Capítulo 18
Análisis de regresión lineal
El procedimiento *Regresión lineal*

Introducción

El análisis de regresión lineal es una técnica estadística utilizada para estudiar la relación entre variables. Se adapta a una amplia variedad de situaciones. En la investigación social, el análisis de regresión se utiliza para predecir un amplio rango de fenómenos, desde medidas económicas hasta diferentes aspectos del comportamiento humano. En el contexto de la investigación de mercados puede utilizarse para determinar en cuál de diferentes medios de comunicación puede resultar más eficaz invertir; o para predecir el número de ventas de un determinado producto. En física se utiliza para caracterizar la relación entre variables o para calibrar medidas. Etc.

Tanto en el caso de dos variables (regresión *simple*) como en el de más de dos variables (regresión *múltiple*), el análisis de regresión lineal puede utilizarse para explorar y cuantificar la relación entre una variable llamada dependiente o criterio (Y) y una o más variables llamadas independientes o predictoras ($X_1, X_2, ..., X_k$), así como para desarrollar una ecuación lineal con fines predictivos. Además, el análisis de regresión lleva asociados una serie de procedimientos de diagnóstico (análisis de los residuos, puntos de influencia) que informan sobre la estabilidad e idoneidad del análisis y que proporcionan pistas sobre cómo perfeccionarlo.

Nuestro objetivo es el de proporcionar los fundamentos del análisis de regresión. Al igual que en los capítulos precedentes, no haremos hincapié en los aspectos más técnicos del análisis, sino que intentaremos fomentar la compresión de cuándo y cómo utilizar el análisis de regresión lineal, y cómo interpretar los resultados. También prestaremos atención a otras cuestiones como el chequeo de los supuestos del análisis de regresión y la forma de proceder cuando se incumplen.

La recta de regresión

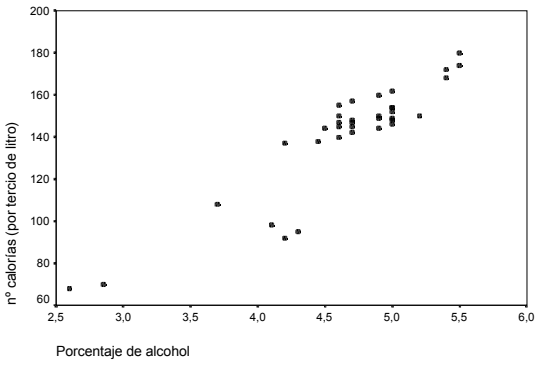
En el capítulo anterior (sobre correlación lineal) hemos visto que un diagrama de dispersión ofrece una idea bastante aproximada sobre el *tipo de relación* existente entre dos variables. Pero, además, un diagrama de dispersión también puede utilizarse como una forma de *cuantificar* el grado de relación lineal existente entre dos variables: basta con observar el grado en el que la nube de puntos se ajusta a una línea recta.

Ahora bien, aunque un diagrama de dispersión permite formarse una primera impresión muy rápida sobre el tipo de relación existente entre dos variables, utilizarlo como una forma

de *cuantificar* esa relación tiene un serio inconveniente: la relación entre dos variables no siempre es perfecta o nula; de hecho, habitualmente no es ni lo uno ni lo otro.

Supongamos que disponemos de un pequeño conjunto de datos con información sobre 35 marcas de cerveza y que estamos interesados en estudiar la relación entre el grado de alcohol de las cervezas y su contenido calórico. Un buen punto de partida para formarnos una primera impresión de esa relación podría ser la representación de la nube de puntos, tal como muestra el diagrama de dispersión de la figura 18.1.

Figura 18.1. Diagrama de dispersión de *porcentaje de alcohol* por *nº de calorías*.



El eje vertical muestra el número de calorías (por cada tercio de litro) y el horizontal el contenido de alcohol (expresado en porcentaje). A simple vista, parece existir una relación positiva entre ambas variables: conforme aumenta el porcentaje de alcohol, también aumenta el número de calorías. En esta muestra no hay cervezas que teniendo alto contenido de alcohol tengan pocas calorías y tampoco hay cervezas que teniendo muchas calorías tengan poco alcohol. La mayor parte de las cervezas de la muestra se agrupan entre el 4,5 % y el 5 % de alcohol, siendo relativamente pocas las cervezas que tienen un contenido de alcohol inferior a ése. Podríamos haber extendido el rango de la muestra incluyendo cervezas sin alcohol, pero el rango de calorías y alcohol considerados parece bastante apropiado: no hay, por ejemplo, cervezas con un contenido de alcohol del 50 %, o cervezas sin calorías.

¿Cómo podríamos describir los datos que acabamos de proponer? Podríamos decir simplemente que el aumento del porcentaje de alcohol va acompañado de un aumento en el número de calorías; pero esto, aunque correcto, es poco específico. ¿Cómo podríamos obtener una descripción más concreta de los resultados? Podríamos, por ejemplo, listar los datos concretos de que disponemos; pero esto, aunque preciso, no resulta demasiado informativo.

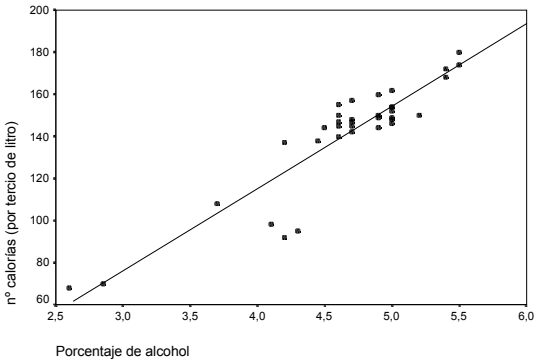
Podríamos hacer algo más interesante. Por ejemplo, describir la pauta observada en la nube de puntos mediante una función matemática simple, tal como una línea recta. A primera vista, una línea recta podría ser un buen punto de partida para describir resumidamente la nube de puntos de la figura 18.1.

Puesto que una línea recta posee una fórmula muy simple,

$$Y_i = B_0 + B_1 X_i$$

podemos comenzar obteniendo los coeficientes B_0 y B_1 que definen la recta. El coeficiente B_1 es la pendiente de la recta: el cambio medio que se produce en el número de calorías (Y_i) por cada unidad de cambio que se produce en el porcentaje de alcohol (X_i) . El coeficiente B_0 es el punto en el que la recta corta el eje vertical: el número medio de calorías que corresponde a una cerveza con porcentaje de alcohol cero. Conociendo los valores de estos dos coeficientes, nuestro interlocutor podría reproducir la recta y describir con ella la relación existente entre el contenido de alcohol y el número de calorías. Aunque no entremos todavía en detalles de cómo obtener los valores de B_0 y B_1, sí podemos ver cómo es esa recta (figura 18.2).

Figura 18.2. Diagrama de dispersión y recta de regresión (*% de alcohol* por *nº de calorías*).



$$Y_i = -33{,}77 + 37{,}65\ X_i$$

nº de calorías = − 33,77 + 37,65 (% de alcohol)

Vemos que, en general, la recta hace un seguimiento bastante bueno de los datos. La fórmula de la recta aparece a la derecha del diagrama. La pendiente de la recta (B_1) indica que, en promedio, a cada incremento de una unidad en el porcentaje de alcohol (X_i) le corresponde un incremento de 37,65 calorías (Y_i). El origen de la recta (B_0) sugiere que una cerveza sin alcohol (grado de alcohol cero) podría contener −33,77 calorías. Y esto, obviamente, no parece posible. Al examinar la nube de puntos vemos que la muestra no contiene cervezas con menos de un 2 % de alcohol. Así, aunque el origen de la recta aporta información sobre lo que podría ocurrir si extrapolamos hacia abajo la pauta observada en los datos hasta llegar a una cerveza con grado de alcohol cero, al hacer esto estaríamos efectuando pronósticos en un rango de valores que va más allá de lo que abarcan los datos disponibles, y eso es algo extremadamente arriesgado en el contexto del análisis de regresión*.

La mejor recta de regresión

En una situación ideal (e irreal) en la que todos los puntos de un diagrama de dispersión se encontraran en una línea recta, no tendríamos que preocuparnos de encontrar la recta que mejor resume los puntos del diagrama. Simplemente uniendo los puntos entre sí obtendríamos la recta

* Debemos aprender una lección de esto: la primera cosa razonable que podríamos hacer es añadir en nuestro estudio alguna cerveza con porcentaje de alcohol cero; probablemente así obtendríamos una recta con un origen más realista.

con mejor ajuste a la nube de puntos. Pero en una nube de puntos más realista (como la de las figuras 18.1 y 18.2) es posible trazar muchas rectas diferentes. Obviamente, no todas ellas se ajustarán igualmente bien a la nube de puntos. Se trata de encontrar la recta capaz de convertirse en el mejor representante del conjunto total de puntos.

Existen diferentes procedimientos para ajustar una función simple, cada uno de los cuales intenta minimizar una medida diferente del grado de ajuste. La elección preferida ha sido, tradicionalmente, la recta que hace *mínima la suma de los cuadrados de las distancias verticales entre cada punto y la recta*. Esto significa que, de todas las rectas posibles, existe una y sólo una que consigue que las distancias verticales entre cada punto y la recta sean mínimas (las distancias se elevan al cuadrado porque, de lo contrario, al ser unas positivas y otras negativas, se anularían unas con otras al sumarlas).

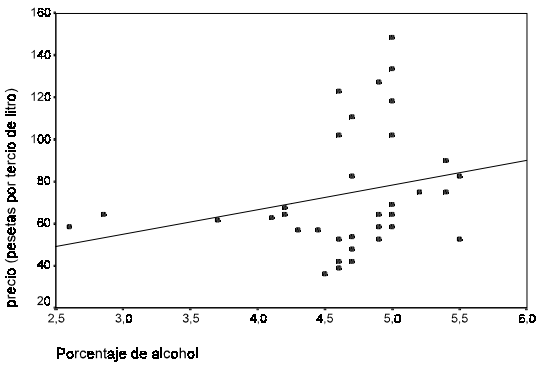
Bondad de ajuste

Además de acompañar la recta con su fórmula, podría resultar útil disponer de alguna indicación precisa del grado en el que la recta se ajusta a la nube de puntos. De hecho, la mejor recta posible no tiene por qué ser buena.

Imaginemos una situación como la presentada en el diagrama de la figura 18.3, en el que la recta consigue un ajuste bastante más pobre que en el caso de la figura 18.2. Ahora hemos representado el porcentaje de alcohol de las cervezas (eje horizontal) y el precio de las mismas (eje vertical). Y no parece existir la misma pauta de asociación detectada entre las variables de la situación anterior.

Así pues, aunque siempre resulta posible, cualquiera que sea la nube de puntos, obtener la recta mínimo-cuadrática, necesitamos información adicional para determinar el grado de fidelidad con que esa recta describe la pauta de relación existente en los datos.

Figura 18.3. Diagrama de dispersión, recta de regresión y ajuste (*% de alcohol* por *precio*).



precio = 20,16 + 11,61 (% de alcohol)

$R^2 = 0{,}06$

¿Cómo podemos cuantificar ese *mejor* o *peor* ajuste de la recta? Hay muchas formas de resumir el grado en el que una recta se ajusta a una nube de puntos. Podríamos utilizar la media de los residuos, o la media de los residuos en valor absoluto, o las medianas de alguna de esas medidas, o alguna función ponderada de esas medidas, etc.

Una medida de ajuste que ha recibido gran aceptación en el contexto del análisis de regresión es el ***coeficiente de determinación*** R^2: el cuadrado del coeficiente de correlación múltiple. Se trata de una medida estandarizada que toma valores entre 0 y 1 (0 cuando las variables son independientes y 1 cuando entre ellas existe relación perfecta).

Este coeficiente posee una interpretación muy intuitiva: representa el grado de ganancia que podemos obtener al predecir una variable basándonos en el conocimiento que tenemos de otra u otras variables. Si queremos, por ejemplo, pronosticar el número de calorías de una cerveza sin el conocimiento de otras variables, utilizaríamos la media del número de calorías. Pero si tenemos información sobre otra variable y del grado de relación entre ambas, es posible mejorar nuestro pronóstico. El valor R^2 del diagrama de la figura 18.2 vale 0,83, lo que indica que si conocemos el porcentaje de alcohol de una cerveza, podemos mejorar en un 83 % nuestros pronósticos sobre su número de calorías si, en lugar de utilizar como pronóstico el número medio de calorías, basamos nuestro pronóstico en el porcentaje de alcohol. Comparando este resultado con el correspondiente al diagrama de la figura 18.3 (donde R^2 vale 0,06) comprenderemos el valor informativo de R^2: en este segundo caso, el conocimiento del contenido de alcohol de una cerveza sólo nos permite mejorar nuestros pronósticos del precio en un 6 %, lo cual nos está indicando, además de que nuestros pronósticos no mejoran de forma importante, que existe un mal ajuste de la recta a la nube de puntos. Parece evidente, sin tener todavía otro tipo de información, que el porcentaje de alcohol de las cervezas está mas relacionado con el número de calorías que con su precio.

Resumen

En este primer apartado introductorio hemos aprendido que el análisis de regresión lineal es una técnica estadística que permite estudiar la relación entre una variable dependiente (VD) y una o más variables independientes (VI) con el doble propósito de: 1) averiguar en qué medida la VD puede estar explicada por la(s) VI y 2) obtener predicciones en la VD a partir de la(s) VI. El procedimiento implica, básicamente, obtener la ecuación mínimo-cuadrática que mejor expresa la relación entre la VD y la(s) VI y estimar mediante el coeficiente de determinación la calidad de la ecuación de regresión obtenida. Estos dos pasos deben ir acompañados de un chequeo del cumplimiento de las condiciones o supuestos que garantizan la validez del procedimiento.

Análisis de regresión lineal simple

Vamos a iniciar nuestro estudio más formal de la regresión con el *modelo de regresión lineal simple* (*simple* = una variable independiente), pero conviene no perder de vista que, puesto que generalmente estaremos interesados en estudiar simultáneamente más de una variable predictora, este análisis es sólo un punto de partida en nuestra explicación del análisis de regresión.

