Unidad didáctica 1:
Procedimientos y preparación de los datos

Unidad didáctica 2:
Análisis descriptivo de datos

Unidad didáctica 3:
Muestreo y estimación

UNIDAD DIDÁCTICA 4:
Análisis inferencial de datos

UNIDAD DIDÁCTICA 5:
Hacia un estudio del modelo

Aquí podrá encontrar información adicional
y actualizada de esta publicación

1. El modelo de regresión. El comando REGRESION

1.1. El modelo de regresión

1.1.1. Introducción

La regresión lineal estudia la relación existente entre una o más variables, denominadas independientes y otra, denominada dependiente, con propósitos tanto descriptivos como predictivos.

Podemos plantear una relación, en principio lineal, entre una variable Y dependiente que trata de ser explicada por k variables independientes y un término de perturbación aleatoria e. De esta forma para cada observación se tendrá:

$$y_i=b_0+b_1x_{i1}+.....+b_kx_{ik}+e_i \quad i=1,....,n \qquad [1]$$

donde:

$b_0,....., b_k$ son parámetros desconocidos a estimar, y
e_i i=1,....,n son variables error, independientes y con distribución $N(0,\sigma^2)$

De forma matricial **Y=XB+E** donde **X** es una matriz con la primera columna unitaria.

El análisis de regresión es una de las técnicas más utilizada en investigación, sus posibilidades son innumerables como lo demuestran las continuas referencias en publicaciones. Sus aplicaciones se pueden agrupar en dos grandes apartados: ***predicción*** y ***explicación***. Estos dos usos no son mutuamente excluyentes y existirán investigaciones donde se utilicen con ambas finalidades.

En la ***predicción***, la combinación lineal de las variables independientes se dirige a maximizar la estimación de la variable dependiente, y es un predictor del poder explicativo de la variable dependiente por las variables independientes. Se deben conseguir niveles adecuados de explicación de la variable dependiente para justificar el modelo de regresión. También, la faceta predictiva del análisis de regresión, sirve para evaluar el conjunto de variables independientes como predictoras de la variable dependiente.

La vertiente ***explicativa*** del análisis de regresión, se utiliza para dar una visión de la importancia relativa de cada variable independiente valorando su magnitud y signo. Además, se puede trabajar para determinar el tipo de relación existente (lineal, cuadrática, logarítmica, exponencial, potencial, etc) con la variable dependiente.

Regresión lineal simple

En el caso particular de una única variable independiente **X**, se habla de regresión lineal simple. La correspondiente función de regresión será del tipo:

$$f(X,b_0,b_1)=b_0+b_1X$$

$$y_i=b_0+b_1x_i+e_i \quad i=1,....,n \qquad [2]$$

Es de destacar la semejanza entre el modelo [1] y el análisis de la varianza modelo factorial con un solo factor, siendo la única diferencia la relativa a que mientras en [2] la variable **X** puede tomar cualquier valor, en el modelo de análisis de la varianza sólo puede tomar los valores 1,0, según se encuentre presente o no el nivel considerado.

Regresión lineal múltiple
En el caso de más de una variable independiente, se habla de regresión lineal múltiple. Su modelo matemático se presentó en [1].

1.1.2. Procedimiento

La consecución del modelo de regresión exige el siguiente procedimiento:

a) Elegir un método de selección de variables
b) Determinar si hay observaciones que desvirtúen el modelo y analizar las condiciones de aplicación. Es decir, evaluar el modelo
c) Evaluación de la significación en el modelo
d) Interpretar los resultados efectuando una valoración del proceso y del ajuste final obtenido

a) Selección de las variables

Existen diversos criterios: unos, emanados del problema de investigación, y con claros tintes teóricos, y otros por criterios empíricos.

En el primer caso se puede dar ***errores de especificación***, tomando variables o incluyendo otras irrelevantes para la investigación. La inclusión de variables irrelevantes afecta a la parsimonia del modelo, y la falta de variables relevantes influye en el poder explicativo del mismo.

Además del error de especificación, las variables pueden tener errores de medida, que en el caso de las independientes, influyan en las predicciones de la dependiente. Los errores de medida se pueden evaluar mediante ***análisis causal***.

