KIWITEC.
HIGH QUALITY TECH COURSES

Curso:

Minería de Datos I con SPSS STATISTICS

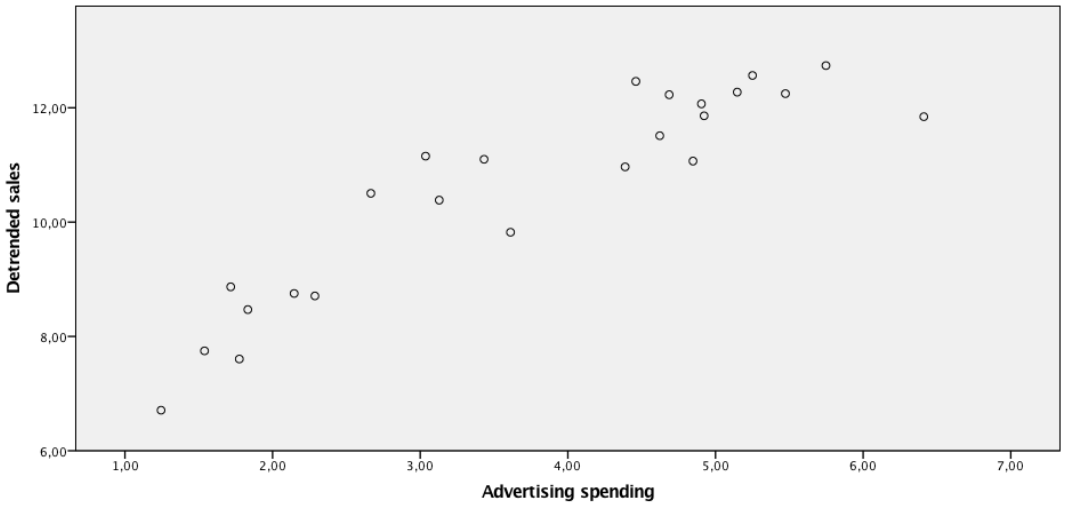
Módulo I:

ANÁLISIS DE REGRESIÓN LINEAL

I. Introducción

El análisis de regresión lineal es una técnica estadística utilizada para estudiar la relación entre diferentes variables.

Una buena forma de comenzar a estudiar la relación existente entre unas variables la constituyen los diagramas de dispersión. Éstos ofrecen una idea bastante aproximada del tipo de relación que existe entre dos o más variables, así como de la intensidad de la misma.



No obstante, si bien un diagrama de dispersión permite formarse una primera impresión sobre el tipo de relación existente entre variables, no puede utilizarse como una forma de cuantificarla.

Los análisis de regresión lineal pueden utilizarse cuando se considera que una variable presenta una relación lineal con otras variables, es decir, cuando puede expresarse como una combinación lineal de dichas variables.

Una regresión lineal simple pretende determinar el grado de relación lineal entre dos variables. Cuando se trabaja con más de dos variables se denomina regresión múltiple. En ambos casos, el análisis de regresión lineal pretende explorar y cuantificar la relación entre una variable llamada dependiente o criterio (*Y*) y una o más variables llamadas independientes o predictoras ($X_1, X_2 \ldots, X_k$).

Además, un análisis de regresión lineal permite desarrollar una ecuación lineal que podrá utilizarse para predecir nuevos valores de la variable criterio a partir de los valores conocidos de las variables predictoras. Esta ecuación recibe el nombre de ecuación o recta de regresión

En el caso de trabajar con una variable criterio y una variable predictora, la menciona ecuación lineal representará una recta, del tipo:

$$Y = B_0 + B_1 X$$

En el caso de trabajar con variables predictoras, la ecuación representará un hiperplano con tantas dimensiones como variables predictoras se estén utilizando.