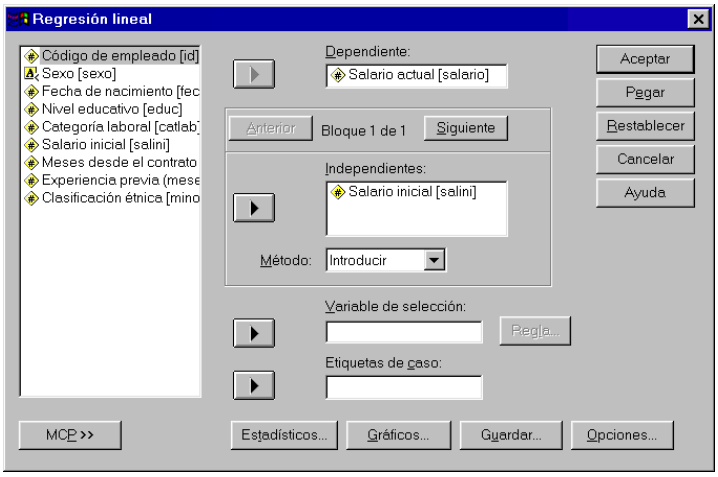
Vamos a seguir utilizando en todo momento el archivo *Datos de empleados* que, como ya sabemos, se instala con el programa en el propio directorio del SPSS. Y comenzaremos utilizando *salario* (salario actual) como variable dependiente y *salini* (salario inicial) como variable independiente o predictora.

Regresión simple

Para llevar a cabo un análisis de regresión simple con las especificaciones que el programa tiene establecidas por defecto:

- Seleccionar la opción **Regresión** > **Lineal** del menú **Analizar** para acceder al cuadro de diálogo *Regresión lineal* que muestra la figura 18.4.

Figura 18.4. Cuadro de diálogo *Regresión lineal*.



- Seleccionar la variable *salario* en la lista de variables del archivo de datos y trasladarla al cuadro **Dependiente**.
- Seleccionar la variable *salini* y trasladarla a la lista **Independientes**.

Con sólo estas especificaciones, al pulsar el botón **Aceptar** el *Visor* ofrece los resultados que muestran las tablas 18.1 a la 18.3.

Bondad de ajuste

La primera información que obtenemos (tabla 18.1) se refiere al coeficiente de correlación múltiple (R) y a su cuadrado. Puesto que sólo tenemos dos variables, el coeficiente de correlación múltiple no es otra cosa que el valor absoluto del coeficiente de correlación de Pearson entre esas dos variables (ver capítulo anterior). Su cuadrado (*R cuadrado*) es el coeficiente de determinación:

$$R^2 = 1 - \frac{\text{Suma de cuadrados de los residuos}}{\text{Suma de cuadrados total}}$$

(los residuos son las diferencias existentes entre las puntuaciones observadas y los pronósticos obtenidos con la recta). Tal como hemos señalado ya, R^2 expresa la proporción de varianza de

la variable dependiente que está explicada por la variable independiente. En nuestro ejemplo (tabla 18.1), R toma un valor muy alto (su máximo es 1); y R^2 nos indica que el 77,5 % de la variación de *salario* está explicada por *salini*. Es importante resaltar en este momento que el análisis de regresión no permite afirmar que las relaciones detectadas sean de tipo causal: sólo es posible hablar de grado de relación.

Tabla 18.1. Resumen del modelo.

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,880	,775	,774	$8,115.36

R cuadrado corregida es una corrección a la baja de R^2 que se basa en el número de casos y de variables independientes:

$$R^2_{\text{corregida}} = R^2 - [p(1-R^2)/(n-p-1)]$$

(p se refiere al número de variables independientes). En una situación con pocos casos y muchas variables independientes, R^2 puede ser artificialmente alta. En tal caso, el valor de R^2 *corregida* será sustancialmente más bajo que el de R^2. En nuestro ejemplo, como hay 474 casos y una sola variable independiente, los dos valores de R^2 (el corregido y el no corregido) son prácticamente iguales.

El *error típico de la estimación* (al que llamaremos S_e) es la desviación típica de los residuos, es decir, la desviación típica de las distancias existentes entre las puntuaciones en la variable dependiente (Y_i) y los pronósticos efectuados con la recta de regresión ($\hat{Y}_i$), aunque no exactamente, pues la suma de las distancias al cuadrado están divididas por $n-2$:

$$\text{Error típico de estimación} = S_e = \sqrt{\sum (Y_i - \hat{Y}_i)^2/(n-2)}$$

En realidad, este error típico es la raíz cuadrada de la *media cuadrática residual* de la tabla 18.2). Representa una medida de la parte de variabilidad de la variable dependiente que no es explicada por la recta de regresión. En general, cuanto mejor es el ajuste, más pequeño es este error típico.

Tabla 18.2. Resumen del ANOVA.

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	106831048750,124	1	106831048750,124	1622,118	,000
	Residual	31085446686,216	472	65858997,217		
	Total	137916495436,340	473			

La tabla resumen del ANOVA (tabla 18.2) nos informa sobre si existe o no relación significativa entre las variables. El estadístico F permite contrastar la hipótesis nula de que el valor poblacional de R es cero, lo cual, en el modelo de regresión simple, equivale a contrastar la hipótesis de que la pendiente de la recta de regresión vale cero. El nivel crítico (*Sig.*) indica que, si suponemos que el valor poblacional de R es cero, es improbable (probabilidad = 0,000) que R, en esta muestra, tome el valor 0,88. Lo cual implica que R es mayor que cero y que, en consecuencia, ambas variables están linealmente relacionadas.

Ecuación de regresión

La tabla 18.3 muestra los coeficientes de la recta de regresión. La columna etiquetada *Coeficientes no estandarizados* contiene los coeficientes de regresión parcial que definen la ecuación de regresión en puntuaciones directas.

Tabla 18.3. Coeficientes de regresión parcial.

	Coeficientes no estandarizados		Coeficientes estandarizados		
	B	Error típ.	Beta	t	Sig.
(Constante)	1928,206	888,680		2,170	,031
Salario inicial	1,909	,047	,880	40,276	,000

El coeficiente correspondiente a la *Constante* es el *origen* de la recta de regresión (lo que hemos llamado B_0):

$$B_0 = \bar{Y} - B_1\bar{X}$$

Y el coeficiente correspondiente a *Salario inicial* es la *pendiente* de la recta de regresión (lo que hemos llamado B_1):

$$B_1 = \frac{\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2}$$

B_1 indica el cambio medio que corresponde a la variable dependiente (*salario*) por cada unidad de cambio de la variable independiente (*salini*). Según esto, la ecuación de regresión queda de la siguiente manera:

Pronóstico en salario = 1928,206 + 1,909 *salini*

A cada valor de *salini* le corresponde un pronóstico en *salario* basado en un incremento constante (1928,206) más 1,909 veces el valor de *salini*.

Coeficientes de regresión estandarizados

Los coeficientes *Beta* (coeficientes de regresión parcial estandarizados) son los coeficientes que definen la ecuación de regresión cuando ésta se obtiene tras estandarizar las variables originales, es decir, tras convertir las puntuaciones directas en típicas. Se obtiene de la siguiente manera: $\beta_1 = B_1 (S_x / S_y)$.

En el análisis de regresión simple, el coeficiente de regresión estandarizado correspondiente a la única variable independiente presente en la ecuación coincide exactamente con el coeficiente de correlación de Pearson. En regresión múltiple, según veremos, los coeficientes de regresión estandarizados permiten valorar la importancia relativa de cada variable independiente dentro de la ecuación.

Pruebas de significación

Finalmente, los estadísticos *t* y sus niveles críticos (*Sig.*) nos permiten contrastar las hipótesis nulas de que los coeficientes de regresión valen cero en la población. Estos estadísticos *t* se obtienen dividiendo los coeficientes de regresión B_0 y B_1 entre sus correspondientes errores típicos:

$$t_{B_0} = \frac{B_0}{S_{B_0}} \qquad \text{y} \qquad t_{B_1} = \frac{B_1}{S_{B_1}}$$

siendo:

$$S_{B_0} = S_e \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2}} \qquad \text{y} \qquad S_{B_1} = \frac{S_e}{\sqrt{\sum (X_i - \bar{X})^2}}$$

Estos estadísticos *t* se distribuyen según el modelo de probabilidad *t* de Student con $n-2$ grados de libertad. Por tanto, pueden ser utilizados para decidir si un determinado coeficiente de regresión es significativamente distinto de cero y, en consecuencia, si la variable independiente está significativamente relacionada con la dependiente.

Puesto que en regresión simple sólo trabajamos con una variable independiente, el resultado del estadístico *t* es equivalente al del estadístico *F* de la tabla del ANOVA (de hecho, $t^2 = F$).

A partir de los resultados de la tabla 18.3, podemos llegar a las siguientes conclusiones:

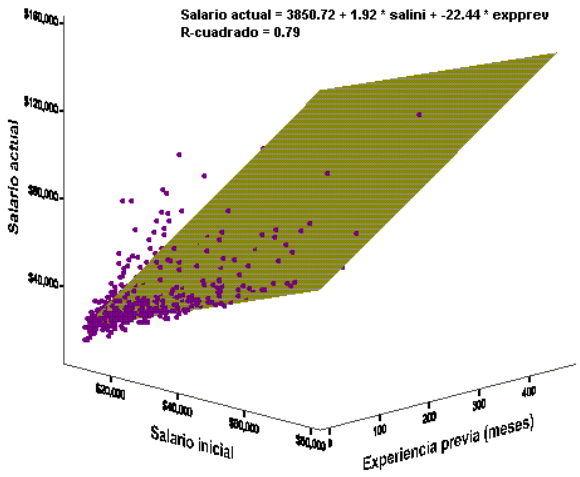
1. El *origen* poblacional de la recta de regresión (β_0) es significativamente distinto de cero (generalmente, contrastar la hipótesis "$\beta_0 = 0$" carece de utilidad, pues no contiene información sobre la relación entre X_i e Y_i).
2. La *pendiente* poblacional de la recta de regresión (el coeficiente de regresión β_1 correspondiente a *salini*) es significativamente distinta de cero, lo cual nos permite concluir que entre *salario* y *salini* existe relación lineal significativa.

Análisis de regresión lineal múltiple

El procedimiento **Regresión lineal** permite utilizar más de una variable independiente y, por tanto, permite llevar a cabo análisis de regresión múltiple. Pero en el análisis de regresión múltiple, la ecuación de regresión ya no define una recta en el plano, sino un hiperplano en un espacio multidimensional.

Imaginemos un análisis de regresión con *salario* como variable dependiente y *salini* (salario inicial) y *expprev* (experiencia previa) como variables independientes. La figura 18.5 muestra el diagrama de dispersión de *salario* sobre *salini* y *expprev*, y el plano de regresión en un espacio tridimensional.

Figura 18.5. Diagrama de dispersión de *salario* sobre *salini* y *expprev*.



Con una variable dependiente y dos independientes, necesitamos tres ejes para poder representar el correspondiente diagrama de dispersión. Y si en lugar de dos variables independientes utilizáramos tres, sería necesario un espacio de cuatro dimensiones para poder construir el diagrama de dispersión. Y un espacio de cinco dimensiones para poder construir el diagrama correspondiente a cuatro variables independientes. Etc.

Por tanto, con más de una variable independiente, la representación gráfica de las relaciones presentes en un modelo de regresión resulta poco intuitiva, muy complicada y nada útil. Es más fácil y práctico partir de la ecuación del modelo de regresión lineal:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

De acuerdo con este modelo o ecuación, la variable dependiente (Y) se interpreta como una combinación lineal de un conjunto de K variables independientes (X_k), cada una de las cuales va acompañada de un coeficiente (β_k) que indica el peso relativo de esa variable en la ecuación. La ecuación incluye además una constante (β_0) y un componente aleatorio (los residuos: ϵ) que recoge todo lo que las variables independientes no son capaces de explicar.

Este modelo, al igual que cualquier otro modelo estadístico, se basa en una serie de supuestos (linealidad, independencia, normalidad, homocedasticidad y no-colinealidad) que estudiaremos en detalle en el siguiente apartado.

La ecuación de regresión mínimo-cuadrática se construye estimando los valores de los coeficientes beta del modelo de regresión. Estas estimaciones se obtienen intentando hacer que las diferencias al cuadrado entre los valores observados (Y) y los pronosticados ($\hat{Y}$) sean mínimas:

$$\hat{Y} = B_0 + B_1 X_1 + B_2 X_2 + \cdots + B_k X_k$$

Regresión múltiple

Al igual que en el análisis de regresión simple del apartado anterior, vamos a seguir utilizando *salario* (salario actual) como variable dependiente. Pero ahora vamos a incluir 3 variables independientes en el modelo: *salini* (salario inicial), *expprev* (experiencia previa) y *educ* (nivel educativo).

Para llevar a cabo un análisis de regresión múltiple con las especificaciones que el programa tiene establecidas por defecto:

- Seleccionar la opción **Regresión** > **Lineal** del menú **Analizar** para acceder al cuadro de diálogo *Regresión lineal* que muestra la figura 18.4.
- Seleccionar la variable *salario* en la lista de variables del archivo de datos y trasladarla al cuadro **Dependiente**.
- Seleccionar las variables *salini*, *expprev* y *educ* y trasladarlas a la lista **Independientes**.

Con estas especificaciones mínimas, al pulsar el botón **Aceptar** el *Visor* ofrece la información que muestran las tablas 18.4 a la 18.6.

Bondad de ajuste

Tomadas juntas (ver tabla 18.4), las tres variables independientes incluidas en el análisis explican un 80 % de la varianza de la variable dependiente, pues R^2 *corregida* = 0,80. Además, el error típico de los residuos (8.115,36 en el análisis de regresión simple) ha disminuido algo (7.631,92 en el análisis de regresión múltiple), lo que indica una pequeña mejora en el ajuste. De nuevo, el valor corregido de R^2 es casi idéntico al valor no corregido.

Tabla 18.4. Resumen del modelo.

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,895	,802	,800	$7,631.92

El estadístico *F* (ver tabla 18.5) contrasta la hipótesis nula de que el valor poblacional de *R* es cero y, por tanto, nos permite decidir si existe relación lineal significativa entre la variable dependiente y el conjunto de variables independientes tomadas juntas. El valor del nivel crítico *Sig.* = 0,000 indica que sí existe relación lineal significativa. Podemos afirmar, por tanto, que el hiperplano definido por la ecuación de regresión ofrece un buen ajuste a la nube de puntos.