Cuando se utilice variables ficticias, los coeficientes del modelo de regresión representarán las diferencias entre la media del grupo y la del grupo de referencia (el de valor nulo).

Métodos de selección de variables
Entre los procedimientos alternativos a calcular todas las posibles ecuaciones de regresión, en función de todas las combinaciones posibles de las variables independientes, destacan los métodos de construcción por pasos:

a) Método Backward: la ecuación comienza con todas las variables incluidas; en cada paso se eliminará una variable
b) Método Forward: en cada paso se introduce una variable
c) Método Stepwise: en cada paso puede eliminarse o introducirse una variable. Dado que una variable puede entrar y salir de la ecuación en más de una ocasión, es conveniente establecer un límite para el número de pasos. En general, se considera el doble del número de variables independientes. En este procedimiento por pasos se debe tener en cuenta la influencia de la multicolinealidad entre las variables independientes. El investigador debe plantear un modelo teórico con la inclusión de las variables más relevantes y los signos de las mismas.

A la hora de calcular los coeficientes, para asegurar que la tasa de error conjunto a lo largo de todos los tests de significación sea razonable, deben emplearse umbrales muy conservadores (0,01) al añadir o quitar variables (Hair 1999,p.173).

b) Supuestos y limitaciones para la construcción de la ecuación de regresión

Identificar el cumplimiento de los condicionantes del modelo, debe considerarse como paso previo y de validación del análisis de regresión.

Identificación de observaciones influyentes

Hair (1999, p.177) las clasifica en tres grupos: datos atípicos, puntos de apalancamiento e influyentes.

Estos puntos "distintos" se basan en alguna de las siguientes condiciones (Hair 1999, p.178):

- Un error en la entrada de observaciones o datos
- Una observación válida, aunque excepcional, explicable por una situación extraordinaria
- Una observación excepcional sin una explicación plausible
- Una observación ordinaria en sus características individuales pero excepcional en su combinación de características

Estas observaciones influyentes deben ser aisladas antes de comenzar la aplicación del método para evitar defectos en las predicciones realizadas con el mismo.

Los casos atípicos han sido muy estudiados, de forma que se han desarrollado métodos de regresión robustos para minimizar su impacto.

Los datos relevantes (de gran peso o importancia en el modelo), son identificados cuando se emplea el SPSS mediante el "Dfajuste" . Se calcula el valor de la predicción para un elemento, cuando el mismo está vinculado a la muestra y cuando no está

incluido en ella, de tal forma la diferencia viene representada por el valor de "Dfajuste" o su valor tipificado "Dfajuste tipificado".Si esta diferencia es grande la observación (x_i,y_i) tendrá mucha importancia en el modelo de regresión, en caso contrario será menor su influencia. También se puede valorar los casos atípicos a través de los residuos estandarizados cuya distribución es N(0,1) y por tanto, valores mayores a 2 o 3, según criterio del investigador, serán considerados datos atípicos.

Comprobación de las hipótesis del modelo

El modelo de regresión debe: a) estar ***bien especificado***; b) las variables medidas sin error sistemático; y c) los errores en la predicción cumplir unas determinadas condiciones (ser independientes con distribución N(0,σ^2)).

El estar bien definido exige tener unas ***variables independientes relevantes***, o de otra manera, el modelo de regresión debe cumplir el principio de parsimonia, es decir, la conformación del modelo con el menor número posible de variables independientes. Para valorar la aportación de cada variable independiente al modelo habrá que observar si el incremento del coeficiente de determinación (R^2) es significativo.

La existencia de errores sistemáticos de medida, en general, dificulta la creación de cualquier modelo predictivo.

