partir de X, es decir el ajuste es el peor posible, y tomando el valor 1 cuando todos los residuos son nulos, es decir el ajuste es perfecto. Para valores intermedios, según estén más próximos a un extremo y otro, nos indicarán un peor o mejor ajuste respectivamente.

En el paquete Statgraphics se incluye, tanto la varianza residual, explicada y total así como el coeficiente de determinación en el estudio de la regresión lineal. Estos resultados aparecen en el propio resumen del estudio, tras la estimación de los parámetros (figura 4.27).

Análisis de la Varianza

Fuente	Suma de cuadrados	GL	Cuadrado medio	Cociente-F	P-Valor
Modelo	2,92276E6	1	2,92276E6	306020,96	0,0000
Residuo	362,932	38	9,55084		
Total (Corr.)	2,92312E6	39			

Coeficiente de Correlación = 0,999938
R-cuadrado = 99,9876 porcentaje

Figura 4.27: Ventana de resultados del resumen del procedimiento Regresión Simple

En el caso que estamos estudiando obtenemos:

$$S_e^2 = 2922760 \qquad S_r^2 = 362.932 \qquad S_y^2 = 2923129 \qquad R^2 = 0.999876$$

El coeficiente de determinación nos indica que el modelo explica un 99.9876 % del salario en el año 2007 a partir del salario en el año 2006.

Si repetimos el análisis de regresión lineal con el segundo caso estudiado en la sección anterior, es decir cambiamos la variable Y a *Seguro07* se obtienen los siguientes resultados: La estimación de los parámetros de la recta de regresión resulta ser ahora de 23.3384 para el parámetro a (u ordenada) de la recta y de 0.192673 para el parámetro b (o pendiente). Por tanto el modelo ajustado es: $Seguro07 = 23.3384 + 0.192673 Salari06$.

En el gráfico de la recta de regresión (figura 4.28) podemos apreciar que no sólo se muestra la recta ajustada, que sería la línea central, sino que también se muestran dos intervalos.

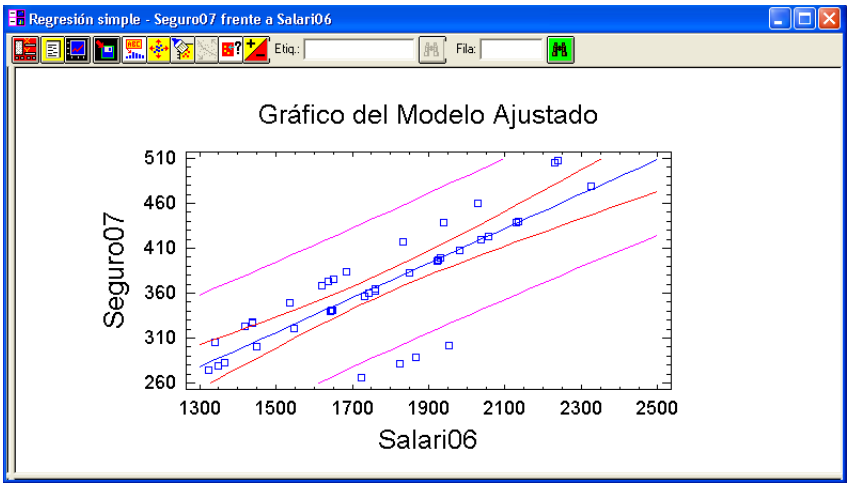

Figura 4.28: Segundo ejemplo de gráfico del modelo ajustado

El intervalo externo es el de predicción para las nuevas observaciones al 95 % de confianza. El intervalo interno es el de confianza para la media de muchas observaciones al 95 % de confianza. Si deseamos obtener sólo el gráfico del ajuste, es decir que no aparezcan los intervalos, debemos hacer clic con el botón derecho

del ratón sobre el gráfico, seleccionar **Opciones de ventana** (figura 4.29), quitar la selección **Límites de Confianza** y/o **Límites de Predicción** y hacer clic en **Aceptar**. En esta opción también se puede modificar el nivel de confianza deseado para los intervalos.