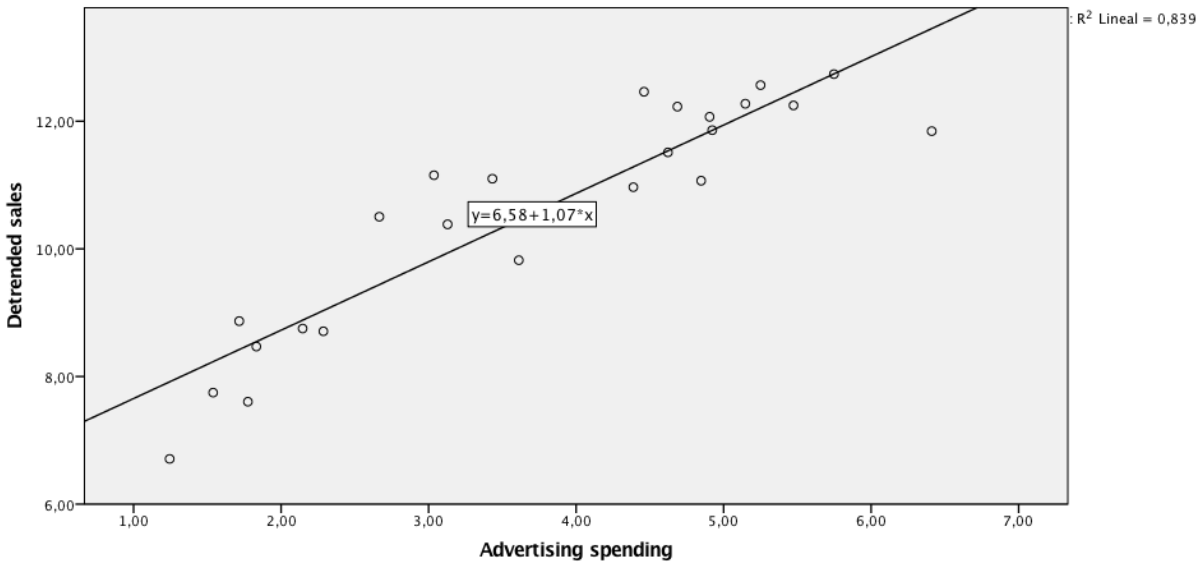
$$Y = B_0 + B_1 X_i$$

Por otro lado, existenten numerosos procedimientos de diagnóstico asociados al análisis de regresión que permiten estudiar la idoneidad del análisis, proporcionando pistas sobre cómo perfeccionarlo.

1. La mejor recta de regresión

En una situación ideal (e irreal) en la que todos los puntos de un diagrama de dispersión se encontrasen 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 con mejor ajuste a la nube de puntos. Pero en una nube de puntos más realista 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).



2. 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.

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.

¿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*** $\boldsymbol{R^2}$: 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.

3. Supuestos del modelo de regresión lineal

Como para todo análisis estadístico, deben cumplirse una serie de supuestos que han de examinarse para garantizar la validez del modelo de regresión:

1.- Linealidad. La ecuación de regresión lineal debe representar un hiperplano. La variable dependiente o criterio se plantea como la suma de un conjunto de items: el origen de la recta, una combinación lineal de variables independientes o predictoras y los residuos. El incumplimiento del supuesto de linealidad es un error de especificación. Algunos casos en los que se puede romper el supuesto de linealidad son la omisión de variables independientes importantes, la inclusión de variables independientes irrelevantes, la no linealidad entre las variables independientes y la dependiente, etc.

2.- Independencia. Las diferencias entre los valores observados y los estimados por el modelo, que se denominan residuos, han de ser independientes entre sí. Los residuos deben ser una variable aleatoria.