Respecto a los residuos y la definición del modelo, se cumplirá:

a) Linealidad

Cada variable independiente tiene una relación lineal con la dependiente; o de otra forma, para cada variable independiente la linealidad indica que el coeficiente de regresión es constante a lo largo de los valores de la variable independiente (regresión lineal simple). O de forma equivalente $E(e_i)=0$ para i=1,…..,n

La comprobación de la linealidad de cada variable independiente se puede hacer por:

- Los residuos no deben presentar ningún patrón sistemático respecto de las predicciones o respecto de cada una de las variables independientes, se observará mediante el gráfico de ***residuos estandarizados***
- La correlación parcial entre la variable dependiente y cada una de las independientes debe ser alta. También los ***gráficos de regresión parcial*** deben presentar una forma lineal.

b) Homocedasticidad

Las varianzas de las distribuciones de Y ligadas a los distintos valores de las variables independientes deben ser iguales. $Var(Y/x_{i1},x_{i2},\ldots..,x_{ik})=\sigma^2$ o de forma equivalente $Var(e_i)= \sigma^2$, para i=1,….,n:

- Los residuos no deben presentar ningún patrón sistemático respecto de las predicciones o respecto de cada una de las variables independientes
- Se puede emplear el test de Levene. Si hay heterocedasticidad se puede utilizar transformaciones en las variables o el método de mínimos cuadrados ponderados

c) Independencia

El valor observado en una variable para un individuo no debe estar influenciado en ningún sentido por los valores de esta variable observados en otros individuos, es decir, cada variable predictor es independiente. En el supuesto de normalidad, equivale a $Cov(Y_i,Y_j)=0$ si $i\neq j$. y para los residuos, con el mismo supuesto de normalidad, será $Cov(e_i,e_j)=0$ si $i\neq j$. Estas condiciones se traducen:

- Los residuos no deben presentar ningún patrón sistemático respecto a la secuencia de casos
- Los residuos deben estar incorrelados; el estadístico de Durbin-Watson, D, debe tener valores próximos a 2, si D es menor que 1,5 existe autocorrelación. Si D se aproxima a 4 los residuos estarán negativamente autocorrelados y, si se aproxima a 0, estarán positivamente autocorrelados

d) Normalidad

Se cumple $Y/x_{i1},x_{i2},.....,x_{ik}$ es $N(b_0+b_1x_{i1}+.....+b_kx_{ik};\ \sigma^2)$, o de forma equivalente, que la distribución de los residuos sea normal, $N(0,\ \sigma^2)$:

- Los residuos observados y los esperados, bajo hipótesis de distribución normal, deben coincidir
- Para su comprobación se puede utilizar métodos gráficos como el diagrama P-P, o métodos analíticos, como la prueba de Kolmogorov-Smirnov

e) Multicolinealidad

El término multicolinealidad influye en la definición del modelo y se utiliza para describir la situación en que un gran número de variables independientes están altamente interrelacionadas. Las variables que sean aproximadamente una combinación lineal de otras se denominan multicolineales.

Si una variable es una combinación lineal perfecta de otras variables independientes, la matriz de correlaciones será singular (matriz singular es aquella cuyo determinante es igual a 0), lo que se traducirá a la hora de calcular la ecuación de regresión, en que no existirá una única solución mínimo-cuadrática insesgada de cálculo de sus coeficientes.

Una matriz de correlaciones con coeficientes muy altos es un indicio de probable multicolinealidad; sin embargo, puede haber multicolinealidad aunque los coeficientes sean relativamente bajos.

Uno de los procedimientos más utilizado para detectar la interdependencia entre variables es el criterio de la tolerancia.

La ***tolerancia*** de una variable X_i con las restantes variables independientes se define como:

$$Tol_i = 1 - R_i^2$$

donde R_i^2 es el cuadrado del coeficiente de correlación múltiple entre X_i y las variables $X_1,.... X_{i-1},X_{i+1},.....,X_k$

- Si $Tol_i=0$ la variable X_i es casi una combinación lineal de las restantes variables y,
- Si Tol=1 la variable X_i puede reducir la parte de variación de Y no explicada por las restantes variables

En el método de selección de variables por pasos, la variable seleccionada debe tener una tolerancia mínima con las variables incluidas en la ecuación para poder entrar en el siguiente paso. Por otro lado, al entrar la variable, ninguna variable en la ecuación debería superar esa mínima tolerancia con las restantes.

Para solucionar los problemas de multicolinealidad se puede: a) aumentar el tamaño muestral, b) a partir de las variables relacionadas construir otra como combinación lineal de las anteriores y c) utilizar un procedimiento jerárquico para introducir las variables y controlar la tolerancia de las mismas.

c) Evaluación de la significación del modelo de regresión

c1) Estimación de los parámetros

Calcular la ecuación de regresión supone deducir la ecuación del plano que mejor se ajusta a la nube de puntos (Etxeberria 1999, p.54).