Tabla 18.5. Resumen del ANOVA.

Modelo: 1

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Regresión	110540801465,350	3	36846933821,783	632,607	,000
Residual	27375693970,990	470	58246157,385		
Total	137916495436,340	473			

Ecuación de regresión

La tabla de coeficientes de regresión parcial (ver tabla 18.6) contiene toda la información necesaria para construir la ecuación de regresión mínimo-cuadrática.

Tabla 18.6. Coeficientes de regresión parcial.

Modelo: 1

	Coeficientes no estandarizados		Coeficientes estandarizados		
	B	Error típ.	Beta	t	Sig.
(Constante)	-3661,517	1935,490		-1,892	,059
Salario inicial	1,749	,060	,806	29,198	,000
Experiencia previa (meses)	-16,730	3,605	-,102	-4,641	,000
Nivel educativo (años)	735,956	168,689	,124	4,363	,000

En la columna encabezada *Coeficientes no estandarizados* se encuentran los coeficientes (B_k) que forman parte de la ecuación en puntuaciones directas:

$$\textit{Pronóstico en salario} = \\ = -3.661{,}517 + 1{,}749\ \textit{salini} - 16{,}730\ \textit{expprev} + 735{,}956\ \textit{educ}$$

Estos coeficientes no estandarizados se interpretan en los términos ya conocidos. Por ejemplo, el coeficiente correspondiente a la variable *salini*, que vale 1,749, indica que, si el resto de variables se mantienen constantes, a un aumento de una unidad (un dólar) en *salini* le corresponde, en promedio, un aumento de 1,749 dólares en *salario*.

Es necesario señalar que estos coeficientes no son independientes entre sí. De hecho, reciben el nombre de coeficientes de regresión *parcial* porque el valor concreto estimado para cada coeficiente se ajusta teniendo en cuenta la presencia del resto de variables independientes. Conviene, por tanto, interpretarlos con cautela.

El signo del coeficiente de regresión parcial de una variable puede no ser el mismo que el del coeficiente de correlación simple entre esa variable y la dependiente. Esto es debido a los ajustes que se llevan a cabo para poder obtener la mejor ecuación posible. Aunque existen diferentes explicaciones para justificar el cambio de signo de un coeficiente de regresión, una de las que deben ser más seriamente consideradas es la que se refiere a la presencia de un alto grado de asociación entre algunas de las variables independientes (colinealidad). Trataremos esta cuestión más adelante.

Coeficientes de regresión estandarizados

Los coeficientes *Beta* están basados en las puntuaciones típicas y, por tanto, son directamente comparables entre sí. Indican la cantidad de cambio, en puntuaciones típicas, que se producirá en la variable dependiente por cada cambio de una unidad en la correspondiente variable independiente (manteniendo constantes el resto de variables independientes).

Estos coeficientes proporcionan una pista muy útil sobre la importancia relativa de cada variable independiente en la ecuación de regresión. En general, una variable tiene tanto más

peso (importancia) en la ecuación de regresión cuanto mayor (en valor absoluto) es su coeficiente de regresión estandarizado. Observando los coeficientes *Beta* de la tabla 18.6 vemos que la variable *salini* es la más importante; después, *educ*; por último, *expprev*. Lo ya dicho sobre la no independencia de los coeficientes de regresión parcial no estandarizados también vale aquí.

Pruebas de significación

Las pruebas *t* y sus niveles críticos (últimas dos columnas de la tabla 18.6: *t* y *Sig.*) sirven para contrastar la hipótesis nula de que un coeficiente de regresión vale cero en la población. Niveles críticos (*Sig.*) muy pequeños (generalmente menores que 0,05) indican que debemos rechazar esa hipótesis nula.

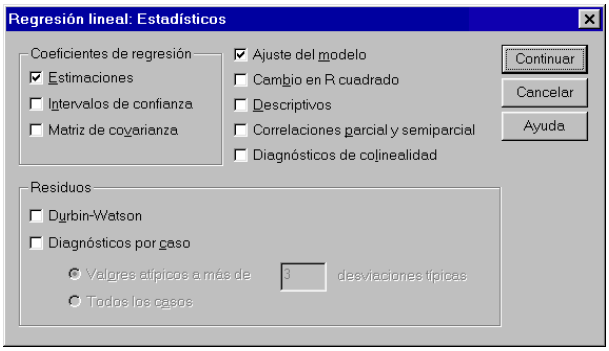
Un coeficiente de cero indica ausencia de relación lineal, de modo que los coeficientes significativamente distintos de cero nos informan sobre qué variables son relevantes en la ecuación de regresión. Observando el nivel crítico asociado a cada prueba *t* (tabla 18.6), vemos que las tres variables utilizadas poseen coeficientes significativamente distintos de cero (en todas, *Sig.* = 0,000). Todas ellas, por tanto, contribuyen de forma significativa a explicar lo que ocurre con la variable dependiente.

Información complementaria

Además de la ecuación de regresión y de la calidad de su ajuste, un análisis de regresión no debe renunciar a la obtención de algunos estadísticos descriptivos elementales como la matriz de correlaciones, la media y la desviación típica de cada variable y el número de casos con el que se está trabajando, etc. Para obtener estos estadísticos:

- Pulsar el botón **Estadísticos...** del cuadro de diálogo *Regresión lineal* (ver figura 18.4) para acceder al subcuadro de diálogo *Regresión lineal: Estadísticos* que muestra la figura 18.6.

Figura 18.6. Subcuadro de diálogo *Regresión lineal: Estadísticos*.



Entre las opciones que ofrece este subcuadro de diálogo, existen dos que se encuentran marcadas por defecto. Estas dos opciones ya marcadas son precisamente las que permiten obtener la información que recogen las tablas 18.1 a la 18.6 cuando pulsamos el botón **Aceptar** del cuadro de diálogo *Regresión lineal* (ver figura 18.4) sin hacer otra cosa que seleccionar la variable dependiente y la independiente:

- **Estimaciones**. Ofrece las estimaciones de los coeficientes de regresión parcial no estandarizados (*B*) y estandarizados (*Beta*), junto con las pruebas de significación *t* individuales para contrastar las hipótesis de que el valor poblacional de esos coeficientes es cero (ver tablas 18.3 y 18.6).
- **Ajuste del modelo**. Muestra el coeficiente de correlación múltiple, su cuadrado corregido y no corregido, y el error típico de los residuos (ver tablas 18.1 y 18.4: R, R^2, R^2 *corregida* y *error típico de la estimación*). Esta opción también incluye la tabla resumen del ANOVA, la cual contiene el estadístico F para contrastar la hipótesis $R = 0$ (ver tablas 18.2 y 18.4).

Al margen de las dos opciones, que se encuentran activas por defecto, el subcuadro de diálogo *Regresión lineal: Estadísticos* (figura 18.6) contiene varias opciones muy interesantes en un análisis de regresión:

- **Intervalos de confianza**. Esta opción, situada en el recuadro **Coeficientes de regresión**, hace que, además de una estimación puntual de los coeficientes de regresión parcial (que ya obtenemos con la opción **Estimaciones**), podamos obtener el intervalo de confianza para esos coeficientes (ver tabla 18.7).

  Estos intervalos nos informan sobre los límites entre los que podemos esperar que se encuentre el valor poblacional de cada coeficiente de regresión. Los límites se obtienen sumando y restando 1,96 errores típicos al valor del correspondiente coeficiente de regresión (decimos 1,96 porque el SPSS trabaja, por defecto, con un nivel de confianza de 0,95).

  Intervalos de confianza muy amplios indican que las estimaciones obtenidas son poco precisas y, probablemente, inestables (cosa que suele ocurrir, por ejemplo, cuando existen problemas de colinealidad; estudiaremos esta cuestión más adelante, en el apartado dedicado a los supuestos del modelo de regresión).

Tabla 18.7. Coeficientes de regresión parcial, incluyendo los Intervalos de confianza.

Modelo: 1

	Coeficientes no estandarizados		Coeficientes estandarizados			Intervalo de confianza para B al 95%	
	B	Error típ.	Beta	t	Sig.	Límite inferior	Límite superior
(Constante)	-3661,5	1935,490		-1,892	,059	-7464,803	141,768
Salario inicial	1,749	,060	,806	29,198	,000	1,631	1,866
Experiencia previa	-16,730	3,605	-,102	-4,641	,000	-23,814	-9,646
Nivel educativo	735,956	168,689	,124	4,363	,000	404,477	1067,434

- **Matriz de covarianza**. Muestra una matriz con las covarianzas y correlaciones existentes entre los coeficientes de regresión parcial (tabla 18.8). Vemos que, efectivamente, los coeficientes de regresión parcial no son independientes entre sí.

Tabla 18.8. Correlaciones entre los coeficientes de regresión.

Modelo: 1

		Nivel educativo	Experiencia previa	Salario inicial
Correlaciones	Nivel educativo	1,000	,363	-,667
	Experiencia previa	,363	1,000	-,274
	Salario inicial	-,667	-,274	1,000
Covarianzas	Nivel educativo	28456,057	220,958	-6,737
	Experiencia previa	220,958	12,997	-5,908E-02
	Salario inicial	-6,737	-5,908E-02	3,587E-03

☐ **Descriptivos**. Ofrece la media y la desviación típica de cada variable y el número de casos utilizados en el análisis (ver tabla 18.9).

Además, ofrece la matriz de correlaciones entre el conjunto de variables utilizadas en el análisis (ver tabla 18.10). En la matriz de correlaciones, cada coeficiente de correlación aparece acompañado de su correspondiente nivel crítico (el cual permite decidir sobre la hipótesis de que el coeficiente de correlación vale cero en la población) y del número de casos sobre el que se ha calculado cada coeficiente.

Lógicamente, en la diagonal de la matriz de correlaciones aparecen unos, pues la relación entre una variable y ella misma es perfecta.

Tabla 18.9. Estadísticos descriptivos.

	Media	Desviación típ.	N
Salario actual	$34,419.57	$17,075.66	474
Salario inicial	$17,016.09	$7,870.64	474
Experiencia previa	95,86	104,59	474
Nivel educativo	13,49	2,88	474

Tabla 18.10. Correlaciones entre variables.

		Salario actual	Salario inicial	Experiencia previa	Nivel educativo
Salario actual	Correlación de Pearson	1,000	,880	-,097	,661
	Sig. (unilateral)	,	,000	,017	,000
	N	474	474	474	474
Salario inicial	Correlación de Pearson	,880	1,000	,045	,633
	Sig. (unilateral)	,000	,	,163	,000
	N	474	474	474	474
Experiencia previa	Correlación de Pearson	-,097	,045	1,000	-,252
	Sig. (unilateral)	,017	,163	,	,000
	N	474	474	474	474
Nivel educativo	Correlación de Pearson	,661	,633	-,252	1,000
	Sig. (unilateral)	,000	,000	,000	,
	N	474	474	474	474

☐ **Correlaciones parcial y semiparcial**. Esta opción permite obtener los coeficientes de correlación parcial y semiparcial entre la variable dependiente y cada variable independiente.

Un coeficiente de *correlación parcial* expresa el grado de relación existente entre dos variables tras eliminar de ambas el efecto debido a terceras variables (ver capítulo 17). En el contexto del análisis de regresión, los coeficientes de correlación parcial expresan el grado de relación existente entre cada variable independiente y la variable dependiente tras eliminar de ambas el efecto debido al resto de variables independientes incluidas en la ecuación.

Un coeficiente de *correlación semiparcial* expresa el grado de relación existente entre dos variables tras eliminar de una de ellas el efecto debido a terceras variables. En el contexto del análisis de regresión, estos coeficientes expresan el grado de relación existente entre la variable dependiente y la parte de cada variable independiente que no está explicada por el resto de variables independientes.

Seleccionando la opción **Correlaciones parcial y semiparcial**, la tabla de coeficientes de regresión (tabla 18.6, ya vista) incluye la información adicional que muestra la tabla 18.11.

Tabla 18.11. Coeficientes de regresión parcial y coeficientes de correlación parcial y semiparcial.

Modelo: 1

	Coeficientes no estandarizados		Coeficientes estandarizados			Correlaciones		
	B	Error típ.	Beta	t	Sig.	Orden cero	Parcial	Semi-parcial
(Constante)	3661,517	1935,490		-1,892	,059			
Salario inicial	1,749	,060	,806	29,198	,000	,880	,803	,600
Experiencia previa	-16,730	3,605	-,102	-4,641	,000	-,097	-,209	-,095
Nivel educativo	735,956	168,689	,124	4,363	,000	,661	,197	,090

Junto con los coeficientes de correlación parcial y semiparcial, aparecen las correlaciones de *orden cero*, es decir, los coeficientes de correlación calculados sin tener en cuenta la presencia de terceras variables (se trata de los mismos coeficientes que aparecen en la tabla 18.10). Comparando entre sí estos coeficientes (de orden cero, parcial y semiparcial) pueden encontrarse pautas de relación interesantes. En los datos de la tabla 18.11 ocurre, por ejemplo, que la relación entre la variable dependiente *salario actual* y la variable independiente *nivel educativo* vale 0,661. Sin embargo, al eliminar de *salario actual* y de *nivel educativo* el efecto atribuible al resto de variables independientes (*salario inicial* y *experiencia previa*), la relación baja hasta 0,197 (parcial); y cuando el efecto atribuible a *salario inicial* y *experiencia previa* se elimina sólo de *salario actual*, la relación baja hasta 0,090 (semiparcial). Lo cual está indicando que la relación entre estas dos últimas variables podría ser espúrea, pues puede explicarse casi por completo recurriendo a las otras dos variables independientes.

El resto de opciones del subcuadro de diálogo *Regresión lineal: Estadísticos* (ver figura 18.6) tienen que ver con los supuestos del modelo de regresión lineal (*estadísticos de colinealidad, residuos*) y con el análisis de regresión por pasos (*cambio en R cuadrado*). Todas estas opciones se tratan más adelante.

Supuestos del modelo de regresión lineal

Los supuestos de un modelo estadístico se refieren a una serie de condiciones que deben darse para garantizar la validez del modelo. Al efectuar aplicaciones prácticas del modelo de regresión, nos veremos en la necesidad de examinar muchos de estos supuestos.