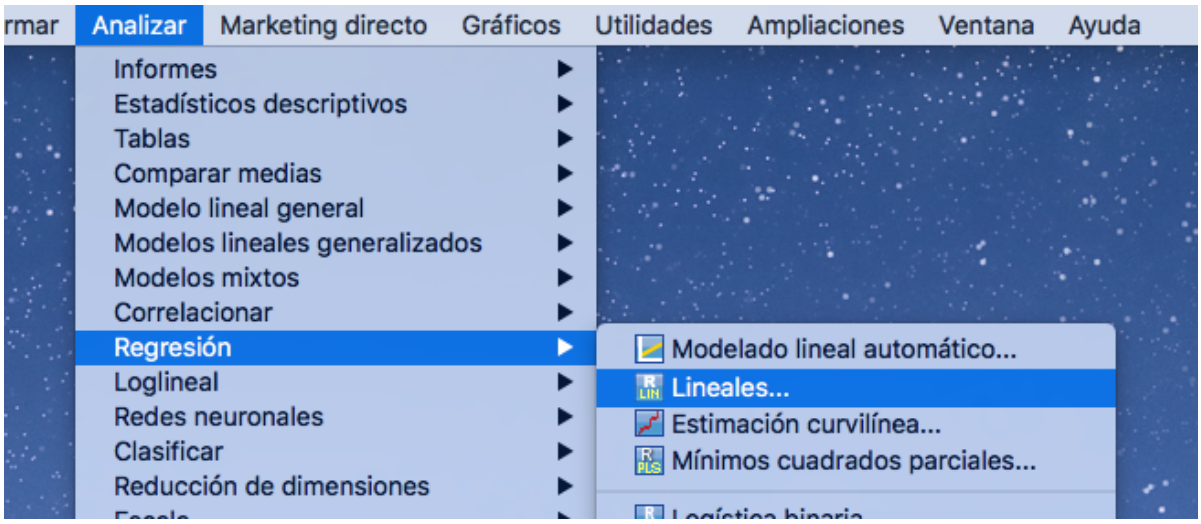
3.- Homocedasticidad. Para cada valor de la variable independiente (o combinación de va- lores de las variables independientes), la varianza de los residuos debe mantenerse constante.

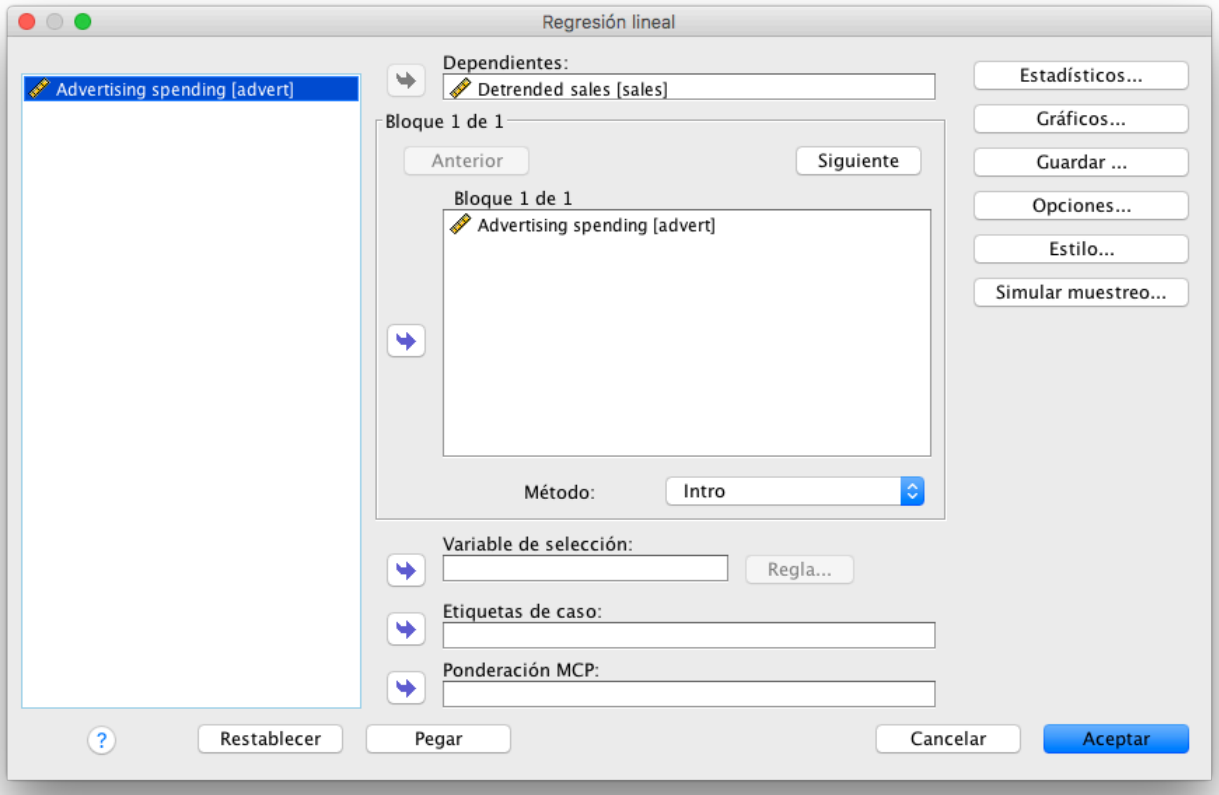
4.- Normalidad. Para cada valor de la variable independiente (o combinación de valores de las variables independientes), los residuos se distribuyen normalmente con media cero.