Sea $\hat{B}$ un estimador del vector de parámetros **B**. Se define el vector de predicciones como

$$\hat{Y} = X\hat{B}$$

El vector de residuos es

$$e = Y - \hat{Y}$$

Uno de los criterios para obtener los coeficientes de regresión $B_0,B_1,....,B_k$, estimaciones de los parámetros desconocidos $b_0,b_1,.....b_k$, es el de mínimos cuadrados, que consiste en minimizar la suma de los cuadrados de los residuos.

Si en el modelo de regresión se calcula [X'X] y es una matriz no singular, es decir, si su determinante |X'X| es distinto de cero, se puede calcular la inversa $[X'X]^{-1}$ y entonces la matriz de los coeficientes será:

$$\hat{B} = [X'X]^{-1}X'Y$$

Los b_i son los ***coeficientes de regresión parciales***, y así, por ejemplo, b_2 nos da la variación de y, inducida por una variación de X_2, suponiendo que las demás variables permanecen constantes.

c2) Propiedades de los estimadores

Estimador de los coeficientes del modelo lineal

Como hemos visto el estimador de B por el método de mínimos cuadrados es:

$$\hat{B} = [X'X]^{-1}X'Y$$

Es un estimador insesgado con $Var(\hat{B}) = \sigma^2[X'X]^{-1}$

El estimador de la varianza

Una hipótesis del modelo es la homocedasticidad, por tanto, $Var(e_i) = \sigma^2$ para i=1,….,n .El parámetro σ^2 habitualmente es desconocido y por tanto es necesario estimarlo. El estimador de este parámetro es la ***varianza residual*** definida como "el cociente entre la suma de residuos al cuadrado (SC_{res}) y el número de grados de libertad del modelo (gl)"

$$S_e^2 = \frac{SC_{res}}{gl} = \frac{SC_{res}}{n-(k+1)} = \frac{1}{n-(k+1)}\sum_{i=1}^{n} e_i^2$$

Si se utiliza la hipótesis de normalidad se obtiene la relación siguiente de la distribución de S_e^2

$$\frac{S_e^2(n-(k+1))}{\sigma^2} \square \chi^2_{n-(k+1)}$$

Obteniéndose como intervalo de confianza de σ^2 el siguiente:

$$\frac{(n-(k+1))S_e^2}{\chi^2_{n-(k+1)}(1-\frac{\alpha}{2})} \leq \sigma^2 \leq \frac{(n-(k+1))S_e^2}{\chi^2_{n-(k+1)}(\frac{\alpha}{2})}$$

c3) El análisis de la varianza

A continuación se verá la descomposición de la variabilidad de la variable Y cuando se ajusta a un modelo de regresión múltiple.

Se puede comprobar la descomposición de cada observación muestral en:

$$\overbrace{\sum_{i=1}^{n}(y_i-\overline{y})^2}^{SC_{total}}=\overbrace{\sum_{i=1}^{n}(y_i^*-\overline{y})^2}^{SC_{reg}}+\overbrace{\sum_{i=1}^{n}(y_i-y_i^*)^2}^{SC_{res}}$$

o de forma matricial:

SC_{res}= [yy'-b'X'y]
SC_{reg}=[b'X'y-n $\overline{y}^2$]
SC_{total}=[y'y-n $\overline{y}^2$]

La descomposición de la suma de cuadrados nos lleva a la siguiente tabla:

Tabla 5.1.1.- Tabla de análisis de la varianza para el modelo de regresión

Fuente de variación	Suma de cuadrados	Grados de Libertad	Media de cuadrados o varianzas	Estadístico F
Regresión	SC_{reg}	k	$S_R^2 = MC_{reg} = \frac{SC_{reg}}{k}$	$\frac{S_R^2}{S_e^2} = \frac{MC_{reg}}{MC_{res}}$
Residual	SC_{res}	n-(k+1)	$S_e^2 = MC_{res} = \frac{SC_{res}}{n-k-1}$	
Total	SC_{total}	n-1	$S_T^2 = MC_{total} = \frac{SC_{total}}{n-1}$	

Contraste múltiple: H_0:B_1=.....=B_k=0 frente a H_1: ∃ i: $B_i \neq 0$

La hipótesis nula significa que las variables independientes no mejoran la predicción de Y sobre $\overline{y^*} = \overline{y}$ La ***tabla anterior de análisis de la varianza*** a través de $F_{k,n-(k+1)}$ permite estudiar la significación en el contraste múltiple. Si resulta significativo algún B_i es distinto de cero.