1. **Linealidad**. La ecuación de regresión adopta una forma particular. En concreto, la variable dependiente es la suma de un conjunto de elementos: el origen de la recta, una combinación lineal de variables independientes o predictoras y los residuos . El incumplimiento del supuesto de linealidad suele denominarse error de especificación. Algunos ejemplos son: omisión de variables independientes importantes, inclusión de variables independientes irrelevantes, no linealidad (la relación entre las variables independientes y la dependiente no es lineal), parámetros cambiantes (los parámetros no permanecen constantes durante el tiempo que dura la recogida de datos), no aditividad (el efecto de alguna variable independiente es sensible a los niveles de alguna otra variable independiente), etc.
2. **Independencia**. Los residuos son independientes entre sí, es decir, los residuos constituyen una variable aleatoria (recordemos que los residuos son las diferencias entre los valores observados y los pronosticados). Es frecuente encontrarse con residuos autocorrelacionados cuando se trabaja con series temporales.
3. **Homocedasticidad**. Para cada valor de la variable independiente (o combinación de valores de las variables independientes), la varianza de los residuos es constante.
4. **Normalidad**. Para cada valor de la variable independiente (o combinación de valores de las variables independientes), los residuos se distribuyen normalmente con media cero.
5. **No-colinealidad**. No existe relación lineal exacta entre ninguna de las variables independientes. El incumplimiento de este supuesto da origen a colinealidad o multicolinealidad.

Sobre el cumplimiento del primer supuesto puede obtenerse información a partir de una inspección del diagrama de dispersión: si tenemos intención de utilizar el modelo de regresión lineal, lo razonable es que la relación entre la variable dependiente y las independientes sea de tipo lineal (veremos que existen *gráficos parciales* que permiten obtener una representación de la relación *neta* existente entre dos variables). El quinto supuesto, *no-colinealidad*, no tiene sentido en regresión simple, pues es imprescindible la presencia de más de una variable independiente. Veremos que existen diferentes formas de diagnosticar la presencia de colinealidad. El resto de los supuestos, *independencia*, *homocedasticidad* y *normalidad*, están estrechamente asociados al comportamiento de los residuos. Por tanto, un análisis cuidadoso de los residuos puede informarnos sobre el cumplimiento de los mismos.

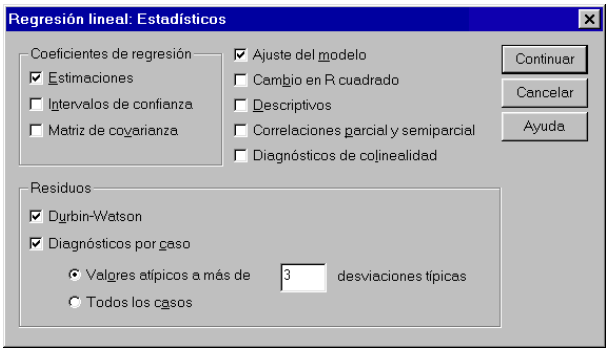
Análisis de los residuos

Llamamos residuos a las diferencias entre los valores observados y los pronosticados: $(Y_i - \hat{Y}_i)$. Pueden obtenerse marcando la opción **No tipificados** dentro del recuadro **Residuos** en el subcuadro de diálogo *Regresión lineal: Guardar nuevas variables* (ver figura 18.12, más adelante).

Los residuos son muy importantes en el análisis de regresión. En primer lugar, nos informan sobre el grado de exactitud de los pronósticos: cuanto más pequeño es el error típico de los residuos (ver tabla 18.1: *error típico de la estimación*), mejores son los pronósticos, o lo que es lo mismo, mejor se ajusta la recta de regresión a la nube de puntos. En segundo lugar, el análisis de las características de los casos con residuos grandes (sean positivos o negativos; es decir, *grandes en valor absoluto*) puede ayudarnos a detectar casos atípicos y, consecuentemente, a perfeccionar la ecuación de regresión a través de un estudio detallado de los mismos.

La opción **Diagnósticos por caso** del cuadro de diálogo *Regresión lineal: Estadísticos* (ver figura 18.6.bis) ofrece un listado de todos los residuos o, alternativamente (y esto es más interesante), un listado de los residuos que se alejan de cero (el valor esperado de los residuos) en más de un determinado número de desviaciones típicas.

Figura 18.6 (bis). Subcuadro de diálogo *Regresión lineal: Estadísticos*.



Por defecto, el SPSS lista los residuos que se alejan de cero más de 3 desviaciones típicas, pero el usuario puede cambiar este valor introduciendo el valor deseado. Para obtener un listado de los residuos que se alejan de cero más de, por ejemplo, tres desviaciones típicas:

- Marcar la opción **Diagnósticos por caso** y seleccionar **Valores atípicos a más de [3] desviaciones típicas**.

Aceptando estas selecciones, el *Visor* ofrece los resultados que muestra la tabla 18.12.

Tabla 18.12. Diagnósticos por caso (listado de los residuos más grandes en valor absoluto).

Número de caso	Residuo tipificado	Salario actual	Valor pronosticado	Residuo
307	5,981	$80,000	$34,350.68	$45,649.32
368	6,381	$103,750	$55,048.80	$48,701.20
395	-3,781	$66,750	$95,602.99	-$28,852.99
401	4,953	$83,750	$45,946.77	$37,803.23
420	3,167	$70,000	$45,829.66	$24,170.34
427	3,095	$110,625	$87,004.54	$23,620.46
438	3,485	$97,000	$70,405.22	$26,594.78
439	3,897	$91,250	$61,505.37	$29,744.63
471	3,401	$90,625	$64,666.70	$25,958.30

Los *residuos tipificados* (residuos divididos por su error típico) tienen una media de 0 y una desviación típica de 1. La tabla recoge los casos con residuos que se alejan de su media (cero) más de 3 desviaciones típicas. Si estos residuos están normalmente distribuidos (cosa que asumimos en el análisis de regresión), cabe esperar que el 95% de ellos se encuentre en el rango [−1,96, +1,96]. Y el 99,9%, en el rango [−3, +3]. Es fácil, por tanto, identificar los casos que poseen residuos grandes.

En la práctica, los casos con residuos grandes deben ser examinados para averiguar si las puntuaciones que tienen asignadas son o no correctas. Si, a pesar de tener asociados residuos grandes, las puntuaciones asignadas son correctas, conviene estudiar esos casos detenidamente para averiguar si difieren de algún modo y de forma sistemática del resto de los casos. Esto último es fácil de hacer en el SPSS pues, según veremos más adelante, es posible salvar los residuos correspondientes a cada caso como una variable más del archivo de datos.

Además de la tabla de *Diagnósticos por caso*, el *Visor* ofrece una tabla resumen con información sobre el valor máximo y mínimo, y la media y la desviación típica de los pronósticos, de los residuos, de los pronósticos tipificados y de los residuos tipificados (ver tabla 18.13). Especialmente importante es señalar que la media de los residuos vale cero.

Tabla 18.13. Estadísticos sobre los residuos.

	Mínimo	Máximo	Media	Desviación típ.	N
Pronóstico	$12,382.90	$146,851.63	$34,419.57	$15,287.30	474
Residuo	-$28,852.99	$48,701.20	$.00	$7,607.68	474
Pronóstico tipificado	-1,442	7,355	,000	1,000	474
Residuo tipificado	-3,781	6,381	,000	,997	474

Independencia

El verdadero interés de los residuos hay que buscarlo en el hecho de que el análisis de los mismos nos proporciona información crucial sobre el cumplimiento de varios supuestos del modelo de regresión lineal: independencia, homocedasticidad, normalidad y linealidad.

Uno de los supuestos básicos del modelo de regresión lineal es el de independencia entre los residuos (supuesto éste particularmente relevante cuando los datos se han recogido siguiendo una secuencia temporal). El estadístico de **Durbin-Watson** (1951) proporciona información sobre el grado de independencia existente entre ellos:

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

(e_i se refiere a los residuos: $e_i = Y_i - \hat{Y}_i$). El estadístico *DW* oscila entre 0 y 4, y toma el valor 2 cuando los residuos son independientes. Los valores menores que 2 indican autocorrelación positiva y los mayores que 2 autocorrelación negativa. Podemos asumir independencia entre los residuos cuando *DW* toma valores entre 1,5 y 2,5.

Para obtener el estadístico de **Durbin-Watson**:

- Seleccionar la opción de **Durbin-Watson** del cuadro de diálogo *Regresión lineal: Estadísticos* (ver figura 18.6.bis).

Esta elección permite obtener en la tabla 18.4 (ya vista) la información adicional que recoge la tabla 18.14.

Tabla 18.14. Resumen del modelo.

Modelo: 1

R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
,895	,802	,800	$7,631.92	1,579

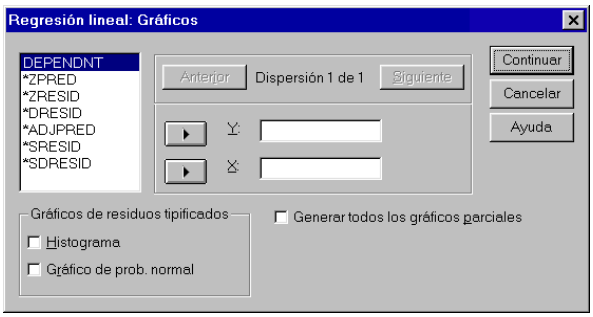
Puesto que el valor $DW = 1{,}579$ se encuentra entre 1,5 y 2,5, podemos asumir que los residuos son independientes.

Homocedasticidad

El procedimiento *Regresión lineal* dispone de una serie de gráficos que permiten, entre otras cosas, obtener información sobre el grado de cumplimiento de los supuestos de homocedasticidad y normalidad de los residuos. Para utilizar estos gráficos:

- Pulsar el botón **Gráficos...** del cuadro de diálogo *Regresión lineal* (ver figura 18.4) para acceder al subcuadro de diálogo *Regresión lineal: Gráficos* que muestra la figura 18.7.

Figura 18.7. Subcuadro de diálogo *Regresión lineal: Gráficos*.



Las variables listadas permiten obtener diferentes gráficos de dispersión. Las variables precedidas por un asterisco son variables creadas por el SPSS; todas ellas pueden crearse en el *Editor de datos* marcando las opciones pertinentes del recuadro **Residuos** del subcuadro de diálogo *Regresión lineal: Guardar nuevas variables* (ver figura 18.12 más adelante):

DEPENDENT: variable dependiente de la ecuación de regresión.

ZPRED (pronósticos típificados): pronósticos divididos por su desviación típica. Son pronósticos transformados en puntuaciones *z* (con media 0 y desviación típica 1).

ZRESID (residuos típificados): residuos divididos por su desviación típica. El tamaño de cada residuo tipificado indica el número de desviaciones típicas que se aleja de su media, de modo que, si están normalmente distribuidos (cosa que asumimos en el análisis de regresión), el 95 % de estos residuos se encontrará en el rango (−1,96, +1,96), lo cual permite identificar fácilmente casos con residuos grandes.

DRESID (residuos eliminados o corregidos): residuos obtenidos al efectuar los pronósticos eliminando de la ecuación de regresión el caso sobre el que se efectúa el pronóstico. El residuo correspondiente a cada caso se obtiene a partir del pronóstico efectuado con una ecuación de regresión en la que no se ha incluido ese caso. Son muy útiles para detectar puntos de influencia (casos con gran peso en la ecuación de regresión).

ADJPRED (pronósticos corregidos): pronósticos efectuados con una ecuación de regresión en la que no se incluye el caso pronosticado (ver residuos eliminados o corregidos). Diferencias importantes entre PRED y ADJPRED delatan la presencia de puntos de influencia (casos con gran peso en la ecuación de regresión).

SRESID (residuos estudentizados): residuos divididos por su desviación típica, basada ésta en cómo de próximo se encuentra un caso a su(s) media(s) en la(s) variable(s) independiente(s). Al igual que ocurre con los residuos estandarizados (a los que se parecen mucho), los estudentizados están escalados en unidades de desviación típica. Se distribuyen según el modelo de probabilidad t de *Student* con $n-p-1$ grados de libertad (p se refiere al número de variables independientes). Con muestras grandes, aproximadamente el 95 % de estos residuos debería encontrarse en el rango (−2, +2).

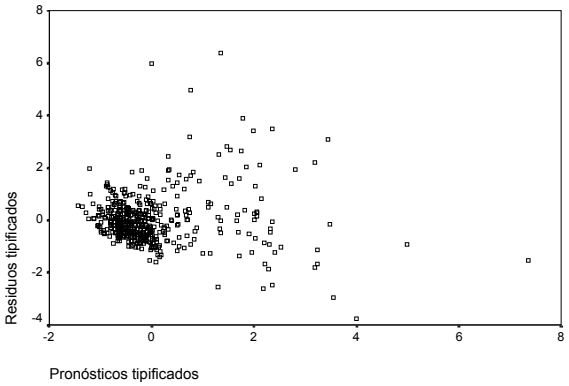
SDRESID (residuos corregidos estudentizados): residuos corregidos divididos por su desviación típica. Útiles también para detectar puntos de influencia.

Algunas de estas variables permiten identificar puntos de influencia (los estudiaremos más adelante), pero hay, entre otras, dos variables cuyo diagrama de dispersión informa sobre el supuesto de homocedasticidad o igualdad de varianzas: ZPRED y ZRESID. El supuesto de igualdad de varianzas implica que la variación de los residuos debe ser uniforme en todo el rango de valores pronosticados. O, lo que es lo mismo, que el tamaño de los residuos es independiente del tamaño de los pronósticos, de donde se desprende que el diagrama de dispersión no debe mostrar ninguna pauta de asociación entre los pronósticos y los residuos. Para obtener un diagrama de dispersión con las variables ZPRED y ZRESID:

- Trasladar la variable ZRESID al cuadro **Y:** del recuadro **Dispersión 1 de 1**.
- Trasladar la variable ZPRED al cuadro **X:** del recuadro **Dispersión 1 de 1**.