5.- No-colinealidad. No debe existir relación lineal exacta entre ninguna de las variables independientes. El incumplimiento de este supuesto da origen a colinealidad o multicolinealidad. Este supuesto no tiene sentido en regresión simple, pues es imprescindible la presencia de más de una variable independiente. Existen diferentes formas de determinar la presencia de colinealidad.

II. Análisis de regresión lineal simple

Para realizar un análisis de regresión simple con SPSS debe seleccionarse la opción de menú: **Analizar > Regresión > Lineal …** Al hacerlo, aparece la siguiente pantalla. En ella debe seleccionarse una variable dependiente o criterio, y una variable independiente o predictora. Pulsando sobre el botón aceptar se procede al análisis de regresión, mostrándose los resultados del mismo sobre el visor de resultados.





1. Resumen del modelo

La primera información que se muestra es el coeficiente de correlación múltiple (*R*) y su cuadrado, el coeficiente de determinación. Al tratarse de una regresión lineal simple, el coeficiente de correlación múltiple se corresponde con el valor absoluto del coeficiente de correlación de Pearson entre esas dos variables.

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

R^2, que se utiliza como medida de la bondad de ajuste de la regresión y por tanto como una medida del tamaño del efecto detectado, expresa la proporción de varianza de la variable dependiente que está explicada por la variable independiente.

Además se muestra *R cuadrado corregida,* que es una corrección a la baja de R^2 basada en el número de casos y de variables independientes

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

Donde *p* se refiere al número de variables independientes.

Además se muestra la desviación típica de los residuos, o error típico de estimación, que será de utilidad a la hora de acotar el error máximo de estimación cometido al pronosticar un valor de la variable dependiente. Este error típico es la raíz cuadrada de la media cuadrática residual de la tabla ANOVA de la regresión. Representa una medida de la parte de variabilidad de la variable dependiente que no es explicada por la recta de regresión. Cuanto mejor es el ajuste, más pequeño es el error típico.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	,916[a]	,839	,832	,73875

a. Predictores: (Constante), Advertising spending

2. Tabla ANOVA

La tabla resumen ANOVA ofrece información que permite aceptar o rechazar la hipótesis nula que determina que las variable no guardan relación lineal estadísticamente significativa a nivel poblacional, o dicho de otra forma, que el valor poblacional de *R* es cero, lo que 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 la probabilidad de que *R* tome el valor estimado en la muestra siendo el valor de R a nivel poblacional 0. Si el valor del nivel crítico es inferior al nivel de significación del estudio, puede concluirse que *R* es mayor que cero y que ambas variables están linealmente relacionadas.

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	62,514	1	62,514	114,548	,000[b]
	Residuo	12,006	22	,546		
	Total	74,520	23			

a. Variable dependiente: Detrended sales

b. Predictores: (Constante), Advertising spending

3. Coeficientes de regresión

Posteriormente se muestra una tabla con 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.