Contraste simple: H_0: B_i=0 frente H_1:$B_i \neq 0$

La hipótesis nula significa que la variable X_i no mejora la predicción de Y sobre la regresión obtenida con las k-1 variables restantes.

El estadístico de contraste $t = \frac{B_i}{s_{B_i}}$ donde s_{B_i} se distribuye bajo H_0 como una t de Student con n-(k+1) grados de libertad. Si el p-valor asociado es menor que α, se rechazará la hipótesis nula al nivel de significación α.

Existe otro procedimiento de realizar esta prueba que presenta la mejora respecto al anterior en permitir ejecutar contraste de varias variables a la vez. Para ello si se quiere contrastar la influencia de la variable X_i se ajusta el modelo de regresión completo con las k variables independientes y se calcula la SC_{reg} (k) . Después se realiza el mismo proceso pero con las k-1 variables, todas menos la X_i y se calcula SC_{reg} (k-x_i). Se define la suma de cuadrados incremental debida a X_i como:

$$\Delta SC_{reg}(x_i) = SC_{reg}(k) - SC_{reg}(k - x_i) \geq 0$$

Se plantea la hipótesis anterior H_0: $B_i=0$ frente H_1:$B_i \neq 0$ y se utiliza como estadístico:

$$F_i = \frac{\frac{\Delta SC_{reg}(x_i)}{1}}{S_e^2(k)} \quad i=0,1,\ldots\ldots,k$$

que se distribuye según $F_{1,n-(k+1)}$. Con este procedimiento se obtiene los mismos resultados que con el contraste t, pero además tiene la ventaja que se puede utilizar para un conjunto $l \leq k$ $\{x_{j1},x_{j2},\ldots..,x_{jl}\}$ de variables independiente, dando:

$$F_l = \frac{\frac{\Delta SC_{reg}(l)}{l}}{S_e^2(k)}$$

que se distribuye según una F con l,n-(k+l) grados de libertad.

c4) Análisis de la asociación entre las variables

Al ajustar un modelo de regresión múltiple a una nube de puntos es importante disponer de medidas que permitan medir la bondad del ajuste. Esto se consigue con los coeficientes de correlación múltiple. Como sabemos la correlación mide el grado o fuerza de relación existente entre variables.

El coeficiente de correlación simple (o de Pearson)

El coeficiente de correlación simple ρ, mide el grado de asociación lineal entre las variables X e Y ρ_{xy} es tal que: $-1 \leq \rho_{xy} \leq 1$

- Si $\rho_{xy}=1$ la asociación será lineal positiva
- Si $\rho_{xy}=-1$ la asociación será lineal negativa y,
- Si $\rho_{xy}=0$ no existirá asociación lineal

El estimador muestral del ρ_{xy} es el coeficiente de correlación muestral r_{xy}

$$r(X,Y) = \frac{S(X,Y)}{S_X S_Y}$$

Donde el numerador es la covarianza muestral entre las variables X e Y; S_X,S_Y son las desviaciones típicas muestrales de X e Y respectivamente.

El coeficiente de determinación

En general cuando se ajusta un modelo estadístico a una nube de observaciones, una medida de la bondad de ajuste es el coeficiente de determinación, definido como:

$$R^2 = \frac{SC_{reg}}{SC_{tot}} = \frac{\sum_{i=1}^{n}(y_i^* - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

R^2 representa el porcentaje de variabilidad de Y que explica el modelo de regresión. Si el modelo de ajuste es el lineal múltiple, R se denomina ***coeficiente de correlación múltiple***.