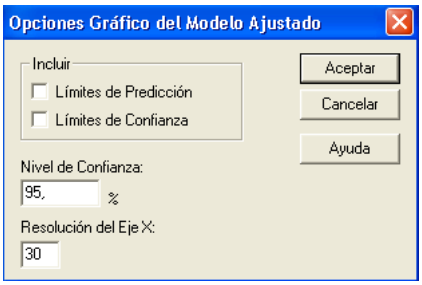

Figura 4.29: Opciones de ventana del gráfico del modelo ajustado

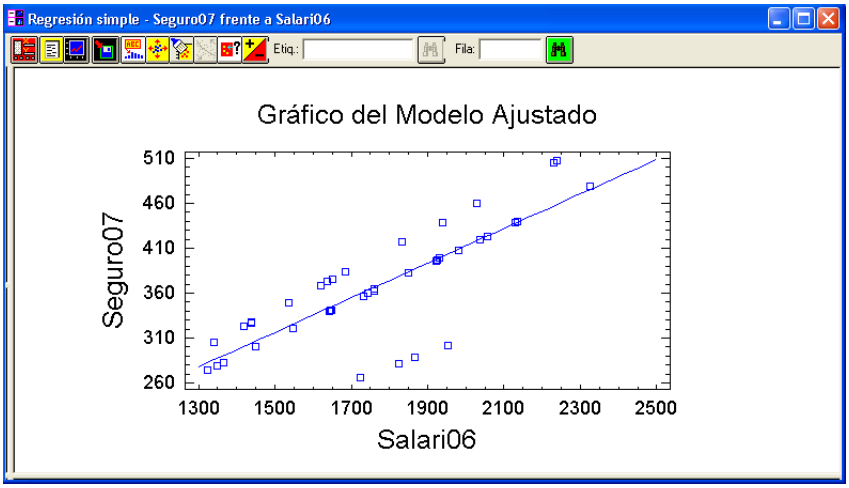

Figura 4.30: Gráfico del modelo ajustado sin intervalos

En este caso la recta del modelo ajustado (figura 4.30) no pasa por todos los puntos, lo cual era de esperar viendo la nube de puntos. En particular pasa, más o menos, próxima a la nube de puntos más central. El Coeficiente de Determinación R^2 =0.658247 nos indica que el modelo sólo explica un 65.8247 % del coste del seguro en el año 2007 a partir del salario en el año 2006, lo cual nos puede indicar que el modelo seleccionado no termina de describir bien los datos observados. Ello puede ser porque existan datos anómalos dentro del estudio. Para detectarlos estudiaremos en la siguiente sección los residuos del ajuste.

Ejemplo 3. Aplicar una regresión lineal simple sobre las variables X =*TiempEmp* e Y =*Edad* y resolver las siguientes cuestiones:

a) Indicar el modelo ajustado.
(Solución: *Edad*=14.2723+1.84122 *TiempEmp*)

b) Visualizar el gráfico del modelo ajustado y comprobar si la recta de regresión, la línea de color azul, para por la mayoría de los puntos. Indicar intuitivamente, a la vista del gráfico, si el ajuste parece adecuado.

(Solución: No pasa por la mayoría de los puntos pero no parece ajustarse demasiado mal)

c) Indicar la varianza residual, la varianza explicada por el modelo y el Coeficiente de Determinación e interpretarlo.

(Solución: σ_r^2 =558.577, σ_e^2 =2100.2, R^2 =0.789912, el modelo explica el 78.99 % de la edad a partir del tiempo en la empresa)

4.3. Estudio de los residuos

Para estudiar los residuos del un ajuste realizado tenemos dos opciones, mediante una tabla que muestra los valores residuales inusuales o mediante el gráfico que muestra los residuos frente a los valores de la variable X.

La tabla la obtenemos seleccionando la opción **Residuos Atípicos** en el cuadro de diálogo que aparece al hacer clic en el icono  de la ventana de resultados de la regresión lineal simple.

El gráfico se obtiene seleccionando la opción **Residuo frente a X** en el cuadro de diálogos que aparece al hacer clic en el icono  de la misma ventana.