El coeficiente correspondiente a la Constante es el origen de la recta de regresión B_0:

$$\mathrm{B}_0 = \overline{\mathrm{Y}} - \mathrm{B}_1\overline{\mathrm{X}}$$

El coeficiente correspondiente a la Variable es la pendiente de la recta de regresión B_1:

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

Los coeficientes de regresión parcial estandarizados son los coeficientes que definen la ecuación de regresión en puntuaciones típicas:

$$\beta_1 = B_1(S_x/S_y)$$

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, los coeficientes de regresión estandarizados permiten valorar la importancia relativa de cada variable independiente dentro de la ecuación.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados		
		B	Error estándar	Beta	t	Sig.
1	(Constante)	6,584	,402		16,391	,000
	Advertising spending	1,071	,100	,916	10,703	,000

a. Variable dependiente: Detrended sales

Junto a los coeficientes, aparecen los estadísticos *t* asociados y sus niveles críticos (*Sig.*). Éstos permiten contrastar las hipótesis nulas de que los coeficientes 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:

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

siendo:

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

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.

En un análisis de regresión simple que sólo trabaja 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*).

Contrastar si el origen poblacional de la recta de regresión (β_0) es significativamente distinto de cero suele carecer de utilidad, pues no contiene información sobre la relación entre X_i e Y_i.

III. 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.

Con una variable dependiente y dos independientes, son necesarios tres ejes para poder representar el correspondiente diagrama de dispersión. Y si en lugar de dos variables independientes se utilizasen tres, sería necesario un espacio de cuatro dimensiones para poder construir el diagrama de dispersión.

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:

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

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.

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:

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

1. Coeficientes de regresión

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

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

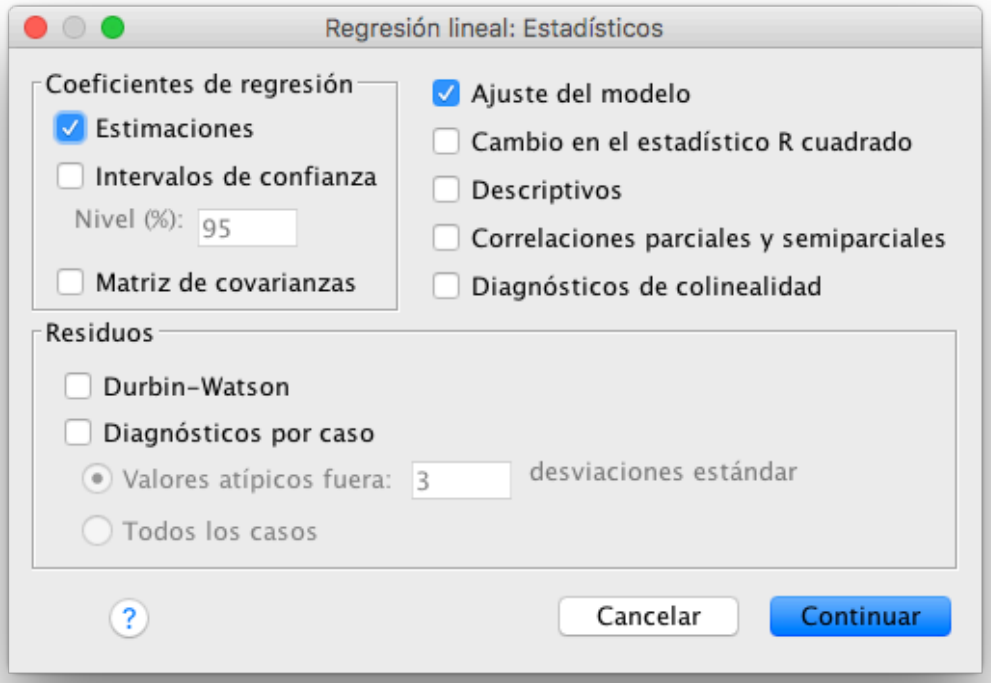
El signo del coeficiente de regresión parcial de una variable puede no ser el mismo que el del coeficiente de correlación simple entre esa variable y la dependiente. Esto es debido a los ajustes que se llevan a cabo para poder obtener la mejor ecuación posible. Aunque existen diferentes explicaciones 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.

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.

1. Otros estadísticos de utilidad

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 hay que pulsar el botón **Estadísticos...** del cuadro de diálogo Regresión lineal para acceder a la subpantalla de diálogo Regresión lineal: Estadísticos.