Además, se puede establecer una relación con la F de la ***tabla 5.1.1.*** Como:

$R^2 = \frac{SC_{reg}}{SC_{tot}}$ y F= $\frac{MC_{reg}}{MC_{res}}$ se pueden relacionar F y R mediante: F= $\frac{R^2 / k}{(1-R^2)/[n-(k+1)]}$

Cuando n es pequeño, R^2 es muy sensible a los valores de n y k, entonces los programas de ordenador dan el R^2 (ajustado) que modula la influencia del tamaño muestral en su valor:

$$R^2(ajustado) = 1-(1-R^2)\frac{n-1}{n-(k+1)}$$

El coeficiente de correlación múltiple

El coeficiente de correlación múltiple $\rho_{Y.1..k}$, es una medida del grado de asociación lineal entre Y y el conjunto de variables independientes X_1,....,X_k, y es tal que: $0 \leq \rho_{Y.1..k} \leq 1$

- Si $\rho_{Y.1..k}$, =1 el ajuste del plano de regresión a la población es casi perfecto y
- Si $\rho_{Y.1..k}$, =0 el plano de regresión no mejora la predicción de Y sobre la predicción con la media muestral de Y.

El estimador muestral del $\rho_{Y.1..k}$, es el ***coeficiente de correlación múltiple*** muestral, R.

Todos los cálculos necesarios para el análisis del grado de asociación lineal se suelen disponer en una tabla como la siguiente:

Tabla 5.1.2.- Tabla de análisis de asociación para el modelo de regresión

Fuente de variación	Suma de cuadrados	Varianza	Correlación
Debida a la regresión	$b'X'y - n\bar{y}^2$	S_R^2	$R^2 = \frac{S_R^2}{S_y^2}$
Debida al error	$y'y - b'X'y$	S_e^2	$1-R^2 = \frac{S_e^2}{S_y^2}$
Total	$y'y - n\bar{y}^2$	S_y^2	-----

Con todo lo anterior, el coeficiente de correlación múltiple será:

$$R = \sqrt{1 - \frac{S_e^2}{S_y^2}} = \sqrt{1 - \frac{y'y - b'X'y}{y'y - n\bar{y}^2}}$$

El coeficiente de correlación parcial

Puede interesar estudiar el grado de asociación existente entre dos variables (por ejemplo Y y X_1) una vez que se ha eliminado la influencia que las restantes independientes ejercen sobre ella. Este problema viene resuelto mediante la determinación del ***coeficiente de correlación parcial*,** que representaremos como

$r_{y1.2,3,4,\ldots,k}$

Una de las expresiones más utilizada es:

$$r^2_{12.3,4,\ldots,k} = -\frac{adj\ \sigma_{12}}{\sqrt{adj\ \sigma_{22}.adj\ \sigma_{11}}}$$

Donde adj σ_{12} representa el adjunto del elemento σ_{12} en la matriz de covarianzas.

c5) Predicción en el modelo de regresión lineal múltiple

Uno de los fines primordiales que se persigue al ajustar una función a una nube de puntos es el de poder extrapolar, esto es, dado el valor de la variable/s "independiente/s" exterior al recorrido que presenta la nube de puntos, calcular el correspondiente valor teórico de la variable "dependiente".

El ajuste será más preciso conforme el valor de la variable independiente esté próximo a los valores primitivos.

d) Interpretación de resultados

Para interpretar los resultados del análisis de regresión múltiple será necesario:

Evaluar el coeficiente de regresión

Para ver la influencia de cada variable en el modelo. Se utiliza los coeficientes beta con los datos estandarizados.

Evaluación de la multicolinealidad

- Valorar el grado de multicolinealidad.
- Determinar su impacto en los resultados

Según hemos comentado, para evaluar la colinealidad de parejas o de múltiples variables se utiliza el valor de la tolerancia o su inverso el factor de influencia de la varianza (VIF).

La multicolinealidad hace inestable los coeficientes de la ecuación de regresión aumentando la variación de los mismos y en consecuencia los intervalos de confianza.

Además de interpretar los resultados, el análisis de regresión exige la validación de resultados como observación del poder de generalización de los mismos.

