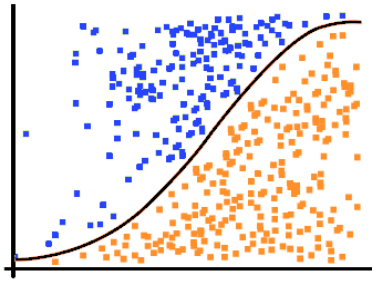
Universidad Autónoma de Madrid

Instrumentos Estadísticos Avanzados
Facultad Ciencias Económicas y Empresariales
Departamento de Economía Aplicada
Profesor: Santiago de la Fuente Fernández

REGRESIÓN LOGÍSTICA



INTRODUCCIÓN

La regresión logística (RL) forma parte del conjunto de métodos estadísticos que caen bajo tal denominación y es la variante que corresponde al caso en que se valora la contribución de diferentes factores en la ocurrencia de un evento simple.

En general, la regresión logística es adecuada cuando la variable de respuesta Y es *politómica* (admite varias categorías de respuesta, tales como *mejora mucho*, *empeora*, *se mantiene*, *mejora*, *mejora mucho*), pero es especialmente útil en particular cuando solo hay dos posibles respuestas (cuando la variable de respuesta es *dicotómica*), que es el caso más común.

La RL es una de las técnicas estadístico-inferenciales más empleadas en la producción científica contemporánea. Surge en la década del 60, su generalización dependía de la solución que se diera al problema de la estimación de los coeficientes. El algoritmo de Walker-Duncan para la obtención de los estimadores de máxima verosimilitud vino a solucionar en parte este problema, pero era de naturaleza tal que el uso de computadoras era imprescindible.

La RL va a contestar a preguntas tales como: ¿Se puede predecir con antelación si un cliente que solicita un préstamo a un banco va a ser un cliente moroso?. ¿Se puede predecir si una empresa va a entrar en bancarrota?. ¿Se puede predecir de antemano que un paciente corra riesgo de un infarto?.

La identificación del mejor modelo de regresión logística se realiza mediante la comparación de modelos utilizando el *cociente de verosimilitud*, que indica a partir de los datos de la muestra cuanto más probable es un modelo frente al otro. La diferencia de los cocientes de verosimilitud entre dos modelos se distribuye según la ley de la Chi-cuadrado con los grados de libertad correspondientes a la diferencia en el número de variables entre ambos modelos.
Si a partir de este coeficiente no se puede demostrar que un modelo resulta mejor que el otro, se considerará como el más adecuado, el más sencillo.

FACTORES DE CONFUSIÓN: Durante el proceso de selección del modelo de regresión más adecuado, el que mejor se ajusta a los datos disponibles, hay que considerar un último aspecto adicional, especialmente si el proceso de selección de variables se hace mediante el *método manual* de obligar a que todas las variables entren en el modelo y es el propio investigador el que paso a paso va construyendo el modelo de regresión más conveniente.

Durante el proceso de incorporación de variables, al eliminar una variable de uno de los modelos de regresión estimados, hay que observar si en el modelo de regresión resultante al excluir esa variable, los coeficientes asociados al resto de variables introducidas en el modelo varían significativamente respecto al modelo de regresión que sí incluía dicha variable. Si así sucede, significa que dicha variable podría ser un *factor de confusión*, al no mostrar una relación significativa con la variable que estamos estudiando directamente, pero sí indirectamente, al relacionarse con otras variables, que en sí mismas pueden estar significativamente relacionadas con la variable de estudio.

En dicho caso, es conveniente no excluir la variable en cuestión del modelo de regresión, aunque no cumpla los requisitos para permanecer en él, obligando a que permanezca, de modo que aunque no se incluya su interpretación al evaluar los resultados del modelo, se ajusta el resultado del resto de variables seleccionadas por su posible efecto.

En la práctica, para incluir o no en la ecuación de regresión una variable de confusión, se utiliza el criterio (incorrectamente) de comprobar si su coeficiente correspondiente es significativamente diferente de cero, por lo que se mira sólo el valor de la probabilidad asociado a ese contraste. Sin embargo, no debe de ser la única razón, hay que considerar si su introducción en la ecuación modifica apreciablemente o no la relación entre la variable dependiente y el otro factor o factores

estudiados. En definitiva, la cuestión debe tratarse con enfoque clínico, puesto que hay que determinar desde ese punto de vista qué se considera como cambio apreciable en el coeficiente de la ecuación de regresión.

Ejemplo: Al estudiar una muestra aleatoria de una población de diabéticos y analizando la posible relación lineal entre la Tensión arterial sistólica (TAS) como variable respuesta y las variables independientes (edad y género de los pacientes), se obtendrá un modelo de regresión donde el género de los pacientes es significativo, es decir, existirá una ecuación diferente de predicción para hombres y otro para mujeres.
Sin embargo, si se controlase también el índice de masa corporal (IMC) introduciéndolo en la ecuación, posiblemente la variable género no sería significativa, mientras que pasaría a serlo el IMC. En ese caso el IMC sería un *factor de confusión* que deberíamos incluir en la ecuación y ello aunque su coeficiente no fuera significativo.

En esta línea, hay que tener cuidado con los términos *relación*, *correlación o significación* y *causalidad*. Que dos factores estén relacionados no implica de ninguna manera que uno sea causa del otro. Es muy frecuente que una alta dependencia indique que las dos variables dependen de una tercera que no ha sido medida (*factor de confusión*).

CONCEPTO DE INTERACCIÓN: Un concepto importante al construir un modelo de regresión es que pueden introducirse términos independientes únicos (una sola variable, por ejemplo efecto del tabaco) y además las interacciones entre variables de cualquier orden (efecto del tabaco según género), si se considera que pueden ser de interés o afectar a los resultados.

Al introducir los términos de interacción en un modelo de regresión es importante para la correcta estimación del modelo respetar un orden jerárquico, es decir siempre que se introduzca un término de interacción de orden superior (x•y•z), deben introducirse en el modelo los términos de interacción de orden inferior (x•y, x•z, y•z) y por supuesto los términos independientes de las variables que participan en la interacción (x, y, z).

Ejemplo: Se desea construir un modelo de regresión para estimar la prevalencia de hipertensos en una muestra y se decide evaluar si la interacción de las variables tabaco, género y edad es significativa o no al estimar dicha prevalencia, por lo que se introduce el término de interacción (tabaco * género * edad).
Automáticamente deberían introducirse igualmente en el modelo los términos de interacción de orden inferiores, es decir, (tabaco*género), (tabaco*edad) y (género*edad), así como los términos independientes tabaco, género y edad para poder estimar el modelo correctamente.
Si se introducen en un modelo de regresión términos de interacción y resultan estadísticamente significativos, no se podrán eliminar del modelo los términos de interacción de orden inferiores ni los términos independientes de las variables que participan en la interacción para simplificarlo, deben mantenerse, aunque no resulten estadísticamente significativos.

VARIABLES DUMMY: Las variables explicativas de tipo nominal con más de dos categorías deben ser incluidas en el modelo definiendo variables *dummy*.

Ejemplo del sentido de las variables *dummy*: Si una variable *nominal* (raza, religión, grupo sanguíneo, etc.) consta de k categorías deben crearse entonces $(k-1)$ variables dicotómicas que son las llamadas variables *dummy* asociadas a la variable nominal. Las $(k-1)$ variables dicotómicas se denotan por $(Z_1, Z_2, \cdots, Z_{k-1})$. A cada categoría o clase de la variable nominal le corresponde un conjunto de valores de los Z_i con el cual se identifica dicha clase.

La manera más usual de definir estas $(k-1)$ variables es la siguiente: si el sujeto pertenece a la *primera* categoría, entonces las $(k-1)$ variables *dummy* valen 0: $(Z_1 = Z_2 = \cdots = Z_{k-1} = 0)$; si el sujeto se halla en la *segunda* categoría, $(Z_1 = 1$ y $Z_2 = \cdots = Z_{k-1} = 0)$; si el sujeto se halla en la *tercera* categoría, $(Z_2 = 1$ y $Z_2 = \cdots = Z_{k-1} = 0)$; y así sucesivamente hasta llegar a la última categoría, para la cual $Z_{k-1} = 1$ y las restantes valen 0.