Aceptando estas elecciones, el *Visor* ofrece el diagrama de dispersión que muestra la figura 18.8. Observando el diagrama de dispersión podemos ver que, aunque los residuos y los pronósticos parecen ser independientes (pues la nube de puntos no sigue ninguna pauta de asociación clara, ni lineal ni de otro tipo), no está claro que las varianzas sean homogéneas. Más bien parece que conforme va aumentando el valor de los pronósticos también lo va haciendo la dispersión de los residuos: los pronósticos menores que la media (con puntuación típica por debajo de cero) están más concentrados que los pronósticos mayores que la media (con puntuación típica mayor que cero).

Figura 18.8. Diagrama de dispersión de *pronósticos tipificados* por *residuos tipificados*.



Cuando un diagrama de dispersión delata la presencia de varianzas heterogéneas, puede utilizarse una transformación de la variable dependiente para resolver el problema (tal como una transformación *logarítmica* o una trasformación *raíz cuadrada*). No obstante, al utilizar una transformación de la variable dependiente, no deben descuidarse los problemas de interpretación que añade el cambio de escala.

El diagrama de dispersión de las variables ZPRED y ZRESID posee la utilidad adicional de permitir detectar relaciones de tipo no lineal entre las variables. Si la relación es, de hecho, no lineal, el diagrama puede contener indicios sobre otro tipo de función de ajuste: por ejemplo, los residuos estandarizados podrían, en lugar de estar homogéneamente dispersos, seguir un trazado curvilíneo.

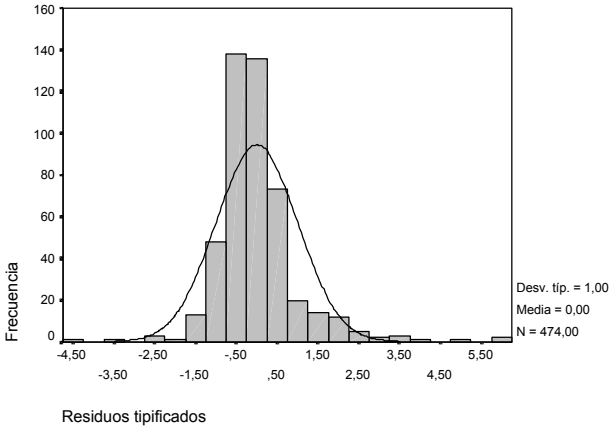
Normalidad

El recuadro **Gráficos de los residuos tipificados** (ver figura 18.7) contiene dos opciones gráficas que informan sobre el grado en el que los residuos tipificados se aproximan a una distribución normal:

☐ **Histograma**. Ofrece un histograma de los residuos tipificados con una curva normal superpuesta (figura 18.9). La curva se construye tomando una media de 0 y una desviación típica de 1, es decir, la misma media y la misma desviación típica que los residuos tipificados.

En el histograma de la figura 18.9 podemos observar, en primer lugar, que la parte central de la distribución acumula muchos más casos de los existen en una curva normal. En segundo lugar, la distribución es algo asimétrica: en la cola positiva de la distribución existen valores más extremos que en la negativa (esto ocurre cuando uno o varios errores muy grandes, correspondientes por lo general a valores atípicos, son contrarrestados con muchos residuos pequeños de signo opuesto). La distribución de los residuos, por tanto, no parece seguir el modelo de probabilidad normal, de modo que los resultados del análisis deben ser interpretados con cautela.

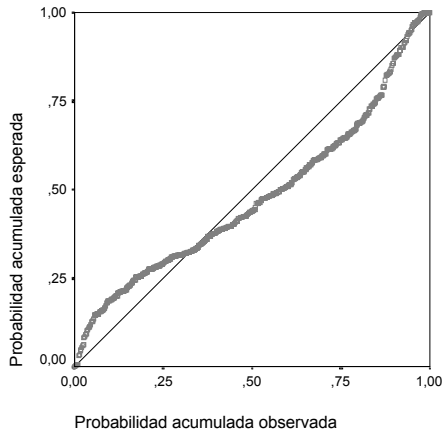
Figura 18.9. *Histograma* de los residuos tipificados.



☐ **Gráfico de prob. normalidad**. Permite obtener un diagrama de probabilidad normal. En el eje de abscisas está representada la probabilidad acumulada que corresponde a cada residuo tipificado. El de ordenadas representa la probabilidad acumulada teórica que corresponde a cada puntuación típica en una curva normal con media 0 y desviación típica 1. Cuando los residuos se distribuyen normalmente, la nube de puntos se encuentra alineada sobre la diagonal del gráfico (ver figura 18.10).

El gráfico de probabilidad normal de la figura 18.10 muestra información similar a la ya obtenida con el histograma de la figura 18.9. Los puntos no se encuentran alineados sobre la diagonal del gráfico, lo cual nos está avisando de nuevo del posible incumplimiento del supuesto de normalidad. Tal vez convenga recordar aquí que el procedimiento **Explorar** del SPSS (ver capítulo 11) contiene estadísticos que permiten contrastar la hipótesis de normalidad.

Figura 18.10. *Gráfico de probabilidad normal* de los residuos.



Linealidad

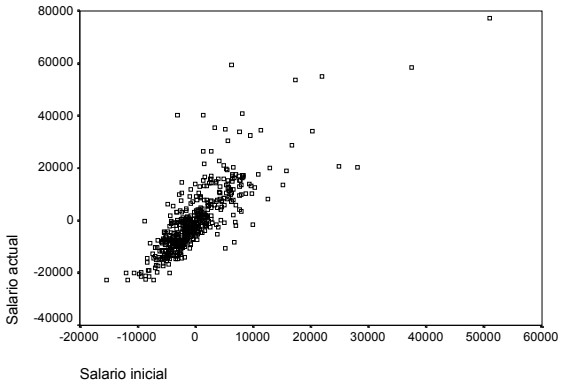
Los diagramas de regresión parcial permiten formarse una idea rápida sobre la forma que adopta una relación. En el contexto del análisis de regresión, permiten examinar la relación existente entre la variable dependiente y cada una de las variables independientes por separado, tras eliminar de ellas el efecto del resto de las variables independientes incluidas en el análisis.

Estos diagramas son similares a los de dispersión ya estudiados, pero no están basados en las puntuaciones originales de las dos variables representadas, sino en los residuos obtenidos al efectuar un análisis de regresión con el resto de las variables independientes. Por ejemplo, en el diagrama de regresión parcial de *salario actual* y *salario inicial* están representados los residuos que resultan de efectuar un análisis de regresión sobre *salario actual* incluyendo todas las variables independientes excepto *salario inicial*, y los residuos que resultan de efectuar un análisis de regresión sobre *salario inicial* incluyendo el resto de variables independientes. La utilidad de estos diagramas está en que, puesto que se controla el efecto del resto de las variables, muestran la relación *neta* entre las variables representadas. Además, las rectas que mejor se ajustan a la nube de puntos de estos diagramas son las definidas por los correspondientes coeficientes de regresión (es justamente en esa nube de puntos en la que se basan los coeficientes de regresión parcial). Para obtener estos diagramas de regresión parcial:

- Marcar la opción **Generar todos los gráficos parciales** del subcuadro de diálogo *Regresión lineal: Gráficos* (ver figura 18.7).

Esta opción genera tantos gráficos parciales como variables independientes se hayan incluido en el análisis. En nuestro ejemplo, por tanto, aparecerán tres de estos gráficos. La figura 18.11 muestra uno de ellos. Podemos observar que la relación entre *salini* (una de las variables independientes) y *salario* (la variable dependiente), tras eliminar el efecto del resto de variables independientes, es claramente lineal y positiva.

Figura 18.11. Gráfico de regresión parcial (*salario* por *salini*).



Así pues, los diagramas de regresión parcial permiten formarse una rápida idea sobre el tamaño y el signo de los coeficientes de regresión parcial (los coeficientes de la ecuación de regresión). En estos diagramas, los valores extremos pueden resultar muy informativos.

Colinealidad

Existe colinealidad perfecta cuando una de las variables independientes se relaciona de forma perfectamente lineal con una o más del resto de variables independientes de la ecuación. Esto ocurre, por ejemplo, cuando se utilizan como variables independientes en la misma ecuación las puntuaciones de las subescalas de un test y la puntuación total en el test (que es la suma de las subescalas y, por tanto, una combinación lineal perfecta de las mismas). Hablamos de colinealidad parcial o, simplemente, colinealidad, cuando entre las variables independientes de una ecuación existen correlaciones altas. Se puede dar, por ejemplo, en una investigación de mercados al tomar registros de muchos atributos de un mismo producto; o al utilizar muchos indicadores económicos para construir una ecuación de regresión. En términos generales, cuantas más variables hay en una ecuación, más fácil es que exista colinealidad (aunque, en principio, bastan dos variables).

La colinealidad es un problema porque, en el caso de colinealidad perfecta, no es posible estimar los coeficientes de la ecuación de regresión; y en el caso de colinealidad parcial, aumenta el tamaño de los residuos típificados y esto produce coeficientes de regresión muy inestables: pequeños cambios en los datos (añadir o quitar un caso, por ejemplo) produce cambios muy grandes en los coeficientes de regresión. Esta es una de las razones por las que podemos encontrarnos con coeficientes con signo cambiado: correlaciones positivas pueden transformarse en coeficientes de regresión negativos (incluso significativamente negativos). Curiosamente, la medida de ajuste R^2 no se altera por la presencia de colinealidad; pero los efectos atribuidos a las variables independientes pueden ser engañosos.

Al evaluar la existencia o no de colinealidad, la dificultad estriba precisamente en determinar cuál es el grado máximo de relación permisible entre las variables independientes. No existe un consenso generalizado sobre esta cuestión, pero puede servirnos de guía la presencia de ciertos indicios que podemos encontrar en los resultados de un análisis de regresión (estos indicios, no obstante, pueden tener su origen en otras causas):

- El estadístico *F* que evalúa el ajuste general de la ecuación de regresión es significativo, pero no lo es ninguno de los coeficientes de regresión parcial.
- Los coeficientes de regresión parcial estandarizados (los coeficientes *beta*) están inflados tanto en positivo como en negativo (adoptan, al mismo tiempo, valores mayores que 1 y menores que −1).
- Existen valores de tolerancia pequeños (próximos a 0,01). La tolerancia de una variable independiente es la proporción de varianza de esa variable que no está asociada (que no depende) del resto de variables independientes incluidas en la ecuación. Una variable con una tolerancia de, por ejemplo, 0,01 es una variable que comparte el 99 % de su varianza con el resto de variables independientes, lo cual significa que se trata de una variable redundante casi por completo.
- Los coeficientes de correlación estimados son muy grandes (por encima de 0,90 en valor absoluto).

Las afirmaciones del tipo “inflados”, “próximos a cero”, “muy grandes” se deben al hecho de que no existe un criterio estadístico formal en el que basar nuestras decisiones. Sólo existen recomendaciones basadas en trabajos de simulación.

Al margen de estos indicios, el SPSS ofrece la posibilidad de obtener algunos estadísticos que pueden ayudar a diagnosticar la presencia de colinealidad. Se trata de estadísticos orientativos que, aunque pueden ayudarnos a determinar si existe mayor o menor grado de colinealidad, no permiten tomar una decisión clara sobre la presencia o no de colinealidad. Para obtener estos estadísticos:

- Seleccionar la opción **Diagnósticos de colinealidad** del subcuadro de diálogo *Regresión lineal: Estadísticos* (ver figura 18.6.bis).

Esta opción permite obtener los estadísticos de colinealidad que recogen las tablas 18.15 y 18.16. La tabla 18.15 es la tabla de coeficientes de regresión parcial ya vista, pero ahora contiene información adicional sobre los niveles de tolerancia y sus inversos (FIV).

El nivel de *tolerancia* de una variable se obtiene restando a 1 el coeficiente de determinación (R^2) que resulta al regresar esa variable sobre el resto de variables independientes. Valores de tolerancia muy pequeños indican que esa variable puede ser explicada por una combinación lineal del resto de variables, lo cual significa que existe colinealidad.

Los *factores de inflación de la varianza* (FIV) son los inversos de los niveles de tolerancia. Reciben ese nombre porque son utilizados en el cálculo de las varianzas de los coeficientes de regresión. Cuanto mayor es el FIV de una variable, mayor es la varianza del correspondiente coeficiente de regresión. De ahí que uno de los problemas de la presencia de colinealidad (tolerancias pequeñas, FIVs grandes) sea la inestabilidad de las estimaciones de los coeficientes de regresión.

Tabla 18.15. Coeficientes de regresión parcial y niveles de tolerancia.

Modelo: 1

	Coeficientes no estandarizados		Coeficientes estandarizados			Estadísticos de colinealidad	
	B	Error típ.	Beta	t	Sig.	Tolerancia	FIV
(Constante)	-3661,5	1935,49		-1,892	,059		
Salario inicial	1,749	,060	,806	29,198	,000	,554	1,804
Experiencia previa	-16,730	3,605	-,102	-4,641	,000	,866	1,154
Nivel educativo	735,956	168,689	,124	4,363	,000	,520	1,923

La tabla 18.16 muestra la solución resultante de aplicar un análisis de componentes principales a la matriz estandarizada no centrada de productos cruzados de las variables independientes.

Tabla 18.16. Diagnósticos de colinealidad.

Modelo: 1

			Proporciones de la varianza			
Dimensión	Autovalor	Indice de condición	(Constante)	Salario inicial	Experiencia previa	Nivel educativo
1	3,401	1,000	,00	,01	,02	,00
2	,489	2,638	,00	,01	,79	,00
3	9,663E-02	5,933	,11	,62	,01	,01
4	1,347E-02	15,892	,88	,36	,18	,98

Los *autovalores* informan sobre cuántas dimensiones o factores diferentes subyacen en el conjunto de variables independientes utilizadas. La presencia de varios autovalores próximos a cero indica que las variables independientes están muy relacionadas entre sí (colinealidad).

Los *índices de condición* son la raíz cuadrada del cociente entre el autovalor más grande y cada uno del resto de los autovalores. En condiciones de no-colinealidad, estos índices no deben superar el valor 15. Índices mayores que 15 indican un posible problema. Índices mayores que 30 delatan un serio problema de colinealidad.