Si lo aplicamos al ejemplo segundo que hemos estudiado en la sección anterior (regresión lineal de X =*Salari06* e Y =*Seguro07*) (figuras 4.31 y 4.32) obtenemos que ambas opciones muestran que, en este caso, hay cuatro puntos cuyos residuos estandarizados están fuera del rango 2, y que por tanto podrían ser anómalos.

Dichos datos coinciden con los puntos correspondientes a los empleados viudos. Si recordamos cuando estudiamos la nube de puntos, en la sección 4.1.1 (figura 4.16), vimos que parecían formar un grupo algo separado.

Residuos Atípicos

Fila	X	Y	Y Predicha	Residuo	Residuo Estudentizado
34	1825,0	281,96	378,967	-97,0066	-2,84
35	1724,0	266,36	359,507	-93,1466	-2,70
39	1867,0	288,45	387,059	-98,6089	-2,90
40	1952,0	301,58	403,436	-101,856	-3,03

Figura 4.31: Ventana de resultados de la opción Residuos Atípicos de la regresión simple

Parecer, por tanto, buena idea eliminar dichos datos del análisis para intentar mejorarlo. Para ello editamos el gráfico del ajuste lineal haciendo doble clic sobre él, seleccionamos uno de los puntos a eliminar y hacemos clic en el icono  . Repetimos el mismo procedimiento hasta eliminar todos los puntos obteniendo un nuevo gráfico (figura 4.33), donde podemos ver como la recta ha cambiado pasando ahora, más o menos, por el centro de las dos nubes de puntos que se observan.

Si miramos ahora el resumen del procedimiento observamos que los valores también han cambiado. El modelo que se ajusta ahora es $Seguro07 = 16.5906 + 0.204959 Salario06$ y el Coeficiente de Determinación a aumentado a R^2 =0.911049, es decir, el modelo ha mejorado considerablemente. Por otro lado, tanto la tabla de valores atípicos como el gráfico de residuos nos indican que ya no existe ningún valor que pueda parecer anómalo.