Validación de resultados

- En primer lugar será necesario tener en cuenta el valor de R^2

- Se puede coger una muestra adicional o dividir la muestra
- Se puede utilizar el estadístico "PRESS" que es una medida parecida al R^2 pero para n-1 modelos de regresión. Es un procedimiento similar a las técnicas de "bootstrapping" de remuestreo
- Comparación de los modelos de regresión. Se utilizará distinto número de predictores y/o distinto ajuste (lineal, cuadrático, cúbico, etc). Será necesario utilizar el R^2 ajustado para evitar la influencia del tamaño muestral

1.1.3. Variables de intervención

En ocasiones se desea incluir en la ecuación de regresión, variables categóricas. Para ello es necesario crear las denominadas variables de intervención.

Si la variable independiente es nominal dicotómica, bastará con crear una variable con el valor 0 para una categoría y 1 para la otra e incluir esta variable en la ecuación como una más.

Si la variable independiente es nominal con más de dos categorías, será necesario crear más de una variable. Por ejemplo, si la variable tiene cuatro categorías, A, B, C y D, será necesario crear tres variables de la siguiente forma:

Tabla 5.1.3.- Ejemplo de variables de intervención en el modelo de regresión

X_i	I_1	I_2	I_3
A	1	0	0
B	0	1	0
C	0	0	1
D	0	0	0

Las variables I_1, I_2 y I_3 se incluirán en la ecuación de regresión junto con las restantes variables independientes.

1.2. El comando REGRESSION

Permite realizar análisis de regresión, tanto simple como múltiple, proporcionando diversos métodos y criterios para la construcción de cada ecuación de regresión.

Problema-ejemplo

El ejemplo propuesto recoge los resultados (simulados) de 200 alumnos en una prueba de aptitud musical con seis variables X1(tono), X2(intensidad), X3(ritmo), X4(tiempo), X5(timbre), X6(memoria tonal). La escala de medición es de 0 a 100 para cada variable. Además se almacenó la valoración en una prueba de entonación vocal (Y) en una escala de 0 a 100 (ver fichero regresión.sav).

Se propone realizar un análisis de regresión lineal con Y como variable dependiente.

Desarrollo del ejemplo

1.2.1. Regresión lineal simple

Para comenzar pensemos en un modelo de regresión lineal simple con Y como variable dependiente y X2 (intensidad) como independiente. Para observar el tipo de relación se dibuja un diagrama de dispersión:

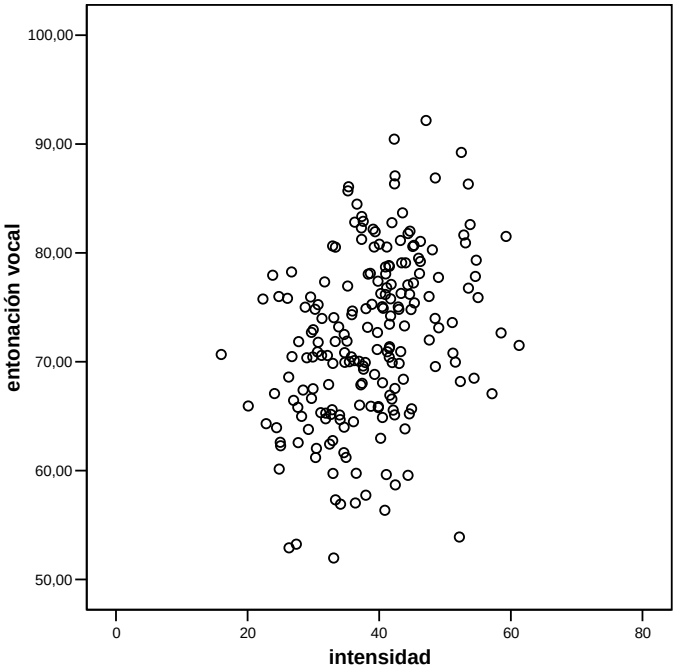

Fig. 5.1.1.- Diagrama de dispersión Y-X2

Como se observa en el gráfico la nube de puntos parece ajustarse a una línea recta, para encontrar la expresión de dicha función mediante SPSS habrá que seleccionar: **Analizar > Regresión >Lineal** y se accederá al cuadro de diálogo de la ***fig. 5.1.2***.