• **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 éstos.

Los intervalos nos informan sobre los límites entre los que podemos esperar que se encuentre el valor poblacional de cada coeficiente de regresión.

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).

• **Matriz de covarianza.** Muestra una matriz con las covarianzas y correlaciones existentes entre los coeficientes de regresión parcia.

• **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. 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.

El resto de opciones del subcuadro de diálogo Regresión lineal: Estadísticos, 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.

IV. Comprobación de los supuestos

1. Análisis de los residuos

Se denominan 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 la subpantalla del diálogo Regresión lineal: Guardar nuevas variables.

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, 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 ayudar 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* 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.

Por defecto, el SPSS lista los residuos que se alejan de cero más de 3 desviaciones típicas, pero puede cambiarse este valor por otro elegido. Para obtener un listado de los residuos que se alejan de cero más de, por ejemplo, tres desviaciones típicas bastaría con marcar la opción **Diagnósticos por caso** y seleccionar **Valores atípicos a más de [3] desviaciones típicas.**

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 (que es un supuesto del 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 cuidadosamente 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. Para ello es de utilidad guardar los residuos correspondientes a cada caso como una variable más del archivo de datos y, consecuentemente, aplicar sobre ella los procedimientos SPSS que se desee.

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. Especialmente importante es señalar que la media de los residuos vale cero.

i. 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:

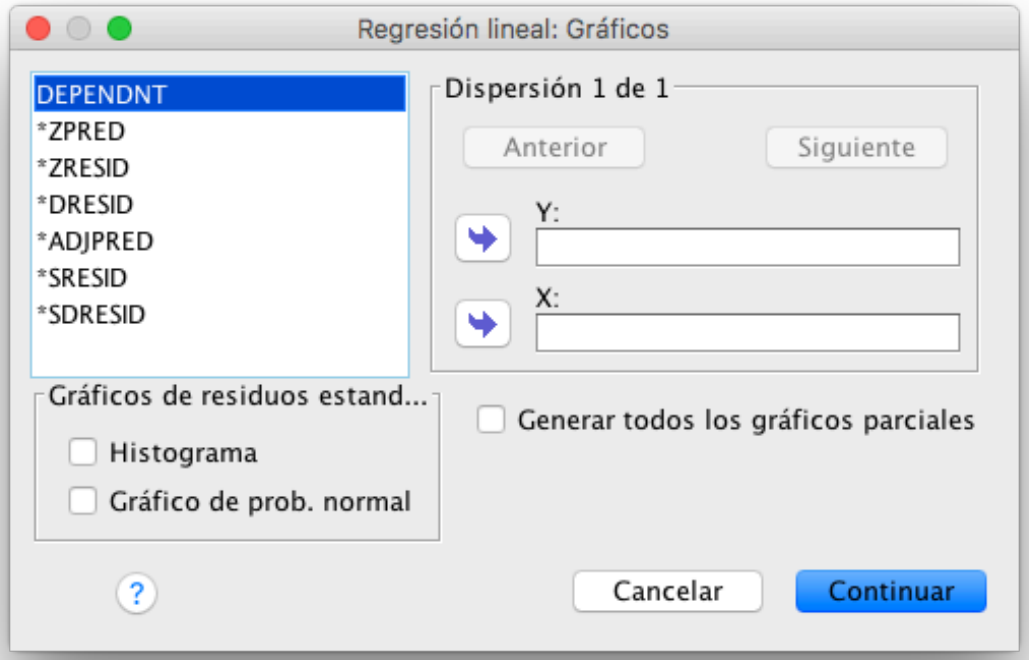
$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

(e_i se refiere a los residuos: $e_i = Y_i - \bar{Y}_i$). 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 hay que seleccionar la opción de Durbin-Watson del cuadro de diálogo Regresión lineal: Estadísticos. El valor del estadístico aparecerá en este caso en la tabla de resumen del modelo.

ii. 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. Estos gráficos pueden generarse desde el botón **Gráficos...** del cuadro de diálogo Regresión lineal.