En esta línea, si la variable nominal de interés es el grupo sanguíneo (tipo 0, tipo A, tipo B, tipo AB), entonces se tendrían los siguientes valores de las variables *dummy* para cada grupo sanguíneo:

Grupo sanguíneo	Z_1	Z_2	Z_3
0	0	0	0
A	1	0	0
B	0	1	0
AB	0	0	1

Si se ajusta un modelo que incluya una variable nominal con k clases, ésta debe ser sustituida por las $(k-1)$ variables *dummy*, y a cada una de ellas corresponderá su respectivo coeficiente.

VARIABLES CUALITATIVAS EN EL MODELO LOGÍSTICO: Como la metodología empleada para la estimación del *modelo logístico* se basa en la utilización de variables cuantitativas, al igual que en cualquier otro procedimiento de regresión, es incorrecto que en él intervengan variables cualitativas, ya sean nominales u ordinales.

La asignación de un número a cada categoría no resuelve el problema. La solución a este problema es crear tantas variables dicotómicas como número de respuestas. Estas nuevas variables, artificialmente creadas, reciben en la literatura anglosajona el nombre de *dummy*, traduciéndose con diferentes denominaciones como pueden ser variables internas, indicadoras, o variables diseño.

Si una variable recoge datos del tabaco con las respuestas (Nunca fumó, Ex-fumador, fuma 20 ó más cigarrillos diarios), hay 4 posibles respuestas por lo que se construyen (4-1=3) variables dummy dicotómicas (valores 0, 1), existiendo diferentes posibilidades de codificación, que conducen a interpretaciones diferentes, siendo la más habitual:

	I1	I2	I3
Nunca fumó	0	0	0
Ex- fumador	1	0	0
< de 20 cigarrillos diarios	0	1	0
≥ 20 cigarrillos diarios	0	0	1

En esta codificación el coeficiente de la ecuación de regresión para cada variable *dummy* (siempre transformado con la función exponencial), se corresponde al *odds-ratio* de esa categoría con respecto al nivel de referencia (la primera respuesta), en el ejemplo cuantifica cómo cambia el riesgo respecto a no haber fumado nunca.

Otra posibilidad es una variable cualitativa de tres respuestas:

	I1	I2
Respuesta 1	0	0
Respuesta 2	1	0
Respuesta 3	1	1

Con esta codificación cada coeficiente se interpreta como una media del cambio del riesgo al pasar de una categoría a la siguiente.

Cuando una categoría no pueda ser considerada de forma natural como nivel de referencia, como por ejemplo el grupo sanguíneo, un posible sistema de clasificación es:

	I1	I2
Respuesta 1	-1	-1
Respuesta 2	1	0
Respuesta 3	0	1

cada coeficiente de las variables *dummy* (indicadoras) tiene una interpretación directa como cambio en el riesgo con respecto a la media de las tres respuestas.

EL MODELO LOGÍSTICO

Sea Y una variable dependiente binaria (con dos posibles valores: 0 y 1). Sean un conjunto de k variables independientes, $(X_1, X_2, \cdots, X_k)$, observadas con el fin de predecir/explicar el valor de Y.

El objetivo consiste en determinar:

$$P[Y=1/X_1,X_2,\cdots,X_k] \;\mapsto\; P[Y=0/X_1,X_2,\cdots,X_k]=1-P[Y=1/X_1,X_2,\cdots,X_k]$$

Para ello, se construye el modelo $P[Y=1/X_1,X_2,\cdots,X_k]=p(X_1,X_2,\cdots,X_k;\beta)$ donde:

$p(X_1,X_2,\cdots,X_k;\beta): R^k \xrightarrow{\text{función de enlace}} [0,1]$ que depende de un vector de parámetros

$\beta=(\beta_1,\beta_2,\cdots,\beta_k)$

FUNCIÓN DE VEROSIMILITUD

Con el fin de estimar $\beta=(\beta_1,\beta_2,\cdots,\beta_k)$ y analizar el comportamiento del modelo estimado se toma una muestra aleatoria de tamaño *n* dada por $(x_i, y_i)_{i=1,2,\cdots,n}$ donde el valor de las variables independientes es $x_i=(x_{i1,}\, x_{i2},\cdots,x_{ik})$ e $y_i \in [0,1]$ es el valor observado de Y en el i-ésimo elemento de la muestra.

Como $(Y/X_1,X_2,\cdots,X_k)\in B[1, p(X_1,X_2,\cdots,X_k;\beta)]$ la función de verosimilitud viene dada por:

$$L[\beta/(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)]=\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \quad \text{donde} \quad p_i=p(x_i;\beta)=p[(x_{i1,}\, x_{i2},\cdots,x_{ik});\beta]_{i=1,2,\cdots,n}$$

- **MODELO LINEAL:** $\begin{cases} 0 & \text{si} \quad \beta_1X_1+\cdots+\beta_kX_k<c_0 \\ \beta_1X_1+\cdots+\beta_kX_k & \text{si} \quad c_0<\beta_1X_1+\cdots+\beta_kX_k\le c_1 \\ 1 & \text{si} \quad \beta_1X_1+\cdots+\beta_kX_k>c_1 \end{cases}$ $\quad c_0, c_1$ son constantes

- **MODELO LOGIT** (*modelo de regresión logística binaria*):

$p(X_1,X_2,\cdots,X_k;\beta)=G[\beta_1X_1+\cdots+\beta_kX_k]$ donde $G(x)=\dfrac{e^x}{1+e^x}$ función distribución función logística

- **MODELO PROBIT:** $p(X_1, X_2, \cdots, X_k; \beta) = \phi[\beta_1 X_1 + \cdots + \beta_k X_k]$ donde ϕ la función de distribución de $N(0,1)$.

MODELO DE REGRESIÓN LOGÍSTICA BINARIA

El modelo logístico establece la siguiente relación entre la probabilidad de que ocurra el suceso, dado que el individuo presenta los valores $(X = x_1, X = x_2, \cdots, X = x_k)$:

$$P[Y = 1/x_1, x_2, \cdots, x_k] = \frac{1}{1 + e^{(-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_k x_k)}}$$

El objetivo es hallar los coeficientes $(\beta_0, \beta_1, \cdots, \beta_k)$ que mejor se ajusten a la expresión funcional.

Se conoce como ***odds*** (*ratio del riesgo*) al cociente de probabilidades:

$$\text{Odds (ratio de riesgo)} = \frac{P[Y = 1/X_1, X_2, \cdots, X_k]}{1 - P[Y = 1/X_1, X_2, \cdots, X_k]} = \frac{p(X_1, X_2, \cdots, X_k; \beta)}{1 - p(X_1, X_2, \cdots, X_k; \beta)} = e^{\beta_1 + \beta_2 X_2 \cdots + \beta_k X_k}$$

se toma como primera variable explicativa a la variable constante que vale 1.

En medicina, por ejemplo, el ratio del riesgo, habitualmente, indica la presencia de una determinada enfermedad objeto de análisis.

Tomando logaritmos neperianos en la expresión anterior, se obtiene una expresión lineal para el modelo:

$$\text{Logit}[P(Y = 1)] = \text{Ln}\left[\frac{P[Y = 1/X_1, X_2, \cdots, X_k]}{1 - P[Y = 1/X_1, X_2, \cdots, X_k]}\right] = \beta_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Aquí se aprecia que el estimador del parámetro β_2 se podrá interpretar como la variación en el *término Logit* (logaritmo neperiano del cociente de probabilidades) originada por una variación unitaria en la variable X_2 (suponiendo constantes el resto de variables explicativas).

Cuando se hace referencia al incremento unitario en una de las variables explicativas del modelo, aparece el concepto de ***oods-ratio*** como el cociente entre los dos odds asociados (el obtenido al realizar el incremento y el anterior al mismo).

Suponiendo que ha habido un incremento unitario en la variable X

$$\text{Odds\_ratio} = \frac{\text{Odds}2}{\text{Odds}1} = e^{\beta_i} \qquad OR = e^{\beta_i}$$

De donde se desprende que, un coeficiente β_i cercano a cero, es decir, un odds-ratio próximo a 1, indicará que cambios en la variable explicativa X_i asociada no tendrán efecto alguno sobre la variable dependiente Y.