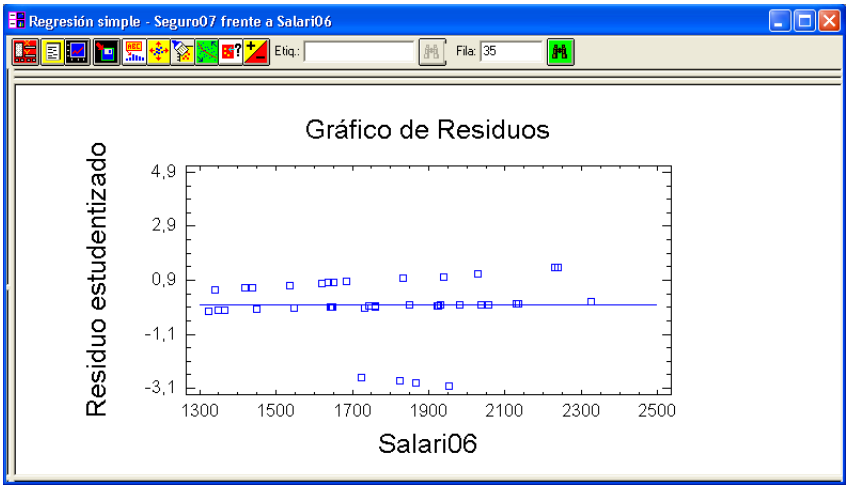

Figura 4.32: Gráfico de Residuos de la regresión simple

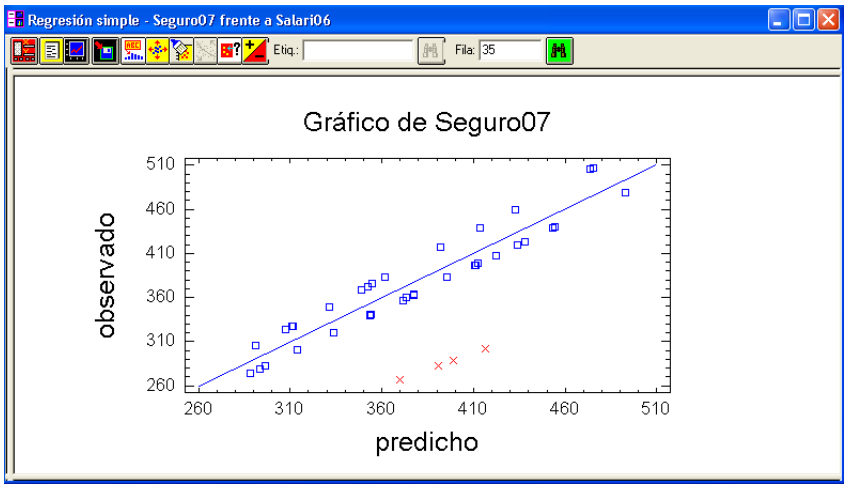

Figura 4.33: Gráfico del modelo ajustado tras eliminar los residuos atípicos

Ejemplo 4. Volver a ajustar una regresión lineal simple sobre las variables $X = TiempEmp$ e $Y = Edad$ y resolver las siguientes cuestiones:

a) Indicar los valores de Y para los residuos atípicos.
(Solución: 34.98 y 42.06 años)

b) Eliminar del gráfico los valores con residuos atípicos. Indicar los nuevos valores de Y para los residuos atípicos.
(Solución: 27.59, 27.84 y 44.52 años)

c) Repetir el procedimiento de eliminar los residuos atípicos hasta que no aparezca ninguno. Indicar el número de datos eliminados del estudio.
(Solución: 8 datos)

d) Indicar el nuevo modelo ajustado.
(Solución: $Edad$=12.085+2.0317$TiempEmp$)

e) Indicar el nuevo Coeficiente de Determinación.
(Solución: R^2 =0.916494)

4.4. Obtención de valores predichos y residuos

Una vez ajustado el modelo, es decir, obtenida la recta de regresión, puede ser útil obtener los valores que se predicen con dicho modelo así como los residuos del ajuste. Recordemos que los valores predichos son $\hat{y}_i = a + bx_i$ y los residuos son $e_{ij} = y_j - \hat{y}_i$.

Para obtener los valores predichos, con el modelo ajustado, para la variable Y a partir de los valores del archivo de datos de la variable X, debemos hacer clic sobre el icono  **Guardar Resultados** de la ventana de resultados de la regresión lineal, para que aparezca una cuadro de diálogos (figura 4.34) donde podemos guardar una serie de resultados, entre los cuales están los residuos y los valores predichos.

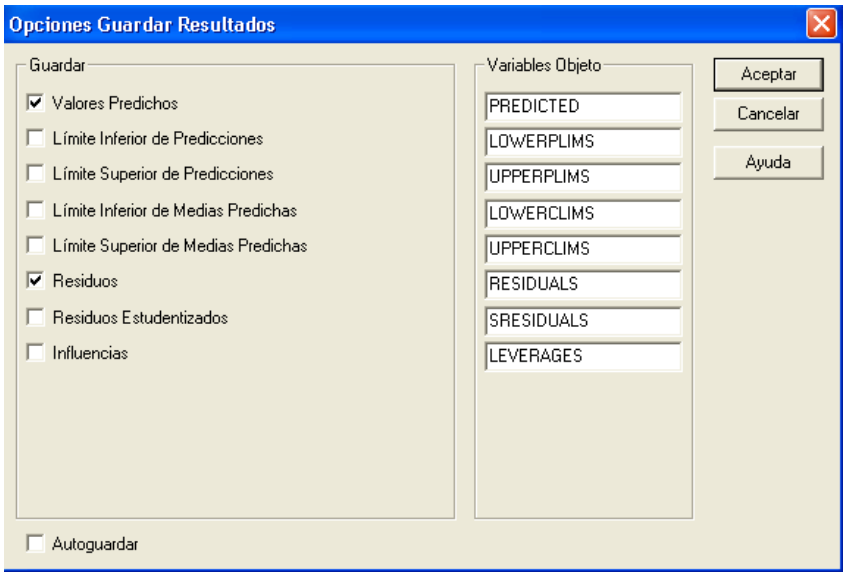

Figura 4.34: Opciones Guardar Resultados de la regresión simple

Seleccionando dichas opciones y haciendo clic en **Aceptar**, aparecen dos nuevas columnas en la ventana de datos con los valores deseados (figura 4.35).