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.*

- **DEPENDENT:** variable dependiente de la ecuación de regresión.
- **ZPRED** (pronósticos tipificados): pronósticos divididos por su desviación típica. Son pronósticos transformados en puntuaciones *z* (con media 0 y desviación típica 1).
- **ZRESID** (residuos tipificados): 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, 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 su diagrama de dispersión. En él puede observarse que, si los residuos y los pronósticos son independientes y homocedásticos.

iii. Normalidad

El recuadro **Gráficos de los residuos tipificados** permite generar un histograma, que informa sobre el grado en el que los residuos tipificados se aproximan a una distribución normal. Contiene los residuos tipificados con una curva normal superpuesta. 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.

2. 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.

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 pueden encontrarse 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 suele utilzarse 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, 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 ayudar 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 debe seleccionarse la opción **Diagnósticos de colinealidad** del subcuadro de diálogo Regresión lineal: Estadísticos.

De esta forma se añaden en la tabla de coeficientes de regresión parcial, 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.

Además se muestra la solución resultante de aplicar un análisis de componentes principales a la matriz estandarizada no centrada de productos cruzados de las variables independientes, que contiene:

- 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.

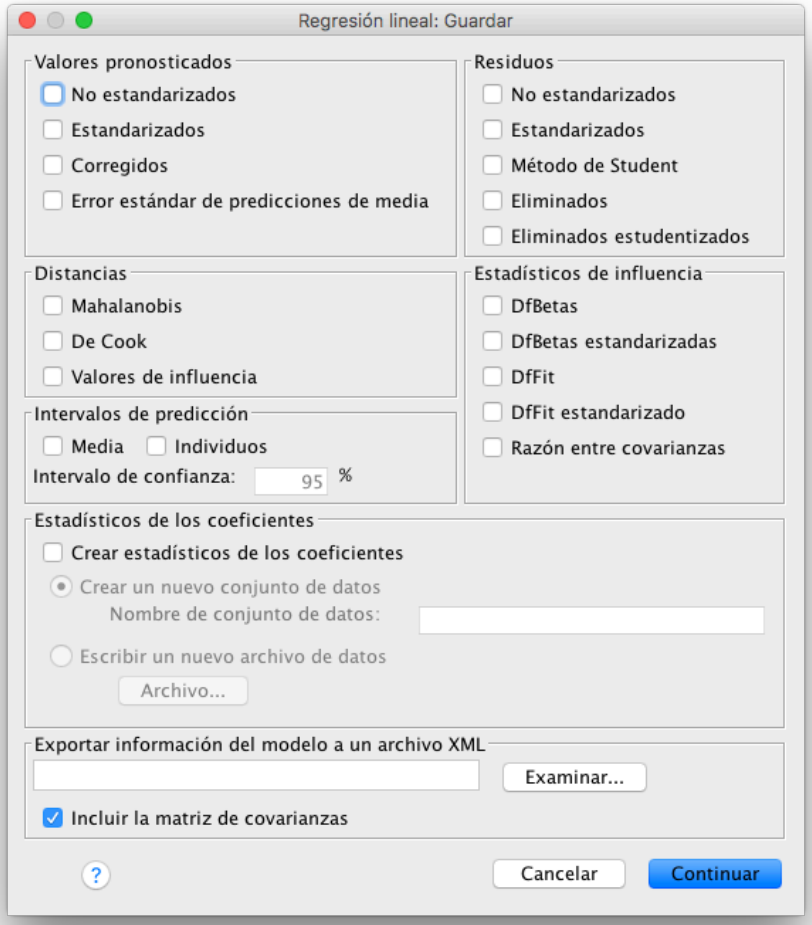
- Las ***proporciones de varianza*** recogen la proporción de varianza de cada coeficiente de regresión parcial que está explicada por cada dimensión o factor. En condiciones de no-colinealidad, cada dimensión suele explicar gran cantidad de varianza de un sólo coeficiente (excepto en lo que se refiere al coeficiente B_0 o *constante*, que siempre aparece asociado a uno de los otros coeficientes). 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, por ejemplo: 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).