Bondad de ajuste del modelo.- Se utilizan dos tipos de contrastes: (a) Contrastes que analizan la bondad de ajuste desde un punto de vista global. (b) Contrastes que analizan la bondad de ajuste paso a paso.

(a) Contraste de bondad de ajuste global de Hosmer-Lemeshow::

☞ El índice de bondad de ajuste:

$z^2=\sum_{i=1}^{n}\frac{(y_i-\hat{p}_i)^2}{\hat{p}_i(1-\hat{p}_i)}$ donde $\hat{p}_i=p(x_{i1},x_{i2},\cdots,x_{ik};\hat{\beta})_{i=1,2,\cdots,n}$, $z^2\approx\chi^2_{n-k}$ si el modelo ajustado es cierto

☞ El *estadístico desviación* viene dado por la expresión:

$$D=2\sum_{i=1}^{n}y_i\,Ln\left[\frac{y_i}{\hat{p}_i}\right]+2\sum_{i=1}^{n-m}(1-y_i)\,Ln\left[\frac{(1-y_i)}{(1-\hat{p}_i)}\right]\quad\begin{cases}m\equiv\text{número observaciones con } y_i=1\\ D\approx\chi^2_{n-k}\ \text{ si el modelo ajustado es cierto}\end{cases}$$

(b) Contraste de bondad de ajuste de Hosmer-Lemeshow:

Evalúa la bondad de ajuste del modelo construyendo una tabla de contingencia a la que aplica un contraste tipo chi-cuadrado.

Calcula los deciles de las probabilidades estimadas $(\hat{p}_i)_{i=1,2,\cdots,n}$, $(D_1,D_2,\cdots,D_9)$, dividiendo los datos observados en diez categorías dadas por: $A_j=\left\{(\hat{p}_i)_{i=1,2,\cdots,n}\in[D_{j-1},D_j)_{j=1,2,\cdots10}\right\}$, donde $D_0=0$ y $D_{10}=1$

El estadístico de contraste:

$$T=\sum_{j=1}^{10}\frac{(e_j-n_j\,\overline{p}_j)^2}{n_j\,\overline{p}_j(1-\overline{p}_j)}\quad\text{donde}\quad\begin{cases}n_j\equiv\text{nº casos en } A_j & (j=1,\cdots,10)\\ e_j\equiv\text{nº } y_i=1 \text{ en } A_j & (j=1,\cdots,10)\end{cases}\qquad\overline{p}_j=\frac{\sum_{i\in A_j}\hat{p}_i}{n_j}$$

p-valor del contraste: $P\left[\chi^2_8\geq T_{observado}\right]$

Diagnósticos del modelo.- Mediante el análisis de los residuos del modelo y de su influencia en la estimación del vector de parámetros se evalúa la bondad del ajuste caso por caso.

☞ *Residuos estandarizados*: $z_i=\frac{y_i-\hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}$

☞ *Residuos studentizados*: $st_i=\frac{y_i-\hat{p}_{(i)}}{\sqrt{\hat{p}_{(i)}(1-\hat{p}_{(i)})}}$, donde $\hat{p}_{(i)}$ es la estimación de p_i obtenida en la observación i-ésima.

☞ Residuos desviación: $(d_i)_{i=1,\cdots,n}=\begin{cases}\sqrt{-2Ln\,\hat{p}_i} & \text{si } y_i=1\\ \sqrt{-2Ln(1-\hat{p}_i)} & \text{si } y_i=0\end{cases}$

Medidas de Influencia.- Cuantifican la influencia que cada observación ejerce sobre la estimación del vector de parámetros o sobre las predicciones hechas a partir del mismo de forma que, cuanto más grande son, mayor es la influencia que ejerce una observación en la estimación del modelo

☞ *Medida de Apalancamiento (Leverage):*

denotando por $W=\text{diagonal}\left[\hat{p}_i(1-\hat{p}_i)\right]$, se calcula a partir de la matriz $H=\sqrt{W}\;X(X'WX)^{-1}X'\sqrt{W}$

El apalancamiento para la observación i-ésima viene dado por el elemento i-ésimo $h_{ii} \in (0,1)$, con un valor medio de p/n.

☞ *Las medidas (distancia de Cook, Dfbeta) miden el impacto que tiene una observación en la estimación de los parámetros.*

Distancia de Cook.- Cuantifica la influencia en la estimación de β:

$COOK_i = \frac{1}{p}\left[\hat{\beta} - \hat{\beta}_{(i)}\right]' (X'WX) \left[\hat{\beta} - \hat{\beta}_{(i)}\right]$ donde $\hat{\beta}_{(i)}$ son estimaciones EMV de β

Dfbeta.- Influencia en la estimación de una componente de β_1: $Dfbeta1_i = \frac{\hat{\beta}_1 - \hat{\beta}_{1(i)}}{st(\hat{\beta}_1)}$

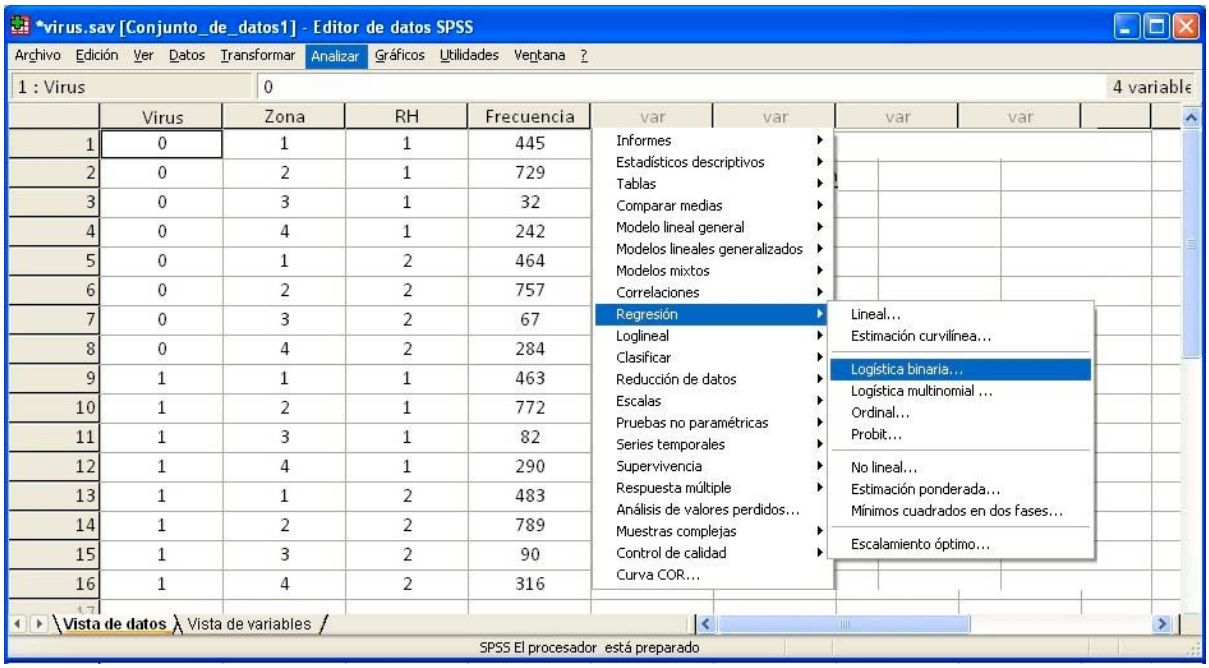
$\hat{\beta}_{1(i)}$ son estimaciones máximo verosímiles (EMV) de β_1

Ejemplo 1.- Se quiere establecer una relación entre el hecho de tener *anticuerpos* a determinado virus con la zona de residencia (norte, sur, este y oeste) y el factor RH.
Para ello, se da la siguiente estructura: *variable nominal* Virus (1-Si, 0-No), *variable nominal* Zona (1-Norte, 2-Sur, 3-Este y 4-Oeste), *variable nominal* RH (1-Positivo, 2-Negativo) y la *variable escalar* Frecuencia.