Los valores predichos que hemos obtenido, como su propio nombre indica, nos muestran las predicciones de la variable Y según los valores que toma la variable X a partir del modelo ajustado. Por ejemplo, en el caso práctico que hemos estudiado del ajuste de una regresión lineal para X =*Salari06* e Y =*Seguro07*, el modelo lineal ajustado predice para el empleado que gana 2230 euros al mes en el año 2006 un coste del seguro en el año 2007 de 473.65 euros. Estos datos podemos observarlos en la tabla de datos, tras guardar las predicciones y los residuos, mirando la columna *Salari06* y la columna *PREDICTED* para el empleado con código de empleado (*Id*) 37. Así mismo podemos observar que el error cometido con dicha predicción, es decir el residuo, es 31.6702 el cual aparece en la columna *RESIDUALS*.

miempresa.sf3

	PREDICTED	RESIDUALS
1	331,203	18,3169
2	291,236	13,884
3	307,633	15,6972
4	296,36	-13,8
5	313,782	-13,6316
6	293,286	-13,8356
7	288,162	-13,8816
8	311,117	16,0829
9	353,749	-13,2286

Figura 4.35: Ventana de resultados con las columnas PREDICTED y RESIDUALS

Pero no tiene mucho sentido ver la predicción para un empleado del cual ya sabemos el valor para dicha variable. Lo usual es querer predecir el valor de la variable Y para valores de X nuevos, distintos de los incluidos en los datos originales. Para conseguir esas predicciones debemos hacer clic en el icono  de la ventana de resultados y seleccionar la opción **Predicciones** del cuadro de diálogos que surge (figura 4.36).

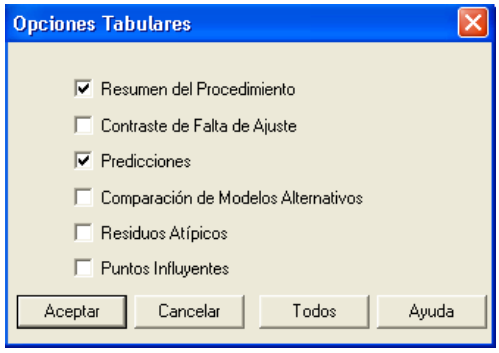

Figura 4.36: Opciones Tabulares de la regresión simple

Al hacer clic en **Aceptar** aparece una nueva subventana en la ventana de resultados del análisis de regresión lineal (figura 4.37) donde aparecen dos predicciones para Y, la del valor mínimo y la del valor máximo de X, junto al intervalo de predicción para las nuevas observaciones y al intervalo de confianza para la media de las observaciones, ambos al nivel de confianza del 95 %.

Valores predichos

		95,00% Límites de Predicción		95,00% Límites de Confianza	
X	Predicho Y	Inferior	Superior	Inferior	Superior
1325,0	288,162	249,366	326,958	276,693	299,63
2324,0	492,916	453,275	532,557	478,851	506,98

Figura 4.37: Ventana de resultados de la opción Predicciones de la regresión simple

Para que en dicha ventana aparezcan las predicciones de los valores de X que deseemos debemos hacer clic con el botón derecho sobre ella. De esta forma aparecerá un cuadro de diálogos (figura 4.38) y al hacer clic ahora sobre **Opciones de Ventana...** aparece un nuevo cuadro de diálogos con las opciones de predicción (figura 4.39). En él aparecen por defecto los valores mínimos y máximo de X. Esos valores de X nos indican el rango de valores del que no debemos salirnos a la hora de usar el modelo ajustado para predecir, pues fuera de ese rango no tenemos asegurada la bondad de la predicción aunque el ajuste sea muy bueno. Es decir, debemos usar el modelo ajustado para predecir valores de Y a partir de valores de X dentro del rango estudiado de dicha variable.