Las *proporciones de varianza* recogen la proporción de varianza de cada coeficiente de regresión parcial que está explicada por cada dimensión o factor. En condiciones de no-colinealidad, cada dimensión suele explicar gran cantidad de varianza de un sólo coeficiente (excepto en lo que se refiere al coeficiente B_0 o *constante*, que siempre aparece asociado a uno de los otros coeficientes; en el ejemplo, el término constante aparece asociado al coeficiente de *Nivel educativo*). La colinealidad es un problema cuando una dimensión o factor con un *índice de condición* alto, contribuye a explicar gran cantidad de la varianza de los coeficientes de dos o más variables.

Si se detecta la presencia de colinealidad en un conjunto de datos, hay que aplicar algún tipo de remedio. A continuación proponemos algunos: aumentar el tamaño de la muestra (esta solución puede resultar útil si existen pocos casos en relación al número de variables); crear indicadores múltiples combinando variables (por ejemplo, promediando variables; o efectuando un análisis de componentes principales para reducir las variables a un conjunto de componentes independientes, y aplicando después el análisis de regresión sobre esos componentes); excluir variables redundantes (es decir, excluir variables que correlacionan muy alto con otras, quedándonos con las que consideremos más importantes); utilizar una técnica de estimación sesgada, tal como la regresión *ridge*.

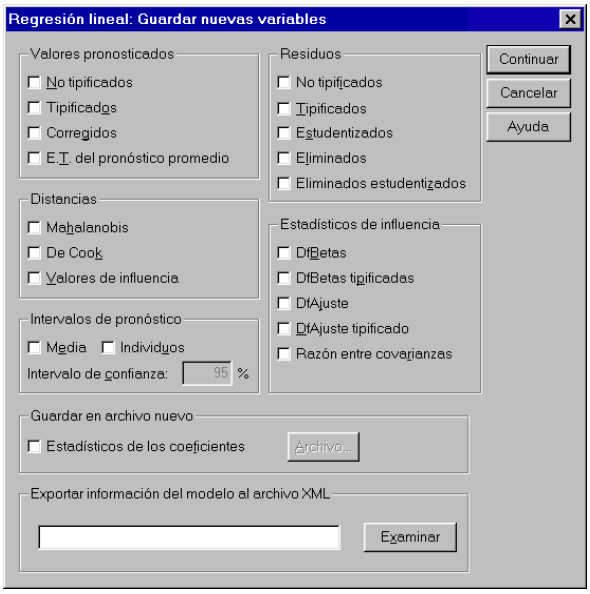
Puntos de influencia

Todos los casos contribuyen a la obtención de la ecuación de regresión, pero no todos lo hacen con la misma fuerza. Los puntos de influencia son casos que afectan de forma importante al valor de la ecuación de regresión. La presencia de puntos de influencia no tiene por qué constituir un problema en regresión: de hecho, lo normal es que en un análisis de regresión no todos los casos tengan la misma importancia (desde el punto de vista estadístico). Sin embargo, el analista debe ser consciente de la presencia de tales puntos pues, entre otras cosas, podría tratarse de casos con valores erróneos. Sólo siendo conscientes de si existen o no puntos de influencia es posible corregir el análisis.

El procedimiento **Regresión lineal** ofrece varias medidas para detectar la presencia de puntos de influencia. Para obtenerlas:

- Pulsar el botón **Guardar...** del cuadro de diálogo *Regresión lineal* (ver figura 18.4) para acceder al subcuadro de diálogo *Regresión lineal: Guardar variables* que muestra la figura 18.12.
- Marcar todas las opciones de los recuadros **Distancias** y **Estadísticos de influencia** (todas estas opciones crean variables nuevas en el archivo de datos).

Figura 18.12. Subcuadro de diálogo *Regresión lineal: Guardar nuevas variables*.



Distancias. Este recuadro recoge tres medidas que expresan el grado en que cada caso se aleja de los demás:

- **Mahalanobis**. La distancia de Mahalanobis mide el grado de distanciamiento de cada caso respecto de los promedios del conjunto de variables independientes. En regresión simple, esta distancia se obtiene simplemente elevando al cuadrado la puntuación típica de cada caso en la variable independiente. En regresión múltiple se obtiene multiplicando por $n-1$ el valor de influencia de cada caso (ver más abajo).
- **Cook**. La distancia de Cook (1977) mide el cambio que se produce en las estimaciones de los coeficientes de regresión al ir eliminando cada caso de la ecuación de regresión. Una distancia de Cook grande indica que ese caso tiene un peso considerable en la estimación de los coeficientes de regresión. Para evaluar estas distancias puede utilizarse la distribución F con $p+1$ y $n-p-1$ grados de libertad (p se refiere al número de variables independientes y n al tamaño de la muestra). En general, un caso con una distancia de Cook superior a 1 debe ser revisado.
- **Valores de influencia**. Representan una medida de la influencia potencial de cada caso. Referido a las variables independientes, un valor de influencia es una medida normalizada del grado de distanciamiento de un punto respecto del centro de su distribución. Los puntos muy alejados pueden influir de forma muy importante en la ecuación de regresión, pero no necesariamente tienen por qué hacerlo.

  Con más de 6 variables y al menos 20 casos, se considera que un valor de influencia debe ser revisado si es mayor que $3p/n$, siendo p el número de variables y n el tamaño de la muestra. Los valores de influencia tienen un máximo de $(n-1)/n$. Como regla general para orientar nuestras decisiones, los valores menores que 0,2 se consi-

deran poco problemáticos; los valores comprendidos entre 0,2 y 0,5 se consideran arriesgados; y los valores mayores que 0,5 deberían evitarse.

Estadísticos de influencia. Este recuadro contiene varios estadísticos que contribuyen a precisar la posible presencia de puntos de influencia:

- ☐ **DfBetas** (diferencia en las betas). Mide el cambio que se produce en los coeficientes de regresión estandarizados (betas) como consecuencia de ir eliminando cada caso de la ecuación de regresión. El SPSS crea en el *Editor de datos* tantas variables nuevas como coeficientes beta tiene la ecuación de regresión (es decir, tantos como variables independientes más uno, el correspondiente a la constante de la ecuación).
- ☐ **DfBetas tipificadas**. Es el cociente entre *Dfbetas* (párrafo anterior) y su error típico. Generalmente, un valor mayor que $2/\sqrt{n}$ delata la presencia de un posible punto de influencia. El SPSS crea en el *Editor de datos* tantas variables nuevas como coeficientes beta tiene la ecuación de regresión.
- ☐ **Df Ajuste** (diferencia en el ajuste). Mide el cambio que se produce en el pronóstico de un caso cuando ese caso es eliminado de la ecuación de regresión.
- ☐ **Df Ajuste tipificado**. Es el cociente entre *DfAjuste* (párrafo anterior) y su error típico. En general, se consideran puntos de influencia los casos en los que *DfAjuste tipificado* es mayor que $2/\sqrt{(p/n)}$, siendo p el número de variables independientes y n el tamaño de la muestra.
- ☐ **Razón entre las covarianzas (RV)**. Indica en qué medida la matriz de productos cruzados (base del análisis de regresión) cambia con la eliminación de cada caso. Se considera que un caso es un punto de influencia si el valor absoluto de RV−1 es mayor que $3+p/n$.

Además de crear las variables correspondientes a cada una de estas opciones, el SPSS ofrece una tabla resumen (ver tabla 18.17) que incluye, para todos los estadísticos del recuadro **Distancias** (ver figura 18.12), el valor mínimo, el máximo, la media, la desviación típica y el número de casos. La tabla también recoge información sobre los pronósticos y los residuos.

Tabla 18.17. Estadísticos sobre los residuos.

	Mínimo	Máximo	Media	Desviación típ.	N
Valor pronosticado	$12,382.90	$146,851.63	$34,419.57	$15,287.30	474
Valor pronosticado tip.	-1,442	7,355	,000	1,000	474
Error típico del valor pronosticado	$375.76	$3,186.70	$645.15	$274.71	474
Valor pronosticado corregido	$12,275.05	$149,354.23	$34,425.49	$15,356.14	474
Residual	-$28,852.99	$48,701.20	$.00	$7,607.68	474
Residuo tip.	-3,781	6,381	,000	,997	474
Residuo estud.	-3,898	6,401	,000	1,003	474
Residuo eliminado	-$30,675.46	$48,999.29	-$5.92	$7,704.67	474
Residuo eliminado estud.	-3,959	6,692	,002	1,015	474
Dist. de Mahalanobis	,149	81,468	2,994	5,172	474
Distancia de Cook	,000	,240	,003	,015	474
Valor de influencia centrada	,000	,172	,006	,011	474

Conviene señalar que los puntos de influencia no tienen por qué tener residuos particularmente grandes, por lo que el problema que plantean no es precisamente de falta de ajuste. No obstante, es muy aconsejable examinarlos por su desproporcionada influencia sobre la ecuación de regresión. Puesto que estos puntos son distintos de los de demás, conviene precisar en qué son distintos.

Una vez identificados y examinados, podríamos eliminarlos del análisis simplemente porque entorpecen el ajuste, o porque su presencia nos está haciendo obtener medidas de ajuste infladas. También podríamos eliminar los casos muy atípicos simplemente argumentando que nuestro objetivo es construir una ecuación para entender lo que ocurre con los casos típicos, corrientes, no con los casos atípicos. Este argumento es más convincente si los casos atípicos representan a una subpoblación especial que se sale del rango de variación normal. Por otro lado, si existe un conjunto de casos que parece formar un subgrupo separado del resto, podría considerarse la posibilidad de incorporar este hecho al modelo de regresión mediante una variable dummy o desarrollando diferentes ecuaciones de regresión para los diferentes subgrupos.

Los estadísticos no se ponen de acuerdo sobre la conveniencia de eliminar o no un caso, pero puede ayudarnos a decidir sobre esto el pensar que la eliminación de un caso cualquiera debe ser justificada ante quien nos pregunte por las razones de tal eliminación.

Análisis de regresión por pasos (regresión *stepwise*)

En los apartados previos hemos utilizado un método de regresión en el que el control sobre las variables utilizadas para construir el modelo de regresión recae sobre el propio analista. Es el analista quien *decide* qué variables independientes desea incluir en la ecuación de regresión seleccionándolas en la lista **Independientes**.

Sin embargo, no es infrecuente encontrarse con situaciones en las que, existiendo un elevado número de posibles variables independientes, no existe una teoría o un trabajo previo que oriente al analista en la elección de las variables relevantes. Este tipo de situaciones pueden afrontarse utilizando procedimientos diseñados para seleccionar, entre una gran cantidad de variables, sólo un conjunto reducido de las mismas: aquellas que permiten obtener el mejor ajuste posible.

Con estos procedimientos de selección, el control sobre las variables que han de formar parte de la ecuación de regresión pasa de las manos del investigador a una regla de decisión basada en criterios estadísticos.

Criterios de selección de variables

Existen diferentes criterios estadísticos para seleccionar variables en un modelo de regresión. Algunos de estos criterios son: el valor del coeficiente de correlación múltiple R^2 (corregido o sin corregir), el valor del coeficiente de correlación parcial entre cada variable independiente y la dependiente, el grado de reducción que se obtiene en el error típico de los residuos al incorporar una variable, etc. De una u otra forma, todos ellos coinciden en intentar maximizar el ajuste del modelo de regresión utilizando el mínimo número posible de variables.

Los métodos por pasos que incluye el SPSS (ver siguiente apartado) basan la selección de variables en dos criterios estadísticos:

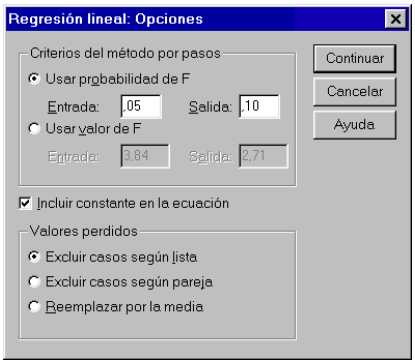
1. **Criterio de significación**. De acuerdo con este criterio, sólo se incorporan al modelo de regresión aquellas variables que contribuyen de forma significativa al ajuste del modelo. La contribución individual de una variable al ajuste del modelo se establece contrastando, a partir del coeficiente de correlación parcial, la hipótesis de independencia entre esa variable y la variable dependiente. Para decidir si se mantiene o rechaza esa hipótesis de independencia, el SPSS incluye dos criterios de significación:

   - *Probabilidad de F*. Una variable pasa a formar parte del modelo de regresión si el nivel crítico asociado a su coeficiente de correlación parcial al contrastar la hipótesis de independencia es menor que 0,05 (probabilidad de *entrada*). Y queda fuera del modelo de regresión si ese nivel crítico es mayor que 0,10 (probabilidad de *salida*).
   - *Valor de F*. Una variable pasa a formar parte del modelo de regresión si el valor del estadístico *F* utilizado para contrastar la hipótesis de independencia es mayor que 3,84 (valor de *entrada*). Y queda fuera del modelo si el valor del estadístico *F* es menor que 2,71 (valor de *salida*).

   Para modificar estos criterios de selección:

   ▶ Pulsar el botón **Opciones** del cuadro de diálogo *Regresión lineal* (ver figura 18.4) para acceder al subcuadro de diálogo *Regresión lineal: Opciones* que muestra la figura 18.13.

Figura 18.13. Subcuadro de diálogo *Regresión lineal: Opciones*.



Las opciones del recuadro **Criterios del método por pasos** permiten seleccionar uno de los dos criterios de significación disponibles y modificar las probabilidades de entrada y salida.

2. **Criterio de tolerancia**. Superado el criterio de *significación*, una variable sólo pasa a formar parte del modelo si su nivel de tolerancia es mayor que el nivel establecido por defecto (este nivel es 0,0001, pero puede cambiarse mediante sintaxis) y si, además, aun correspondiéndole un coeficiente de correlación parcial significativamente distinto de cero, su

incorporación al modelo hace que alguna de las variables previamente seleccionadas pase a tener un nivel de tolerancia por debajo del nivel establecido por defecto. El concepto de tolerancia se ha descrito ya en el apartado sobre colinealidad.