3. 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 hay que pulsar el botón **Guardar**... del cuadro de diálogo Regresión lineal y marcar todas las opciones de los recuadros **Distancias** y **Estadísticos de influencia** (todas estas opciones crean variables nuevas en el archivo de datos).



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.

- **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 consideran 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. 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, SPSS ofrece una tabla resumen que incluye, para todos los estadísticos del recuadro **Distancias**, 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.

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, podrían ser eliminados del análisis, simplemente porque entorpecen el ajuste, o porque su presencia está haciendo obtener medidas de ajuste infladas. También podrían eliminarse los casos muy atípicos simplemente argumentando que el 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.

A la hora de decidir sobre la conveniencia de eliminar o no un caso, puede ayudar el pensar sobre los motivos que justifiquen su eliminación.

V. 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.

1. 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.

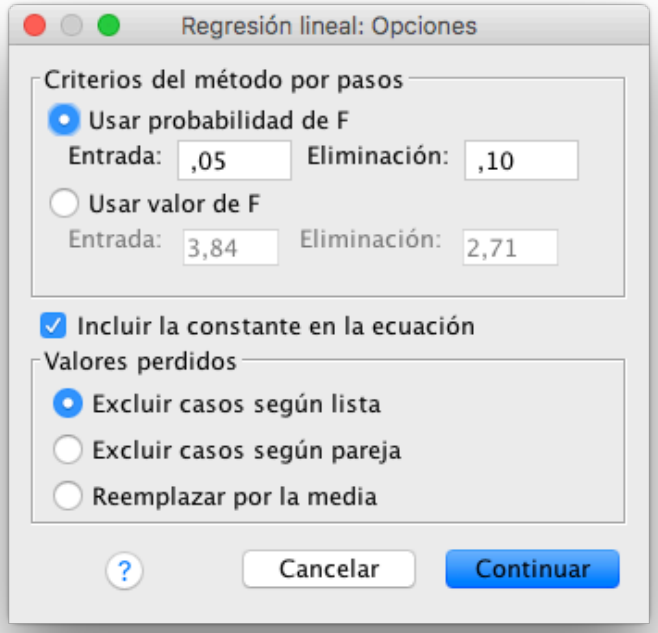
Los métodos por pasos que incluye el SPSS basan la selección de variables en dos criterios estadísticos:

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, 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 la probabilidad de *entrada*. Y queda fuera del modelo de regresión si ese nivel crítico es mayor que 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 el valor de entrada. Y queda fuera del modelo si el valor del estadístico *F* es menor que el valor de salida.

Para modificar estos criterios de selección hay que pulsar el botón **Opciones** del cuadro de diálogo *Regresión lineal* para acceder al 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 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.

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. Este cambio se define como $R^2_{cambio} = R^2 - R^2_i$, 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 R^2_{cambio} de y su significación (el grado en que el cambio observado en R^2 difiere de cero). SPSS calsulará este valor al marcar la opción **Cambio en R cuadrado** del cuadro de diálogo *Regresión lineal*: Estadísticos.

2. Métodos definidos por SPSS

El procedimiento *Regresión lineal* del SPSS incluye varios 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*. 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. 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.

En los métodos explicados anteriormente, 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, y se denominan 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 siguientes 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 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*.

VI. 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, el objetivo principal del análisis puede ser el de realizar pronósticos para nuevos casos.

Con los coeficientes de regresión parcial (*B*) puede construirse la ecuación de regresión del modelo analizado.

Conociendo los pesos de la ecuación de regresión, puede utilizarse la opción **Calcular** del menú **Transformar** para obtener los pronósticos que la ecuación asigna a cada caso. Pero esto no es siempre necesario. El subcuadro de diálogo *Regresión lineal: Guardar nuevas variables* 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 pro- nóstico 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_#.* Al efectuar pronósticos es posible optar entre:
    - efectuar un pronóstico individual $Y_i^{'}$ para cada caso concreto X_i
    - 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.

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.

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 $\mu_{Y|X_0}$

Intervalos de pronóstico: 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).

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.

VII. 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.