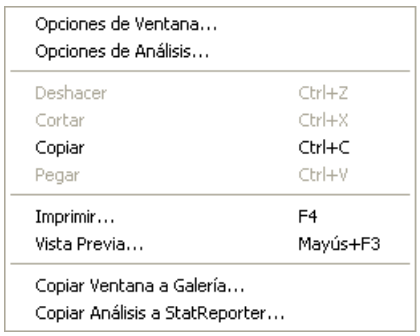

Figura 4.38: Opciones de ventana

Si incluimos los valores de X para los que queremos conocer la predicción de Y y hacemos clic en **Aceptar** aparecerán dichas predicciones en la ventana anterior.

Por ejemplo, para obtener las predicciones de los costes de los seguros en el año 2007 para un empleado cuyo sueldo sea 1500 euros y otro que sea 2000 euros en el año 2006, debemos incluir dichos sueldos en cualquiera de las casillas vacías de las predicciones de X (figura 4.39) y obtenemos que las predicciones son (figura 4.40) de 324.03 euros el coste del seguro para un empleado con salario 1500 euros y de 426.509 euros para un empleado con salario 2000 euros.

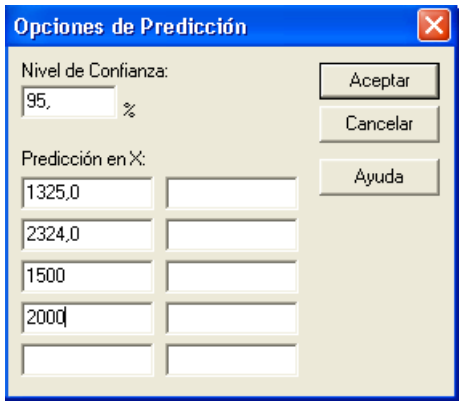

Figura 4.39: Opciones de Predicción de la regresión simple

Valores predichos

		95,00%		95,00%	
	Predicho	Límites de Predicción		Límites de Confianza	
X	Y	Inferior	Superior	Inferior	Superior
1325,0	288,162	249,366	326,958	276,693	299,63
2324,0	492,916	453,275	532,557	478,851	506,98
1500,0	324,03	286,018	362,041	315,586	332,473
2000,0	426,509	388,549	464,469	418,302	434,716

Figura 4.40: Ventana de resultados de las Predicciones para la regresión simple

Ejemplo 5. Volver al ajuste lineal realizado en el ejercicio anterior, tras eliminar todos los datos que daban residuos atípicos, y resolver las siguientes cuestiones:

a) Guardar los residuos y las predicciones para los datos del estudio. Indicar la predicción, y error de la misma, que da el modelo para la *Edad* de los empleados que llevan en la empresa aproximadamente 16 años. (Para saber qué filas ocupan dichos empleados identificarlos previamente en el gráfico del modelo).

(Solución: La predicción es de 43 años para ambos empleados con errores de -1.71 y 1.26)

b) Indicar la predicción de *Edad* que da el modelo para el empleado que menos tiempo lleva en la empresa y para el que más tiempo lleva.

(Solución: Las predicciones son de 20 y 53 años)

c) Indicar la predicción de *Edad* que da el modelo para un empleado que lleva 14 años en la empresa.

(Solución: La predicción es de 40 años)

d) Indicar la predicción de *Edad* que da el modelo para un empleado que lleva 1 año en la empresa y si dicha predicción tiene sentido.

(Solución: La predicción es de 14 años, lo cual no tiene sentido, se deben tener al menos 16 años para poder trabajar)

e) Indicar la predicción de *Edad* que da el modelo para un empleado que lleva 30 año en la empresa y si dicha predicción tiene sentido.

(Solución: La predicción es de 73 años, el empleado ya estaría jubilado)

4.5. Otros modelos de regresión simple no lineal

Como ya se especificó en la introducción del tema, el Statgraphics permite no sólo el estudio de la regresión lineal simple, sino que admite otras funciones para el ajuste como la exponencial o el modelo multiplicativo entre otros. Al igual que con el modelo lineal, se pueden obtener las estimaciones de los parámetros, los residuos y las predicciones para el resto de los modelos siguiendo los mismos pasos.