Se seleccionará como variable ***dependiente*** Y (entonación vocal) y como ***independiente*** X2 (intensidad).

Como en otros procedimientos de SPSS, cuando se pulse pegar se añade al fichero de sintaxis. En este caso dicho fichero tomará la expresión:

```
**** Diagrama de dispersión ****************.
GRAPH
  /SCATTERPLOT(BIVAR)=x2 WITH y
  /MISSING=LISTWISE.

******** Análisis de regresión**************.
REGRESSION
```

```
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT y
/METHOD=ENTER x2.
```

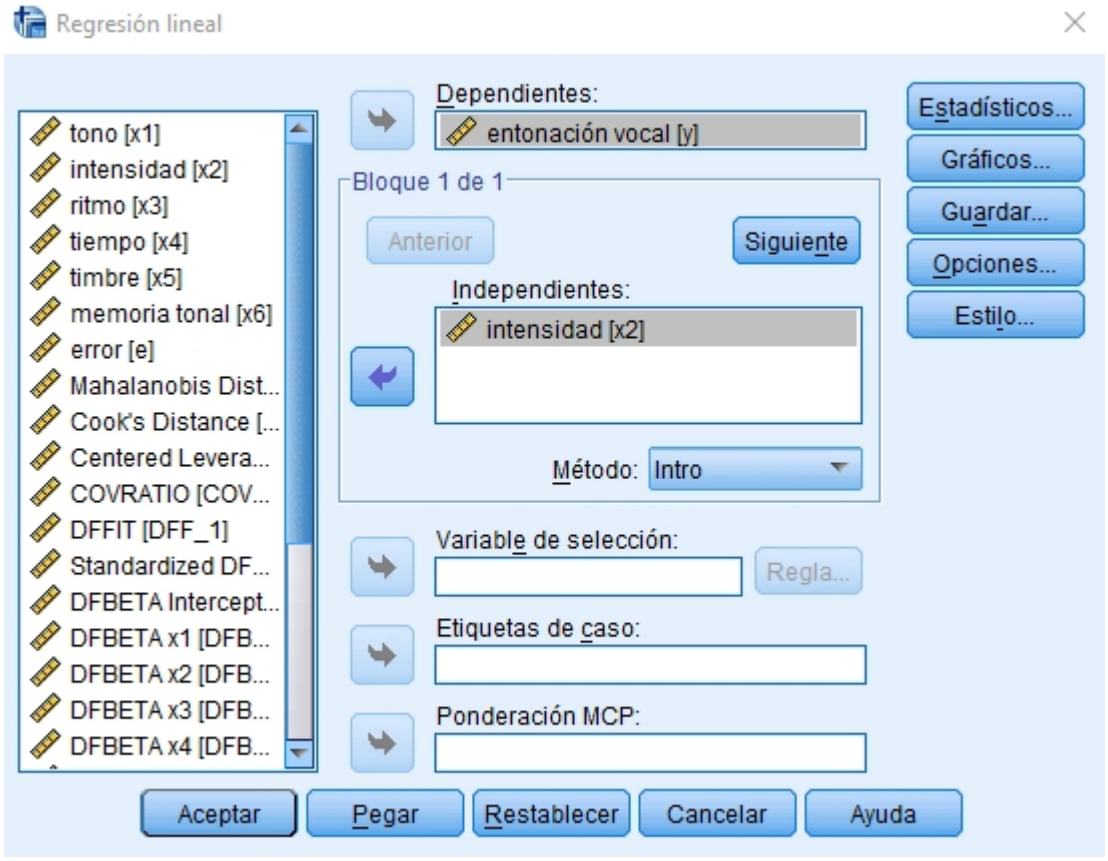

Fig. 5.1.2.- Cuadro de diálogo. Regresión lineal

Los resultados de ejecutar el procedimiento análisis de regresión, con las opciones dadas por defecto por el programa serán:

Tabla 5.1.4.- Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,351(a)	,123	,118	7,31940

a Variables predictoras: (Constante), intensidad

Donde R es el ***coeficiente de correlación*** y R cuadrado el ***coeficiente de determinación***. El valor de éste último permite afirmar que de la variación de la variable dependiente el 12,3% se puede explicar por la variable independiente X2. El error típico de