Señalar que la variable nominal Zona tiene cuatro categorías y debería ser sustituida por 3 variables *dummy*:

Zona Madrid	Z_1	Z_2	Z_3
Norte	0	0	0
Sur	1	0	0
Este	0	1	0
Oeste	0	0	1

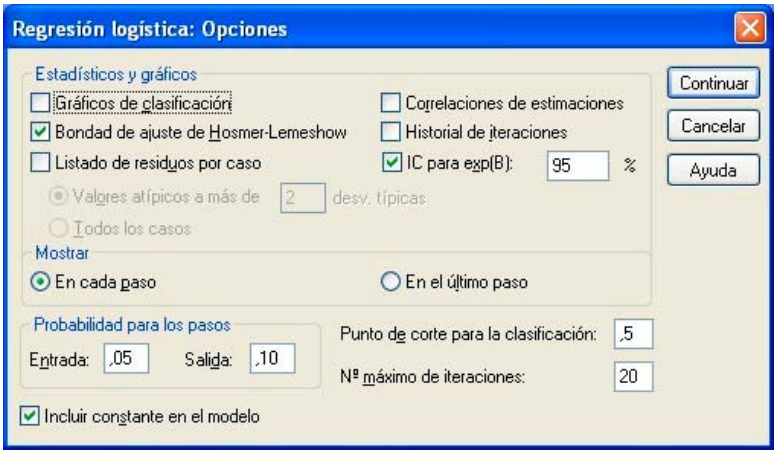
Sin considerar este hecho, introduzcamos los datos en SPSS. Después se ponderan los datos (*Datos/Ponderar casos/frecuencia*).



Se selecciona la variable dependiente (*Virus)* y las covariables (variables independientes: Zona y RH). Ahora tenemos que indicarle al SPSS las variables categóricas, se pulsa el botón [Categóricas].

Se elige el Método *Introducir* (procedimiento en el que todas las variables de un bloque se introducen en un solo paso). Se podía haber utilizado el Método Adelante RV (método automático por pasos, hacia delante, que utiliza la prueba de la Razón de Verosimilitud para comprobar las covariables a incluir o excluir), en este modelo se habría anulado la variable RH de la ecuación.

En [**Opciones**] están disponibles:



El Visor de resultados de SSPS:

Resumen del procesamiento de los casos

Casos no ponderados[a]		N	Porcentaje
Casos seleccionados	Incluidos en el análisis	16	100,0
	Casos perdidos	0	,0
	Total	16	100,0
Casos no seleccionados		0	,0
Total		16	100,0

a. Si está activada la ponderación, consulte la tabla de clasificación para ver el número total de casos.

Aparece un cuadro con el número de casos introducidos (16), los seleccionados para el análisis y los excluidos (casos perdidos, por tener algún valor faltante).

La tabla especifica la codificación de la variable dependiente (que debe ser dicotómica).
Internamente el programa asigna el valor 0 al menor de los dos códigos, y el valor 1 al mayor.

Codificación de la variable dependiente

Valor original	Valor interno
No	0
Si	1

Codificaciones de variables categóricas

		Frecuencia	Codificación de parámetros (1)	(2)	(3)
Zona	Norte	4	1,000	,000	,000
	Sur	4	,000	1,000	,000
	Este	4	,000	,000	1,000
	Oeste	4	,000	,000	,000
RH	Positivo	8	1,000		
	Negativo	8	,000		

La tabla muestra la codificación empleada en las variables independientes y de control (covariables). Se han seleccionado dos variables independientes (Zona, RH) y se refleja la categoría codificada. Además se refleja la frecuencia absoluta de cada valor.

Si en el cuadro de definir *Variables Categóricas* se ha seleccionado en Contraste *Indicador* y en Categoría de referencia *última* (opciones que da el programa por defecto), la categoría codificada con el valor interno más bajo (0) será la de referencia, la 'última' para el SPSS.

La sucesión de estimadores ha convergido, el número de iteraciones necesarias son 3.

Bloque 0: Bloque inicial

Historial de iteraciones[a,b,c]

Iteración		-2 log de la verosimilitud	Coeficientes Constant
Paso 0	1	8729,445	,084
	2	8729,445	,084

a. En el modelo se incluye una constante.

b. -2 log de la verosimilitud inicial: 8729,445

c. La estimación ha finalizado en el número de iteración 2 porque las estimaciones de los parámetros han cambiado en menos de ,001.

En este bloque inicial se calcula la verosimilitud de un modelo que sólo tiene el término constante β_0.

Puesto que la *verosimilitud L* es un número muy pequeño (comprendido entre 0 y 1), se suele ofrecer el *logaritmo neperiano de la verosimilitud* (*LL*), que es un número negativo, o *menos dos veces el logaritmo neperiano de la verosimilitud* (*-2LL*), que es un número positivo.

El estadístico (*-2LL*) mide hasta qué punto un modelo se ajusta bien a los datos. El resultado de esta medición recibe también el nombre de *desviación*. Cuanto más pequeño sea el valor, mejor será el ajuste.

Como en [Opciones] se había solicitado el historial de iteraciones, la salida del ordenador muestra un resumen del proceso iterativo de estimación del primer parámetro β_0, como se observa el proceso ha necesitado dos ciclos para estimar correctamente el término constante $\beta_0 = 0{,}084$, porque la variación de (*-2LL*) entre el primer y segundo bucle ha cambiado en menos del criterio fijado por el programa (0,001).

La tabla permite evaluar el ajuste del modelo de regresión (hasta este momento, con un solo parámetro en la ecuación), comparando los valores predichos con los valores observados.

Tabla de clasificación[a]

Observado			Pronosticado: Virus: No Anticuerpos	Pronosticado: Virus: Anticuerpos	Pronosticado: Porcentaje correcto
Paso 1	Virus	No Anticuerpos	0	3020	,0
		Anticuerpos	0	3285	100,0
	Porcentaje global				52,1

a. El valor de corte es ,500

Por defecto se ha empleado un punto de corte (0,5) de la probabilidad de Y para clasificar a los individuos. Esto significa que aquellos sujetos para los que la ecuación – con éste único término – calcula una probabilidad < 0,5 se clasifican como Virus=0 (No tienen anticuerpos), mientras que si la probabilidad resultante es $\geq 0{,}5$ se clasifican como Virus=1 (tienen anticuerpos).
En este primer paso el modelo ha clasificado correctamente a un 52,1% de los casos, y ningún caso de 'No hay virus' ha sido clasificado correctamente.

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 0	Constante	,084	,025	11,131	1	,001	1,088

En este primer bloque, en la ecuación de regresión sólo aparece el parámetro estimado $\beta_0 = 0{,}084$, el error estándar $E.T = 0{,}025$ y la significación estadística con la prueba de Wald, que es un estadístico que sigue una ley Chi-cuadrado con 1 grado de libertad, y la estimación de la $OR = e^{\beta_0} = e^{0{,}084} = 1{,}088$.

En la tabla de variables que no están asociadas en la ecuación figura la significación estadística asociada al índice de Wald.

Variables que no están en la ecuación

			Puntuación	gl	Sig.
Paso 0	Variables	Zona	16,794	3	,001
		Zona(1)	1,284	1	,257
		Zona(2)	1,792	1	,181
		Zona(3)	14,662	1	,000
		RH(1)	,596	1	,440
	Estadísticos globales		17,602	4	,001

Bloque 1: Método = Introducir

Historial de iteraciones[a,b,c,d]

Iteración		-2 log de la verosimilitud	Coeficientes				
			Constant	Zona(1)	Zona(2)	Zona(3)	RH(1)
Paso 1	1	8711,635	,120	-,102	-,093	,400	,045
	2	8711,623	,120	-,103	-,093	,413	,046
	3	8711,623	,120	-,103	-,093	,413	,046

a. Método: Introducir

b. En el modelo se incluye una constante.

c. -2 log de la verosimilitud inicial: 8729,445

d. La estimación ha finalizado en el número de iteración 3 porque las estimaciones de los parámetros han cambiado en menos de ,001.

En la tabla se muestra el proceso de iteración, que ahora se realiza para tres coeficientes, la constante (ya incluida en el anterior paso), la variable Zona (definida con tres variables *dummy:* Z_1, Z_2 y Z_3), y la variable RH.