El problema surge al decidir que modelo debemos ajustar. Para ello podemos observar la nube de puntos, pero esto no siempre nos dará una idea de que modelo debemos usar. Otra opción, que siempre podremos usar, es la de comparar todos los modelos. Para ello debemos aplicar el modelo lineal, siguiendo los pasos indicados anteriormente, y a continuación haciendo clic en el icono  de la ventana de resultados obtenida en la regresión lineal podemos seleccionar la opción **Comparación de Modelos Alternativos** (figura 4.41). Haciendo ahora clic en **Aceptar** obtenemos los resultados deseados.

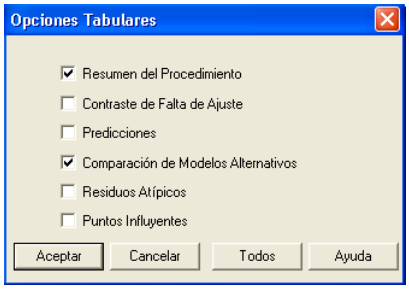

Figura 4.41: Opciones Tabulares de la regresión simple

Si aplicamos las últimas indicaciones sobre la ventana de resultados que obtuvimos al ajustar una regresión lineal simple con las variables X =*Salari06* e Y =*Seguro07*, con todos los datos, obtenemos una subventana dentro de la ventana de resultados con la comparación de los modelos (figura 4.42).

Comparación de Modelos Alternativos

Modelo	Correlación	R-cuadrado
Lineal	0,8113	65,82%
Raiz cuadrada-X	0,8045	64,73%
Raiz cuadrada-Y	0,8022	64,36%
Logarítmico-X	0,7968	63,48%
Exponencial	0,7909	62,56%
Multiplicativo	0,7813	61,04%
Inverso-X	-0,7785	60,61%
curva-S	-0,7680	58,98%
Inverso-Y	-0,7627	58,17%
Doble inverso	0,7492	56,13%
Logístico	<sin ajuste>	
Log Probit	<sin ajuste>	

Figura 4.42: Ventana de resultados de Modelos Alternativos de la regresión simple

Según estos resultados el modelo más adecuado el es lineal con un R^2 =0.6582, es decir es el más adecuado pero no termina de ajustar bien los datos. Como ya estudiamos en la sección 3, si eliminamos del estudio los valores correspondientes a los empleados viudos el ajuste mejora considerablemente. Por tanto si repetimos el procedimiento descrito en dicha sección, y eliminamos dichos cuatro valores, los resultados de la comparación de modelos cambian (figura 4.43).

Comparación de Modelos Alternativos

Modelo	Correlación	R-cuadrado
Multiplicativo	0,9557	91,34%
Raiz cuadrada-Y	0,9554	91,28%
Doble inverso	0,9550	91,20%
Lineal	0,9545	91,10%
Exponencial	0,9539	91,00%
Raiz cuadrada-X	0,9529	90,80%
curva-S	-0,9522	90,67%
Logarítmico-X	0,9499	90,23%
Inverso-Y	-0,9440	89,11%
Inverso-X	-0,9401	88,38%
Logístico	<sin ajuste>	
Log Probit	<sin ajuste>	

Figura 4.43: Segundo ejemplo de ventana de resultados de Modelos Alternativos de la regresión simple

Ahora el modelo más adecuado es el multiplicativo ($Y = aX^b$ ó $\ln Y = \ln a + b \ln X$) con R^2 =0.9134, aunque la mejora frente al modelo lineal (R^2 =0.911) no es muy destacable. En realidad, como puede observarse, cualquiera de los modelos que se permiten ajustar son más o menos adecuados en este caso.

Una vez decidido el modelo más adecuado para obtener la estimación de los parámetros del modelo ajustado debemos hacer clic con el botón derecho sobre la ventana de resultados de forma que aparecerá un cuadro de diálogos (figura 4.44). Haciendo clic sobre **Opciones de Análisis...** surge un nuevo cuadro de diálogos (figura 4.5) con todas las opciones de los distintos análisis que admite el Statgraphics.