Una forma muy intuitiva de comprender y valorar el efecto resultante de aplicar estos criterios de selección consiste en observar el cambio que se va produciendo en el coeficiente de determinación R^2 a medida que se van incorporando (o eliminando) variables al modelo. Podemos definir este cambio como $R^2_{\text{cambio}} = R^2 - R^2_i$, donde R^2_i se refiere al coeficiente de determinación obtenido con todas las variables independientes excepto la *i*-ésima. Un cambio grande en R^2 indica que esa variable contribuye de forma importante a explicar lo que ocurre con la variable dependiente. Para obtener los valores de R^2_{cambio} y su significación (el grado en que el cambio observado en R^2 difiere de cero):

- Marcar la opción **Cambio en R cuadrado** del cuadro de diálogo *Regresión lineal: Estadísticos* (ver figura 18.6).

Según veremos, estas opciones permiten obtener el valor de R^2_{cambio} resultante de la incorporación de cada variable independiente, el valor del estadístico *F* al contrastar la hipótesis de que el valor poblacional de R^2_{cambio} es cero, y el nivel crítico asociado al estadístico *F*.

Métodos de selección de variables

Existen diferentes métodos para seleccionar las variables independientes que debe incluir un modelo de regresión, pero los que mayor aceptación han recibido son los métodos de selección por pasos (*stepwise*). Con estos métodos, se selecciona en primer lugar la *mejor* variable (siempre de acuerdo con algún criterio estadístico); a continuación, la mejor de las restantes; y así sucesivamente hasta que ya no quedan variables que cumplan los criterios de selección.

El procedimiento *Regresión lineal* del SPSS incluye varios de estos métodos de selección de variables. Todos ellos se encuentran disponibles en el botón de menú desplegable de la opción **Método** del cuadro de diálogo *Regresión lineal* (ver figura 18.4). Dos de estos métodos permiten incluir o excluir, en un sólo paso, todas las variables independientes seleccionadas (no son métodos de selección por pasos):

- **Introducir**. Este método construye la ecuación de regresión utilizando todas las variables seleccionadas en la lista **Independientes** (ver figura 18.4). Es el método utilizado por defecto.
- **Eliminar**. Elimina en un sólo paso todas las variables de la lista **Independientes** y ofrece los coeficientes de regresión que corresponderían a cada variable en el caso de que pasaran a formar parte de la ecuación de regresión.

El resto de métodos de selección de variables son métodos por pasos, es decir, métodos que van incorporando o eliminando variables paso a paso dependiendo de que éstas cumplan o no los criterios de selección:

- **Hacia adelante**. Las variables se incorporan al modelo de regresión una a una. En el primer paso se selecciona la variable independiente que, además de superar los criterios de *entrada*, más alto correlaciona (positiva o negativamente) con la dependiente. En los si-

guientes pasos se utiliza como criterio de selección el coeficiente de correlación parcial: van siendo seleccionadas una a una las variables que, además de superar los criterios de *entrada*, poseen el coeficiente de correlación parcial más alto en valor absoluto (la relación se parcializa controlando el efecto de las variables independientes previamente seleccionadas).

La selección de variables se detiene cuando no quedan variables que superen el criterio de *entrada*. (Utilizar como criterio de *entrada* el tamaño, en valor absoluto, del coeficiente de correlación parcial, es equivalente a seleccionar la variable con menor *probabilidad de F* o mayor *valor de F*).

- **Hacia atrás**. Comienza incluyendo en el modelo todas las variables seleccionadas en la lista **Independientes** (ver figura 18.4) y luego procede a eliminarlas una a una. La primera variable eliminada es aquella que, además de cumplir los criterios de *salida*, posee el coeficiente de regresión más bajo en valor absoluto. En cada paso sucesivo se van eliminado las variables con coeficientes de regresión no significativos, siempre en orden inverso al tamaño de su nivel crítico.

  La eliminación de variables se detiene cuando no quedan variables en el modelo que cumplan los criterios de salida.

- **Pasos sucesivos**. Este método es una especie de mezcla de los métodos *hacia adelante* y *hacia atrás*. Comienza, al igual que el método *hacia adelante*, seleccionando, en el primer paso, la variable independiente que, además de superar los criterios de *entrada*, más alto correlaciona (en valor absoluto) con la variable dependiente. A continuación, selecciona la variable independiente que, además de superar los criterios de *entrada*, posee el coeficiente de correlación parcial más alto (en valor absoluto). Cada vez que se incorpora una nueva variable al modelo, las variables previamente seleccionadas son, al igual que en el método *hacia atrás*, evaluadas nuevamente para determinar si siguen cumpliendo o no los criterios de *salida*. Si alguna variable seleccionada cumple los criterios de salida, es eliminada del modelo.

  El proceso se detiene cuando no quedan variables que superen los criterios de *entrada* y las variables seleccionadas no cumplen los criterios de *salida*.

Regresión por pasos

Para ilustrar el funcionamiento del análisis de regresión por pasos, vamos a presentar un ejemplo con el método *pasos sucesivos*. Utilizaremos el salario actual (*salario*) como variable dependiente y, como variables independientes, la fecha de nacimiento (*fechnac*), el nivel educativo (*educ*), el salario inicial (*salini*), la experiencia previa (*expprev*), y la clasificación étnica (*minoría*). El objetivo del análisis es encontrar un modelo de regresión que explique, con el mínimo número posible de variables independientes, la mayor cantidad posible de la varianza de la variable *salario*. Para llevar a cabo el análisis:

- Seleccionar la variable *salario* y trasladarla al cuadro **Dependiente** (figura 18.4).
- Seleccionar las variables *fechnac*, *educ*, *salini*, *expprev* y *minoría*, y trasladarlas a la lista **Independientes**.

- Pulsar el botón de menú desplegable del cuadro **Método** y seleccionar la opción **Pasos sucesivos**.
- Pulsar el botón **Estadísticos...** para acceder al subcuadro de diálogo *Regresión lineal: Estadísticos* (ver figura 18.6) y marcar la opción **Cambio en R cuadrado**.

Aceptando todas estas elecciones, el *Visor* ofrece los resultados que muestran las tablas 18.18 a la 18.22.

La tabla 18.18 ofrece un resumen del modelo final al que se ha llegado. En la columna *Modelo* se indica en número de pasos dados para construir el modelo de regresión: tres pasos. En el primer paso se ha seleccionado la variable *salario inicial*, en el segundo, *experiencia previa* y, en el tercero, *nivel educativo*. También se indica si en alguno de los pasos se ha eliminado alguna variable previamente seleccionada; en este ejemplo no se han eliminado variables. Por último, se informa sobre el método de selección utilizado (*Por pasos*) y sobre los criterios de *entrada* y *salida*: una variable es incorporada al modelo si su coeficiente de regresión parcial es significativamente distinto de cero al 5 % y, una vez seleccionada, sólo es eliminada del modelo si con la incorporación de otra u otras variables en un paso posterior su coeficiente de regresión parcial deja de ser significativamente distinto de cero al 10 %.

Tabla 18.18. Variables introducidas/eliminadas.

Modelo	Variables introducidas	Variables eliminadas	Método
1	Salario inicial	.	Por pasos (criterio: Probabilidad de F para entrar <= ,050, Probabilidad de F para salir >= ,100).
2	Experiencia previa	.	Por pasos (criterio: Probabilidad de F para entrar <= ,050, Probabilidad de F para salir >= ,100).
3	Nivel educativo	.	Por pasos (criterio: Probabilidad de F para entrar <= ,050, Probabilidad de F para salir >= ,100).

La tabla 18.19 recoge el valor de R^2 en cada paso, el cambio experimentado por R^2 en cada paso, y el estadístico F y su significación. El estadístico F permite contrastar la hipótesis de que el cambio en R^2 vale cero en la población. Al seleccionar la primera variable (*Modelo 1*), el valor de R^2 es 0,775. Lógicamente, en el primer paso, $R^2_{\text{cambio}} = R^2$. Al contrastar la hipótesis de que el valor poblacional de R^2_{cambio} es cero se obtiene un estadístico F de 1.620,83 que, con 1 y 471 grados de libertad, tiene una probabilidad asociada de 0,000. Puesto que este valor es menor que 0,05, podemos afirmar que la proporción de varianza explicada por la variable *salario inicial* (la variable seleccionada en el primer paso) es significativamente distinta de cero.

En el segundo paso (*Modelo 2*), el valor de R^2 aumenta hasta 0,794. Esto supone un cambio de 0,019 (aproximadamente un 2 por ciento). La tabla muestra el valor del estadístico F (43,180) obtenido al contrastar la hipótesis de que el valor poblacional de R^2_{cambio} es cero, y su significación (0,000). Aunque se trata de un incremento muy pequeño (un 2 por ciento), el valor del nivel crítico nos permite afirmar que la variable *experiencia previa* (la variable incorporada al modelo en el segundo paso) contribuye significativamente a explicar lo que ocurre con la variable dependiente.

En el tercer y último paso (*Modelo 3*), R^2 toma un valor de 0,802, lo cual supone un incremento de 0,008 (aproximadamente un 1 por ciento). De nuevo se trata de un incremento muy pequeño, pero al evaluar su significación se obtiene un estadístico F de 19,293 y un nivel crí-

tico de 0,000, lo cual nos está indicando que la variable *nivel educativo* (variable incorporada en el tercer paso), también contribuye de forma significativa a explicar el comportamiento de la variable dependiente. Las tres variables seleccionadas en el modelo final consiguen explicar un 80 por ciento ($R^2 = 0{,}802$) de la variabilidad observada en el *salario actual*.

Tabla 18.19. Resumen del modelo.

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Estadísticos del cambio				
					Cambio en R cuadrado	Cambio en F	gl1	gl2	Sig. del cambio en F
1	,880	,775	,774	$8,119.79	,775	1620,83	1	471	,000
2	,891	,794	,793	$7,778.94	,019	43,180	1	470	,000
3	,896	,802	,801	$7,631.84	,008	19,293	1	469	,000

La tabla resumen del ANOVA (tabla 18.20) contiene el valor del estadístico *F* obtenido al contrastar la hipótesis de que el valor poblacional de R^2 en cada paso es cero. Ahora no se evalúa el cambio que se va produciendo en el valor de R^2 de un paso a otro, sino el valor de R^2 en cada paso. Lógicamente, si R^2 es significativamente distinta de cero en el primer paso, también lo será en los pasos sucesivos.

Tabla 18.20. Resumen del ANOVA.

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	106862706669,340	1	106862706669,340	1620,826	,000
	Residual	31053506813,535	471	65931012,343		
	Total	137916213482,875	472			
2	Regresión	109475617434,356	2	54737808717,178	904,579	,000
	Residual	28440596048,519	470	60511906,486		
	Total	137916213482,875	472			
3	Regresión	110599332927,334	3	36866444309,111	632,955	,000
	Residual	27316880555,541	469	58244947,880		
	Total	137916213482,875	472			

La tabla 18.21 contiene los coeficientes de regresión parcial de las variables incluidas en el modelo de regresión; es decir, la información necesaria para construir la ecuación de regresión en cada paso (incluyendo el término constante). Las primeras columnas recogen el valor de los coeficientes de regresión parcial (*B*) y su error típico. A continuación aparecen los coeficientes de regresión parcial estandarizados (*Beta*), los cuales proporcionan una idea acerca de la importancia relativa de cada variable dentro de la ecuación. Las dos últimas columnas muestran el estadístico *t* y el nivel crítico (*Sig*) obtenidos al contrastar las hipótesis de que los coeficientes de regresión parcial valen cero en la población. Un nivel crítico por debajo de 0,05 indica que la variable contribuye significativamente a mejorar la calidad del modelo de regresión.

Utilizar el estadístico *t* para contrastar la hipótesis de que un coeficiente de regresión parcial vale cero es exactamente lo mismo que utilizar el estadístico *F* para contrastar la hipótesis de que el valor poblacional del cambio observado en R^2 vale cero. De hecho, elevando al cua-

drado los valores del estadístico *t* de la tabla 18.21 obtenemos los valores del estadístico *F* de la tabla 18.19. De las dos formas se está intentando evaluar la contribución individual de una variable a la proporción de varianza explicada por el conjunto de variables independientes.

Tabla 18.21. Coeficientes de regresión parcial.

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados		
		B	Error típ.	Beta	t	Sig.
1	(Constante)	1929,517	889,168		2,170	,031
	Salario inicial	1,910	,047	,880	40,259	,000
2	(Constante)	3856,955	900,928		4,281	,000
	Salario inicial	1,924	,045	,887	42,279	,000
	Experiencia previa	-22,500	3,424	-,138	-6,571	,000
3	(Constante)	-3708,904	1936,045		-1,916	,056
	Salario inicial	1,748	,060	,806	29,192	,000
	Experiencia previa	-16,752	3,605	-,103	-4,647	,000
	Nivel educativo	741,307	168,772	,125	4,392	,000

Por último, la tabla 18.22 muestra los coeficientes de regresión parcial de las variables no seleccionadas para formar parte de la ecuación de regresión en cada paso. La información que contiene esta tabla permite conocer en detalle por qué unas variables han sido seleccionadas y otras no. En el primer paso se ha seleccionado la variable *salario inicial* porque es la que más alto correlaciona, en valor absoluto, con la variable dependiente (esta información la tenemos en la tabla 18.10). En ese primer paso, todavía están fuera del modelo el resto de variables independientes. La columna *Beta dentro* contiene el valor que tomaría el coeficiente de regresión estandarizado de una variable en el caso de que fuera seleccionada en el siguiente paso. Las columnas *t* y *Sig.* nos informan sobre si ese valor que adoptaría el coeficiente de regresión de una variable en el caso de ser incorporada al modelo sería o no significativamente distinto de cero.

Tabla 18.22. Variables excluidas.