Se observa como disminuye el (*-2LL*) respecto al paso anterior (el modelo sólo con la constante tenía un valor de este estadístico de 8729,445, mientras que ahora se reduce a 8711,623), y el proceso termina con tres bucles.
Los coeficientes calculados son para la constante $\beta_0 = 0{,}120$, para la variable Zona, respectivamente, los coeficientes de Z_1, Z_2 y Z_3 (0,103 ; 0,093 ; 0,413), y para la variable RH el coeficiente 0,046.

Se muestra una tabla chi-cuadrado que evalúa la hipótesis nula de que los coeficientes β_i de todos los términos (excepto la constante) incluidos en el modelo son cero.

Pruebas omnibus sobre los coeficientes del modelo

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	17,822	4	,001
	Bloque	17,822	4	,001
	Modelo	17,822	4	,001

El estadístico chi-cuadrado para este contraste es la diferencia entre el valor de (*-2LL*) para el modelo sólo con la constante (-2LL = 8729,445) y el valor (*-2LL*) para el modelo actual (-2LL = 8711,623), es decir, el cociente o razón de verosimilitudes:

$$RV = \chi^2_4 = (-2LL_{MODELO\ 0}) - (-2LL_{MODELO\ 1}) = 8711{,}623 - 8711{,}623 = 17{,}822$$

En general, la razón de verosimilitudes (RV) es útil, para determinar si hay una diferencia significativa entre incluir en modelo todas las variables y no incluir ninguna, dicho de otro modo, RV sirve para evaluar sí las variables tomadas en conjunto, contribuyen efectivamente a 'explicar' las modificaciones que se producen en $P(Y = 1)$.

Prueba Omnibus, SSPS ofrece tres entradas (Paso, Bloque y Modelo):

- La fila primera (PASO) es la correspondiente al cambio de verosimilitud (de *-2LL*) entre pasos sucesivos en la construcción del modelo, contrastando la hipótesis nula H_0 de que los coeficientes de las variables añadidas en el último paso son cero.
- La segunda fila (BLOQUE) es el cambio en *-2LL* entre bloques de entrada sucesivos durante la construcción del modelo. Si como es habitual en la práctica se introducen las variables en un solo bloque, la Chi-Cuadrado del Bloque es el mismo que la Chi-Cuadrado del Modelo.

- La tercera fila (MODELO) es la diferencia entre el valor de *-2LL* para el modelo sólo con la constante y el valor de -2LL para el modelo actual.

Resumen de los modelos

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	8711,623[a]	,003	,004

a. La estimación ha finalizado en el número de iteración 3 porque las estimaciones de los parámetros han cambiado en menos de ,001.

Seguidamente, tres medidas *Resumen de los modelos*, para evaluar de forma global su validez.

Los coeficientes de determinación tienen valores muy pequeños, indicando que sólo el 0,3% o el 0,4% de la variación de la variable dependiente es explicada por las variables incluidas en el modelo, y debe mejorar cuando se vayan incluyendo variables más explicativas del resultado o términos de interacción.

- – 2 logaritmo de la verosimilitud (*–2LL*) mide hasta qué punto un modelo se ajusta bien a los datos. El resultado de esta medición recibe también el nombre de *desviación*. Cuanto más pequeño sea el valor, mejor será el ajuste.

- La R cuadradro de *Cox y Snell* es un coeficiente de determinación generalizado que se utiliza para estimar la proporción de varianza de la variable dependiente explicada por las variables predictoras (independientes).
  La R cuadrado de *Cox y Snell* se basa en la comparación del logaritmo de la verosimilitud (*LL*) para el modelo respecto al logaritmo de la verosimilitud (*LL*) para un modelo de línea base. Los valores oscilan entre 0 y 1.

- La R cuadrado de *Nagelkerke* es una versión corregida de la R cuadrado de *Cox y Snell*.
  La R cuadrado de *Cox y Snell* tiene un valor máximo inferior a 1, incluso para un modelo "perfecto". La R cuadrado de *Nagelkerke* corrige la escala del estadístico para cubrir el rango completo de 0 a 1.

Prueba de Hosmer y Lemeshow

Paso	Chi-cuadrado	gl	Sig.
1	,459	4	,977

La bondad de ajuste ha resultado excelente, basta notar la similitud entre valores esperados y observados en el procedimiento de Hosmer y Lemeshow.

Tabla de contingencias para la prueba de Hosmer y Lemeshow

		Virus = No Anticuerpos		Virus = Anticuerpos		
		Observado	Esperado	Observado	Esperado	Total
Paso 1	1	464	469,327	483	477,673	947
	2	757	762,632	789	783,368	1546
	3	445	439,673	463	468,327	908
	4	729	723,368	772	777,632	1501
	5	284	281,990	316	318,010	600
	6	341	343,010	462	459,990	803

La prueba de *Hosmer-Lemeshow* es otra prueba para evaluar la *bondad del ajuste* de un modelo de regresión logística (RL).

Parte de la idea de que si el ajuste es bueno, un valor alto de la probabilidad predicha (p) se asociará con el resultado 1 de la variable binomial dependiente, mientras que un valor bajo de p (próximo a cero) corresponderá (en la mayoría de las ocasiones) con el resultado Y=0.

Para cada observación del conjunto de datos, se trata de calcular las probabilidades de la variable dependiente que predice el modelo, ordenarlas, agruparlas y calcular, a partir de ellas, las frecuencias esperadas, y compararlas con las observadas mediante una prueba chi-cuadrado.

Señalar que esta prueba de *bondad de ajuste* tiene algunas 'inconvenientes': El estadígrafo de *Hosmer-Lemeshow* no se computa cuando, para algunos grupos, e_i (valores esperados) ó $e_i*(n_i - e_i)$ son nulos o muy pequeños (menores que 5). Por otra parte, lo que se desea en esta prueba es que no haya significación (lo contrario a lo que suele ser habitual). Por este motivo, muchos autores proponen simplemente cotejar valores observados y esperados mediante simple inspección y evaluar el grado de concordancia entre unos y otros a partir del sentido común.

Sobre este razonamiento, una forma de evaluar la ecuación de regresión y el modelo obtenido es construir una tabla 2×2 clasificando a todos los individuos de la muestra según la concordancia de los valores observados con los predichos o estimados por el modelo, de forma similar a como se evalúan las pruebas diagnósticas.

Una ecuación sin poder de clasificación alguno tendría una especificidad, sensibilidad y total de clasificación correctas igual al 50% (por el simple azar). Un modelo puede considerarse aceptable si tanto la especificidad como la sensibilidad tienen un nivel alto, de al menos el 75%.

Tabla de clasificación[a]

Observado			Pronosticado		
			Virus		
			No Anticuerpos	Anticuerpos	Porcentaje correcto
Paso 1	Virus	No Anticuerpos	0	3020	,0
		Anticuerpos	0	3285	100,0
	Porcentaje global				52,1

a. El valor de corte es ,500

El modelo tiene una especificidad alta (100%) y una sensibilidad nula (0%). Con la constante y una única variable predictora (Virus), clasifica mal a los individuos que no tienen anticuerpos cuando el punto de corte de la probabilidad de Y calculada se establece (por defecto) en 50% (0,5).

Por último, SPSS ofrece las variables de la ecuación, los coeficientes de regresión con sus correspondientes errores estándar (ET), el valor del *estadístico de Wald* para evaluar la hipótesis nula ($p_i = 0$), la significación estadística asociada, y el valor de la OR= exp(β_i) con sus intervalos de confianza.

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95,0% para EXP(B)	
								Inferior	Superior
Paso 1[a]	Zona			16,746	3	,001			
	Zona(1)	-,103	,076	1,843	1	,175	,903	,778	1,047
	Zona(2)	-,093	,070	1,792	1	,181	,911	,794	1,044
	Zona(3)	,413	,140	8,762	1	,003	1,511	1,150	1,987
	RH(1)	,046	,051	,811	1	,368	1,047	,948	1,156
	Constante	,120	,064	3,512	1	,061	1,128		

a. Variable(s) introducida(s) en el paso 1: Zona, RH.