Figura 4.44: Opciones de Análisis

Figura 4.45: Opciones de análisis de la regresión simple

Si se selecciona el modelo deseado y se hace clic en **Aceptar** se obtienen los mismo resultados que para el modelo lineal pero en el nuevo modelo. Por ejemplo, siguiendo con nuestro análisis, hemos visto que el mejor era el modelo multiplicativo. Si seleccionamos dicha opción y hacemos clic en **Aceptar** la ventana de resultados cambia mostrando el nuevo modelo ajustado (figura 4.46) y el gráfico del mismo (figura 4.47). En este caso el modelo ajustado es ln*Seguro07* = -1.17955+0.951698 ln*Salari06*.

Al igual que en el estudio del modelo lineal podemos analizar los residuos de este nuevo ajuste mediante su gráfico o su tabla siguiendo los mismos pasos que se indicaron en la sección 3. En el gráfico podemos observar (figura 4.48) que los residuos están dentro del rango adecuado (-2,2) y que por tanto no hay ninguno que pueda parecer anómalo.

Análisis de Regresión - Modelo Multiplicativo: Y = a*X^b

Variable dependiente: Seguro07
Variable independiente: Salari06

Parámetro	Estimación	Error estándar	Estadístico T	P-Valor
Ordenada	-1,17955	0,374914	-3,14619	0,0034
Pendiente	0,951698	0,050249	18,9396	0,0000

Figura 4.46: Ventana de resultados del ajuste del modelo multiplicativo de la regresión simple

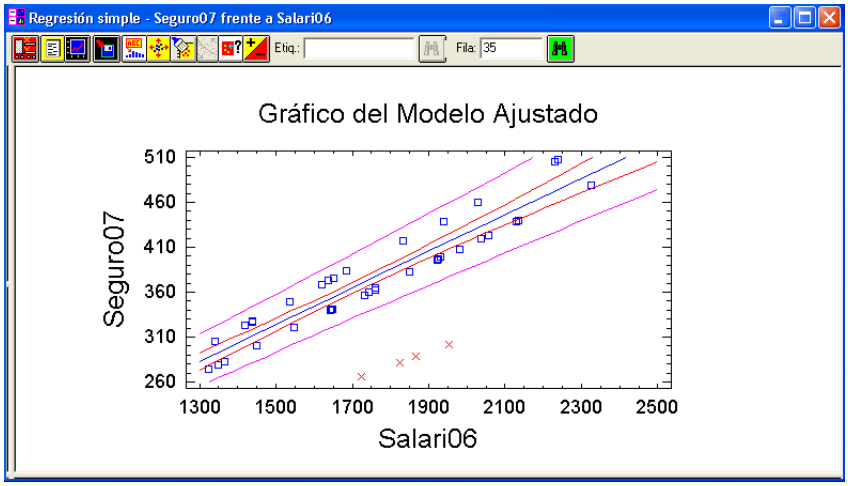

Figura 4.47: Gráfico del modelo multiplicativo ajustado en la regresión simple

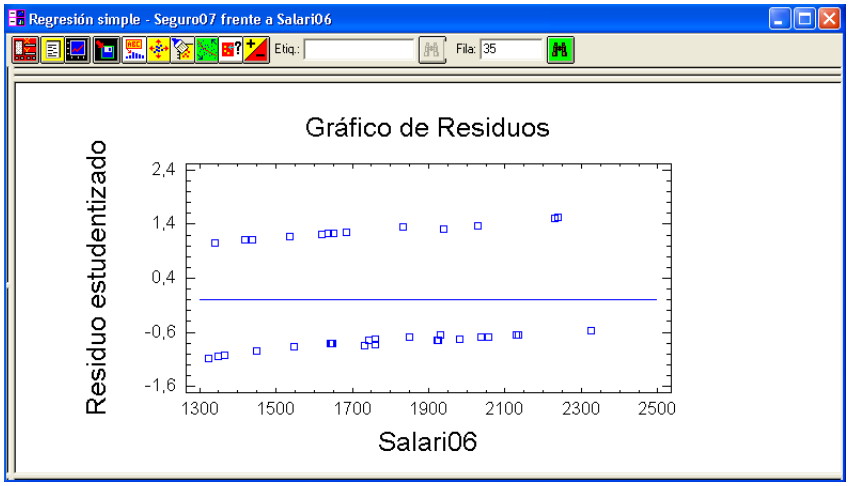

Figura 4.48: Gráfico de residuos del ajuste del modelo multiplicativo