Modelo		Beta dentro	t	Sig.	Correlación parcial	Estadísticos de colinealidad
						Tolerancia
1	Nivel educativo	,173	6,385	,000	,283	,599
	Experiencia previa	-,138	-6,571	,000	-,290	,998
	Clasificación étnica	-,040	-1,809	,071	-,083	,975
	Fecha de nacimiento	,136	6,471	,000	,286	1,000
2	Nivel educativo	,125	4,392	,000	,199	,520
	Clasificación étnica	-,019	-,885	,377	-,041	,952
	Fecha de nacimiento	,071	2,020	,044	,093	,354
3	Clasificación étnica	-,020	-,965	,335	-,045	,952
	Fecha de nacimiento	,054	1,556	,120	,072	,350

Vemos que, en el primer paso, hay tres variables todavía no seleccionadas (*nivel educativo*, *experiencia previa* y *fecha de nacimiento*) cuyos coeficientes de regresión poseen niveles críticos por debajo de 0,05 (criterio de *entrada*). Entre ellas, la que posee un coeficiente de

correlación parcial mayor en valor absoluto (*experiencia previa* = −0,290) y, además, un nivel de tolerancia por encima de 0,001 (*tolerancia mínima* establecida por defecto), es la variable que ha sido seleccionada en el segundo paso. En el segundo paso todavía quedan fuera de la ecuación dos variables cuyos coeficientes de regresión serían significativos en caso de ser seleccionadas: *nivel educativo* y *fecha de nacimiento*. En el tercer paso ha sido seleccionada la variable *nivel educativo* porque, teniendo un nivel de tolerancia por encima de 0,001, es la que posee el coeficiente de correlación parcial más alto. Después del tercer paso ya sólo quedan dos variables fuera de la ecuación: *fecha de nacimiento* y *clasificación étnica*. Puesto que ninguna de las dos supera el criterio de *entrada* (*Sig.* < 0,05), el proceso se detiene y ambas variables quedan fuera del modelo.

Qué variables debe incluir la ecuación de regresión

El método de selección por pasos nos ha llevado a construir una ecuación de regresión con tres variables. Las tres variables seleccionadas poseen coeficientes de regresión parcial significativos. Sin embargo, la primera variable explica el 78 % de la varianza de la variable dependiente, la segunda el 2 %, y la tercera el 2 %. Si en lugar del método *pasos sucesivos* hubiéramos utilizado el método *introducir*, habríamos obtenido los resultados que muestran las tablas 18.23 y 18.24.

Tabla 18.23. Resumen del modelo.

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,902	,813	,811	$7,437.80

Tabla 18.24. Coeficientes de regresión parcial.

Modelo: 1

	Coeficientes no estandarizados		Coeficientes estandarizados		
	B	Error típ.	Beta	t	Sig.
(Constante)	-38768,134	18686,944		-2,075	,039
Código de empleado	-12,426	2,526	-,099	-4,920	,000
Fecha de nacimiento	3,275E-06	,000	,071	2,091	,037
Nivel educativo	630,255	166,453	,106	3,786	,000
Salario inicial	1,760	,059	,811	29,826	,000
Experiencia previa	-7,703	5,596	-,047	-1,376	,169
Clasificación étnica	-994,637	846,844	-,024	-1,175	,241

Por un lado, la ganancia que obtenemos en R^2 utilizando las 5 variables en lugar de las tres seleccionadas con el método por pasos es extremadamente pequeña (0,813−0,802 = 0,011). No parece que tenga mucho sentido añadir dos variables para obtener una mejora de once milésimas en la proporción de varianza explicada. Es cierto que R^2 nunca disminuye cuando se van incorporando nuevas variables al modelo de regresión, sino que aumenta o se queda como está. Sin embargo, esto no significa, necesariamente, que la ecuación con más variables se ajuste

mejor a los datos poblacionales. Generalmente, conforme va aumentando la calidad del modelo, va disminuyendo el error típico de los residuos (*Error típ. de la estimación*). Pero el incremento que se va produciendo en R^2 al ir añadiendo variables no se corresponde necesariamente con una disminución del error típico de los residuos. Con cada variable nueva que se incorpora al modelo, la suma de cuadrados de la regresión gana un grado de libertad y la suma de cuadrados de los residuos lo pierde. Por tanto, el error típico de los residuos puede aumentar cuando el descenso de la variación residual es demasiado pequeño para compensar la pérdida de un grado de libertad en la suma de cuadrados de los residuos. Estas consideraciones sugieren la conveniencia de utilizar modelos parsimoniosos, es decir, modelos con un número reducido de variables independientes.

Por otro lado, las variables que tienen pesos significativos en la ecuación de regresión obtenida con el método *pasos sucesivos* no son las mismas que las que tienen pesos significativos en la ecuación obtenida con el método *introducir*. Esta diferencia entre métodos de selección de variables debe ser tenida muy en cuenta. ¿Cuáles son las variables *buenas*? Atendiendo a criterios puramente estadísticos, la ecuación de regresión con las tres variables seleccionadas por el método *pasos sucesivos*, es la mejor de las posibles con el mínimo número de variables. Pero en la práctica, la decisión sobre cuántas variables debe incluir la ecuación de regresión puede tomarse teniendo en cuenta, además de los criterios estadísticos, otro tipo de consideraciones. Si, por ejemplo, resulta muy costoso (tiempo, dinero, etc.) obtener las unidades de análisis, un modelo con sólo una variable podría resultar lo bastante apropiado. Si las consecuencias de los residuos de los pronósticos fueran muy graves, deberían incluirse en el modelo las tres variables del método *pasos sucesivos* o las cuatro con pesos significativos del método *introducir*. Así pues, para decidir con qué modelo de regresión nos quedamos, casi siempre es conveniente tomar en consideración criterios adicionales a los puramente estadísticos.

Por supuesto, los contrastes estadísticos sirven de apoyo para tomar decisiones. Pero, dado que la potencia de un contraste se incrementa conforme lo hace el tamaño de la muestra, debemos ser cautelosos con las conclusiones a las que llegamos. Esto significa que, con muestras grandes, efectos muy pequeños desde el punto de vista de su importancia teórica o práctica pueden resultar estadísticamente significativos. Por el contrario, con muestras pequeñas, para que un efecto resulte significativo, debe tratarse de un efecto importante (con muestras pequeñas, existe mayor grado de coincidencia entre la significación estadística y la importancia práctica). Por esta razón, en la determinación de la ecuación de regresión final, debe tenerse en cuenta, cuando se trabaja con muestras grandes, la conveniencia de considerar elementos de decisión adicionales a la significación estadística.

Puesto que la utilización de los métodos de selección por pasos está bastante generalizada, conviene también alertar sobre el peligro de alcanzar un resultado falsamente positivo (un error de tipo I). Es decir, si examinamos un número de variables lo bastante grande, tarde o temprano una o más pueden resultar significativas sólo por azar. Este riesgo es tanto mayor cuanto más variables se incluyen en el análisis. Para evitar este problema, si la muestra es lo bastante grande, puede dividirse en dos, aplicar el análisis a una mitad y verificar en la otra mitad si se confirma el resultado obtenido. Si la muestra es pequeña, esta solución es inviable y, por tanto, el riesgo de cometer un error de tipo I permanece.

Cómo efectuar pronósticos

Si el objetivo del análisis de regresión es el de evaluar la capacidad de un conjunto de variables independientes para dar cuenta del comportamiento de una variable dependiente, no es necesario añadir nada más a lo ya estudiado. Sin embargo, si el objetivo principal del análisis es el de poder efectuar pronósticos en casos nuevos, todavía nos falta saber algunas cosas.

Recordemos que ya hemos utilizado los coeficientes de regresión parcial (*B*) para construir la ecuación de regresión (ver tabla 18.6). Reproducimos a continuación la ecuación de regresión obtenida:

$$Pronóstico\ (salario) = -3.661{,}517 + 1{,}749\ salini - 16{,}730\ expprev + 735{,}956\ educ$$

Conociendo los pesos de la ecuación de regresión, podríamos utilizar la opción **Calcular** del menú **Transformar** para obtener los pronósticos que la ecuación asigna a cada caso. Pero esto no es necesario. El subcuadro de diálogo *Regresión lineal: Guardar nuevas variables* (ver figura 18.12) contiene varias opciones relacionadas con los pronósticos:

Valores pronosticados. Las opciones de este recuadro generan, en el *Editor de datos*, cuatro nuevas variables. Estas nuevas variables reciben automáticamente un nombre seguido de un número de serie: *nombre_#*. Por ejemplo, la primera vez que se solicitan durante una sesión los *pronósticos tipificados*, la nueva variable con los pronósticos tipificados recibe el nombre "zpr_1". Si se vuelven a solicitar los pronósticos tipificados durante la misma sesión, la nueva variable recibe el nombre "zpr_2". Etc.

- ☐ **No tipificados**: pronósticos que se derivan de la ecuación de regresión en puntuaciones directas. Nombre: *pre_#*.
- ☐ **Tipificados**: pronósticos convertidos en puntuaciones típicas (restando a cada pronóstico la media de los pronósticos y dividiendo la diferencia por la desviación típica de los pronósticos). Nombre: *zpr_#*.
- ☐ **Corregidos**: pronóstico que corresponde a cada caso cuando la ecuación de regresión se obtiene sin incluir ese caso. Nombre: *adj_#*.
- ☐ **E.T. del pronóstico promedio**: error típico de los pronósticos correspondientes a los casos que tienen el mismo valor en las variables independientes. Nombre: *sep_#*.

  Aclaremos esto. Al efectuar pronósticos es posible optar entre: 1) efectuar un pronóstico individual Y_i' para cada caso concreto X_i, o (2) pronosticar para cada caso la media de los pronósticos (Y_0') correspondientes a todos los casos en con el mismo valor X_0 en la(s) variable(s) independiente(s); a esta media es a la que llamamos *pronóstico promedio*. En ambos casos se obtiene el mismo pronóstico ($Y_i' = Y_0'$), pero cada tipo de pronóstico (ambos son variables aleatorias) tiene un error típico distinto. La figura 18.14 puede ayudarnos a comprender la diferencia entre estos dos errores típicos.

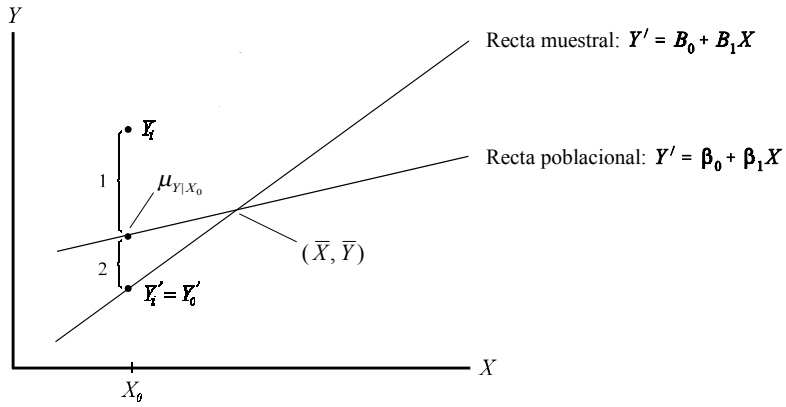
  Al efectuar un pronóstico individual para un determinado valor de X_i, el error de estimación o variación residual ($Y_i - Y_i'$) puede contener dos fuentes de error (identificadas en la figura 18.14 con los números 1 y 2):

- La diferencia entre el valor observado en la variable dependiente (Y_i) y la media poblacional correspondiente a X_0 ($\mu_{Y|X_0}$).
- La diferencia entre el pronóstico para ese caso (Y_i' o Y_0') y la media poblacional correspondiente a X_0 ($\mu_{Y|X_0}$).

En un pronóstico *individual* entran juego ambas fuentes de error. Pero en un pronóstico *promedio* sólo entra en juego la segunda fuente de error. Por tanto, para un valor dado de X_0, el error típico del pronóstico *promedio* siempre será menor o igual que el error típico del pronóstico *individual*. En consecuencia, al construir intervalos de confianza para los pronósticos, la amplitud del intervalo cambiará dependiendo del error típico que se tome como referencia.

Además, observando la figura 18.14, puede intuirse fácilmente que los errores típicos del pronóstico promedio (que ya sabemos que están basados en las distancias entre Y_0' y $\mu_{Y|X_0}$) serán tanto menores cuanto más se parezcan X_0 y $\bar{X}$, pues cuanto más se parezcan, más cerca estará la recta muestral de la poblacional y, consecuentemente, más cerca estarán Y_0' y $\mu_{Y|X_0}$.

Figura 18.14. Tipos de error en los pronósticos de la regresión.



Intervalos de pronóstico (seguimos en el subcuadro de diálogo *Regresión lineal: Guardar nuevas variables*; ver figura 18.12). Las opciones de este recuadro permiten obtener los intervalos de confianza para los pronósticos:

- **Media**. Intervalo de confianza basado en los errores típicos de los pronósticos promedio.
- **Individuos**. Intervalo de confianza basado los errores típicos de los pronósticos individuales.

La opción **Intervalo de confianza *k* %** permite establecer el nivel de confianza con el que se construyen los intervalos de confianza.

Lógicamente, estos dos intervalos son distintos. Para un valor dado de *X*, el primer intervalo (media) es más estrecho que el segundo (individuos). Recuérdese lo dicho en este mismo apartado sobre los errores típicos de los pronósticos.

Cada una de estas dos opciones (media e individuos) genera en el *Editor de datos* dos nuevas variables con el límite inferior y superior del intervalo. Estas nuevas variables reciben los siguientes nombres:

- *lmci_#*: límite inferior del intervalo de confianza para el pronóstico medio.
- *umci_#*: límite superior del intervalo de confianza para el pronóstico medio.
- *lici_#*: límite inferior del intervalo de confianza para el pronóstico individual.
- *uici_#*: límite superior del intervalo de confianza para el pronóstico individual.

Validez del modelo de regresión

El modelo de regresión puede ser validado utilizando casos nuevos. Para ello, basta con obtener los pronósticos para esos casos nuevos y, a continuación, calcular el coeficiente de correlación entre los valores observados en la variable dependiente y los valores pronosticados para esos casos nuevos. En teoría, el coeficiente de correlación así obtenido debería ser igual al coeficiente de correlación múltiple del análisis de regresión (*R*). En la práctica, si el modelo es lo bastante bueno, encontraremos pequeñas diferencias entre esos coeficientes, atribuibles únicamente al azar muestral. Es muy importante que los nuevos casos representen a las mismas poblaciones que los casos originalmente utilizados para obtener la ecuación de regresión.

En ocasiones, es posible que no tengamos acceso a nuevos datos o que sea muy difícil obtenerlos. En esos casos, todavía es posible validar el modelo de regresión si la muestra es lo bastante grande. Basta con utilizar la mitad de los casos de la muestra (aleatoriamente seleccionados) para obtener la ecuación de regresión y la otra mitad de la muestra para efectuar los pronósticos. Un modelo fiable debería llevarnos a obtener una correlación similar entre los valores observados y pronosticados de ambas mitades.