El modelo ajustado resulta: $P[\text{anticuerpos}] = \frac{1}{1 + e^{(-0{,}120 + 0{,}103Z_1 + 0{,}093Z_2 - 0{,}413Z_3 + 0{,}046RH)}}$

Para estimar, mediante el modelo, la tasa de anticuerpos entre sujetos del ESTE (1-Norte, 2-Sur, 3-Este y 4-Oeste) que tienen RH negativo, se tendría que sustituir en la ecuación los valores $(Z_1 = 0, Z_2 = 1, Z_3 = 0, RH = 2)$

$$P[\text{anticuerpos}] = \frac{1}{1 + e^{(-0{,}120 + 0{,}093.1 + 0{,}046.2)}} = 0{,}937$$

Computando la tasa de sujetos con anticuerpos en esta subpoblación (Este, RH negativo) utilizando la información original, siendo $\overline{RH}_{(anticuerpo\ s,\ zona)}$, la razón sería:

$$\frac{\overline{RH}_{(1,3)}}{\overline{RH}_{(1,3)}+\overline{RH}_{(0,3)}}=\frac{90}{90+67}=0{,}573$$

virus2.sav [Conjunto_de_datos1] - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?

21 : Visible

	Virus	Zona	RH	Zona_RH	Frecuencia
1	0	1	1	1	445
2	0	2	1	2	729
3	0	3	1	3	32
4	0	4	1	4	242
5	0	1	2	2	464
6	0	2	2	4	757
7	0	3	2	6	67
8	0	4	2	8	284
9	1	1	1	1	463
10	1	2	1	2	772
11	1	3	1	3	82
12	1	4	1	4	290
13	1	1	2	2	483
14	1	2	2	4	789
15	1	3	2	6	90
16	1	4	2	8	316
17					

Vista de datos | Vista de variables

SPSS El procesador está preparado

Si el modelo contempla la interacción (Zona_RH) se debe incluir como una variable más el producto de las dos variables (Zona*RH), sin codificar los valores de la nueva variable, sino simplemente el producto de ambas.

Sin embargo, puesto que Zona ha de tratarse a través de las variables dummy (indicadoras), en este caso crear la variable (Zona*RH) sería incorrecto. Para hacer el ajuste incorporando la interacción de Zona y RH no se debe indicar a SPSS que maneje Zona a través de variables dummy, sino que deben construirse las tres variables dummy previamente y luego los tres productos procedentes de éstas con RH. La tabla de contingencia resultante sería:

virus3.sav [Conjunto_de_datos1] - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?

19 : Frecuencia Visible: 9 de 9 varia

	Virus	Z1	Z2	Z3	RH	Z1_RH	Z2_RH	Z3_RH	Frecuencia
1	0	0	0	0	1	0	0	0	445
2	0	1	0	0	1	1	0	0	729
3	0	0	1	0	1	0	1	0	32
4	0	0	0	1	1	0	0	1	242
5	0	0	0	0	2	0	0	0	464
6	0	1	0	0	2	2	0	0	757
7	0	0	1	0	2	0	2	0	67
8	0	0	0	1	2	0	0	2	284
9	1	0	0	0	1	0	0	0	463
10	1	1	0	0	1	1	0	0	772
11	1	0	1	0	1	0	1	0	82
12	1	0	0	1	1	0	0	1	290
13	1	0	0	0	2	0	0	0	483
14	1	1	0	0	2	2	0	0	789
15	1	0	1	0	2	0	2	0	90
16	1	0	0	1	2	0	0	2	316
17									

Vista de datos | Vista de variables

SPSS El procesador está preparado

Después de ponderar los datos por la frecuencia, al realizar el contraste en SPSS:

En el Visor de resultados del Bloque 1: Método Introducir

Historial de iteraciones[a,b,c,d]

Iteración		-2 log de la verosimilitud	Coeficientes						
			Constant	Z1(1)	Z2	Z3	RH(1)	Z1_RH	Z2_RH	Z3_RH
Paso 1	1	8705,929	,074	-,034	1,422	,215	,000	-,016	-,585	-,074
	2	8705,834	,074	-,034	1,546	,216	,000	-,016	-,645	-,075
	3	8705,834	,074	-,034	1,548	,216	,000	-,016	-,646	-,075
	4	8705,834	,074	-,034	1,548	,216	,000	-,016	-,646	-,075

a. Método: Introducir

b. En el modelo se incluye una constante.

c. -2 log de la verosimilitud inicial: 8729,445

d. La estimación ha finalizado en el número de iteración 4 porque las estimaciones de los parámetros han cambiado en menos de ,001.

El proceso de iteración se realiza para ocho coeficientes. Los coeficientes calculados son, respectivamente, para la constante $\beta_0 = 0{,}074$, y para la variables Z_1, Z_2, Z_3, RH, Z1_RH, Z2_RH, Z3_RH.

Se muestra una tabla chi-cuadrado que evalúa la hipótesis nula de que los coeficientes β_i de todos los términos (excepto la constante) incluidos en el modelo son cero.

Pruebas omnibus sobre los coeficientes del modelo

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	23,610	7	,001
	Bloque	23,610	7	,001
	Modelo	23,610	7	,001

El estadístico chi-cuadrado para este contraste es la diferencia entre el valor de (*-2LL*) para el modelo sólo con la constante (-2LL = 8729,445) y el valor (*-2LL*) para el modelo actual (-2LL = 8705,834), es decir, el cociente o razón de verosimilitudes:

$$RV = \chi^2_4 = (-2LL_{\text{MODELO 0}}) - (-2LL_{\text{MODELO 1}}) = 8729{,}445 - 8705{,}8334 = 23{,}610$$

En general, la razón de verosimilitudes (RV) es útil, para determinar si hay una diferencia significativa entre incluir en modelo todas las variables y no incluir ninguna, dicho de otro modo, RV sirve para evaluar sí las variables tomadas en conjunto, contribuyen efectivamente a 'explicar' las modificaciones que se producen en $P(Y = 1)$.

Resumen de los modelos

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	8705,834[a]	,004	,005

a. La estimación ha finalizado en el número de iteración 4 porque las estimaciones de los parámetros han cambiado en menos de ,001.

El coeficiente de determinación tiene un valor muy pequeño, indicando que sólo el 0,4% de la variación de la variable dependiente es explicada por las variables incluidas en el modelo.

Prueba de Hosmer y Lemeshow

Paso	Chi-cuadrado	gl	Sig.
1	,000	4	1,000

La bondad de ajuste ha resultado excelente, basta notar la similitud entre valores esperados y observados en el procedimiento de Hosmer y Lemeshow.

Tabla de contingencias para la prueba de Hosmer y Lemeshow

		Virus = No Anticuerpos		Virus = Anticuerpos		
		Observado	Esperado	Observado	Esperado	Total
Paso 1	1	445	445,000	463	463,000	908
	2	464	464,000	483	483,000	947
	3	757	757,000	789	789,000	1546
	4	729	729,000	772	772,000	1501
	5	284	284,000	316	316,000	600
	6	341	341,000	462	462,000	803

Por último, SPSS ofrece las variables de la ecuación, los coeficientes de regresión con sus correspondientes errores estándar (ET), el valor del *estadístico de Wald* para evaluar la hipótesis nula ($p_i = 0$), la significación estadística asociada, y el valor de la OR= exp(β_i) con sus intervalos de confianza.

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95,0% para EXP(B)	
								Inferior	Superior
Paso 1[a]	Z1(1)	-,034	,187	,033	1	,856	,967	,669	1,395
	Z2	1,548	,471	10,806	1	,001	4,701	1,868	11,828
	Z3	,216	,243	,792	1	,373	1,241	,771	1,997
	RH(1)	,000	,093	,000	1	,996	1,000	,833	1,199
	Z1_RH	-,016	,118	,019	1	,889	,984	,781	1,239
	Z2_RH	-,646	,279	5,348	1	,021	,524	,303	,906
	Z3_RH	-,075	,151	,243	1	,622	,928	,690	1,248
	Constante	,074	,219	,115	1	,734	1,077		

a. Variable(s) introducida(s) en el paso 1: Z1, Z2, Z3, RH, Z1_RH, Z2_RH, Z3_RH.

El modelo ajustado resulta:

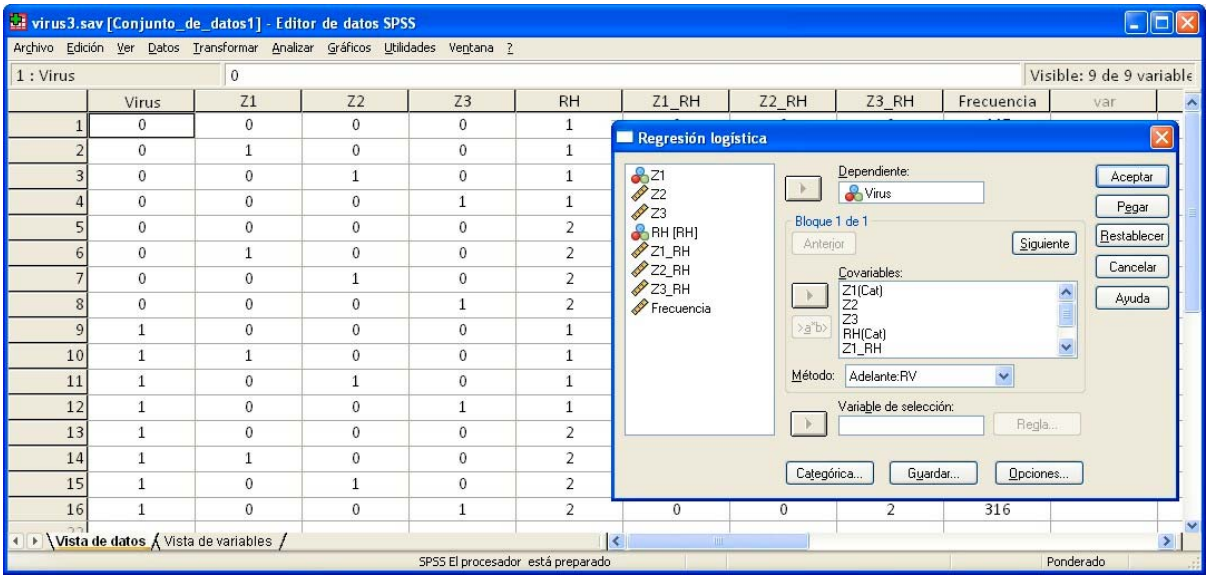
$$P[\text{anticuerpos}] = \frac{1}{1 + e^{(-0,074 + 0,034 Z_1 - 1,548 Z_2 - 0,216 Z_3 + 0,016 Z_1\_RH + 0,646 Z_2\_RH + 0,075 Z_3\_RH)}}$$

NOTA.- Las variables con un error estándar mayor que 1 no entrarían en el modelo sean o no significativas, o las que tienen un OR muy grande o cercano a cero.
El OR= exp(β_i) es una medida estadística que cuantifica el riesgo que representa poseer el factor correspondiente o no poseerlo, suponiendo que el resto de variables del modelo permanecen constantes. Un *odds-ratio* próximo a 1 ($OR = e^{\beta_i}$), es decir, un coeficiente β_i cercano a cero, indicará que cambios en la variable explicativa asociada no tendrán efecto alguno sobre la variable dependiente. Para determinar si el OR es significativamente distinto de 1 se calcula su intervalo de confianza [$OR < 1$ es un factor protector, $OR = 1$ es un factor que no es protector ni de riesgo, $OR > 1$ es un factor de riesgo]. Es significativo cuando su p_valor (*Signatura*) < 0,05

Las variables Z1, Z3, RH, Z1_RH, Z3_RH tienen intervalos de confianza que cubre el 1, por lo que no tienen efecto alguno sobre la variable respuesta (*anticuerpos*).

Las variables que entran en la ecuación son Z2, Z2_RH, sólo hay que analizar estas dos variables y se inicia el procedimiento de nuevo con el **Método Introducir**.

A la misma conclusión se hubiera llegado si se hubiera elegido el **Método Adelante RV** (método automático por pasos, hacia delante, que utilizará la prueba de la Razón de Verosimilitud para comprobar las covariables a incluir o excluir).



SPSS ofrece las variables que dejará en la ecuación, sus coeficientes de regresión con sus correspondientes errores estándar, el valor del estadístico de Wald para evaluar la hipótesis nula (Pi=0), la significación estadística asociada, y el valor de la OR (exp(B)) con sus intervalos de confianza.

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95,0% para EXP(B)	
								Inferior	Superior
Paso 1[a]	Z2	,489	,129	14,407	1	,000	1,630	1,267	2,098
	Constante	,064	,026	6,107	1	,013	1,066		
Paso 2[b]	Z2	1,523	,448	11,573	1	,001	4,587	1,907	11,032
	Z2_RH	-,646	,264	6,004	1	,014	,524	,313	,879
	Constante	,064	,026	6,107	1	,013	1,066		

a. Variable(s) introducida(s) en el paso 1: Z2.

b. Variable(s) introducida(s) en el paso 2: Z2_RH.

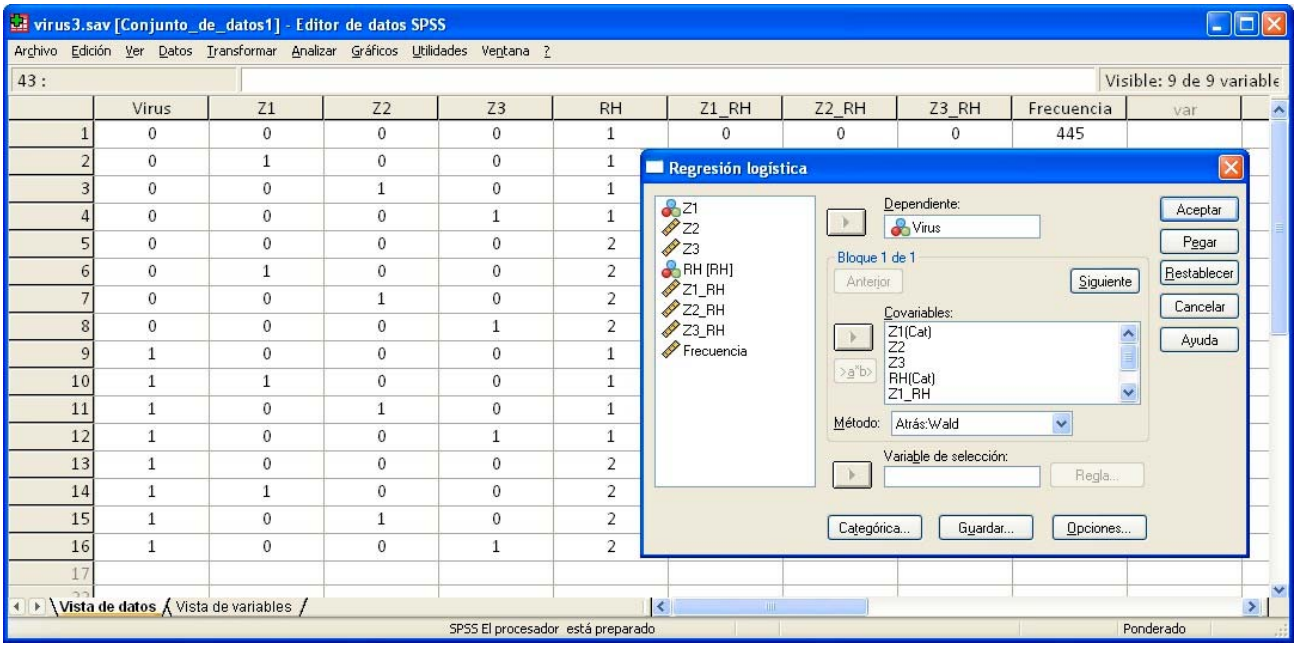
Modelo si se elimina el término

Variable		Log verosimilitud del modelo	Cambio en -2 log de la verosimilitud	gl	Sig. del cambio
Paso 1	Z2	-4364,722	14,878	1	,000
Paso 2	Z2	-4360,396	12,391	1	,000
	Z2_RH	-4357,283	6,166	1	,013

Se muestra una evaluación de cuánto perdería el modelo obtenido si se eliminara la variable incluida en este paso, ya que en los métodos automáticos de construcción del modelo por pasos el proceso evalúa la inclusión y la exclusión de variables.

La tabla presenta, para cada variable del modelo, los cambios en la verosimilitud si dichas variables se eliminan; si la significación estadística asociada (*Sig. del cambio*) fuese mayor que el criterio de exclusión establecido, la variable se eliminaría del modelo en el paso siguiente.
Como el cambio de verosimilitud es estadísticamente significativo (< 0,05), las variables quedan en el modelo.

Sí se hubiera elegido el **Método Atrás Wald** (método automático de selección por pasos hacia atrás, el contraste para la eliminación se basa en la probabilidad del estadístico de Wald). Se selecciona hacia atrás porque se desea que el modelo incluya en un principio todas las variables independientes y vaya quitando variables en cada paso hasta solo quedar las variables explicativas.



Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95,0% para EXP(B)	
								Inferior	Superior
Paso 1[a]	Z1(1)	-,034	,187	,033	1	,856	,967	,669	1,395
	Z2	1,548	,471	10,806	1	,001	4,701	1,868	11,828
	Z3	,216	,243	,792	1	,373	1,241	,771	1,997
	RH(1)	,000	,093	,000	1	,996	1,000	,833	1,199
	Z1_RH	-,016	,118	,019	1	,889	,984	,781	1,239
	Z2_RH	-,646	,279	5,348	1	,021	,524	,303	,906
	Z3_RH	-,075	,151	,243	1	,622	,928	,690	1,248
	Constante	,074	,219	,115	1	,734	1,077		
Paso 2[a]	Z1(1)	-,033	,124	,072	1	,788	,967	,758	1,234
	Z2	1,547	,449	11,848	1	,001	4,697	1,947	11,334
	Z3	,215	,198	1,183	1	,277	1,240	,841	1,828
	Z1_RH	-,016	,072	,048	1	,826	,984	,854	1,134
	Z2_RH	-,646	,264	6,004	1	,014	,524	,313	,879
	Z3_RH	-,074	,119	,386	1	,535	,929	,735	1,173
	Constante	,073	,115	,404	1	,525	1,076		
Paso 3[a]	Z1(1)	-,009	,059	,025	1	,874	,991	,883	1,112
	Z2	1,547	,449	11,848	1	,001	4,697	1,947	11,334
	Z3	,215	,198	1,183	1	,277	1,240	,841	1,828
	Z2_RH	-,646	,264	6,004	1	,014	,524	,313	,879
	Z3_RH	-,074	,119	,386	1	,535	,929	,735	1,173
	Constante	,049	,036	1,846	1	,174	1,050		
Paso 4[a]	Z2	1,541	,448	11,838	1	,001	4,670	1,941	11,235
	Z3	,209	,194	1,159	1	,282	1,233	,842	1,805
	Z2_RH	-,646	,264	6,004	1	,014	,524	,313	,879
	Z3_RH	-,074	,119	,386	1	,535	,929	,735	1,173
	Constante	,046	,029	2,559	1	,110	1,047		
Paso 5[a]	Z2	1,541	,448	11,838	1	,001	4,670	1,941	11,235
	Z3	,096	,066	2,105	1	,147	1,101	,967	1,253
	Z2_RH	-,646	,264	6,004	1	,014	,524	,313	,879
	Constante	,046	,029	2,559	1	,110	1,047		
Paso 6[a]	Z2	1,523	,448	11,573	1	,001	4,587	1,907	11,032
	Z2_RH	-,646	,264	6,004	1	,014	,524	,313	,879
	Constante	,064	,026	6,107	1	,013	1,066		

a. Variable(s) introducida(s) en el paso 1: Z1, Z2, Z3, RH, Z1_RH, Z2_RH, Z3_RH.

- En el paso 3 habían entrado las variables (Z1, Z2, Z3, Z2_RH, Z3_RH), en el paso 4 queda eliminada la variable Z1 porque tiene el mayor OR próximo a cero, el intervalo de confianza del OR cubre el 1 (no tiene efecto alguno sobre la variable dependiente).

- En el paso 4 habían entrado las variables (Z2, Z3, Z2_RH, Z3_RH), en el paso 5 queda eliminada la variable Z3_RH porque tiene el mayor OR próximo a cero, el intervalo de confianza del OR cubre el 1 (no tiene efecto alguno sobre la variable dependiente).

- En el paso 5 habían entrado las variables (Z2, Z3, Z2_RH), en el paso 6 queda eliminada la variable Z3 porque el intervalo de confianza del OR cubre el 1, en consecuencia, no tiene efecto alguno sobre la variable dependiente.

Las variables que entran en la ecuación son Z2, Z2_RH, sólo hay que analizar estas dos variables y se inicia el procedimiento de nuevo con el **Método Introducir**.

Si se tuviera alguna otra variable que podría modificar el modelo se podría introducir en covariables (variables independientes).
En Método Introducir.

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95,0% para EXP(B)	
								Inferior	Superior
Paso 1[a]	Z2	1,523	,448	11,573	1	,001	4,587	1,907	11,032
	Z2_RH	-,646	,264	6,004	1	,014	,524	,313	,879
	Constante	,064	,026	6,107	1	,013	1,066		

a. Variable(s) introducida(s) en el paso 1: Z2, Z2_RH.

El modelo ajustado resulta: $P[\text{anticuerpos}] = \frac{1}{1 + e^{(-0,064 \; - \; 1,523 Z_2 \; + \; 0,646 Z_2\_RH)}}$

Ejemplo 2.- Se estudia la infección hospitalaria posquirúrgica en pacientes operados de rodilla a lo largo de la primera semana. Para evaluar un nuevo régimen de la atención de enfermería que se dispensa a los pacientes se hace un estudio a ochenta pacientes de edades diferentes, donde 36 se ubican en la atención establecida y 44 en la atención en estudio. Obteniendo la tabla siguiente:

	Infección	
Atención	Sí (1)	No (0)
Estudio (1)	7	37
Establecida (0)	14	22

$$OR = \frac{7.22}{14.37} = 0{,}279$$

Si se considera la variable *edad del paciente* (< 40 años, ≥40 años), se introduce una *variable de confusión* en la relación que pudiera existir en la relación (atención - desarrollar infección).

La distribución de pacientes según régimen de atención enfermera, condición respecto de la infección y grupo de edad, viene dada por la tabla adjunta:

		Infección	
	Atención	Sí (1)	No (0)
Edad < 40 (1)	Estudio (1)	2	22
	Establecida (0)	2	9
Edad ≥ 40 (2)	Estudio (1)	5	15
	Establecida (0)	12	13

$$OR_1 = \frac{2.9}{2.22} = 0{,}41$$

$$OR_2 = \frac{5.13}{12.15} = 0{,}36$$

Los datos de la tabla de contingencia de 3 entradas con 8 celdas:

Infección	Atención	Edad	Frecuencia
0	0	1	9
0	0	2	13
0	1	1	22
0	1	2	15
1	0	1	2
1	0	2	12
1	1	1	2
1	1	2	5

La asociación entre la atención y la infección puede ser omitida o falsamente detectada en caso de que exista un factor de confusión. Un factor de confusión es el que se asocia con la atención de enfermería y la infección de los pacientes.

Para analizar la asociación entre la atención de enfermería y la infección a lo largo de la semana de los pacientes: *Analizar/Estadísticos descriptivos/Tablas de contingencia*

En [Estadísticos] se selecciona *Riesgo*.
En [Casillas] se selecciona *Porcentaje en columnas*.

Tabla de contingencia Atención * Infección

			Infección		
			No se infecta	Se infecta	Total
Atención	Atención establecida	Recuento	22	14	36
		% de Infección	37,3%	66,7%	45,0%
	Atención nueva	Recuento	37	7	44
		% de Infección	62,7%	33,3%	55,0%
Total		Recuento	59	21	80
		% de Infección	100,0%	100,0%	100,0%

Se calcula el OR de la atención establecida respecto a la atención nueva.

El OR es 0,297 y su intervalo de confianza no contiene la unidad, por lo tanto es un OR significativo.

Estimación de riesgo

	Valor	Intervalo de confianza al 95%	
		Inferior	Superior
Razón de las ventajas para Atención (Atención establecida / Atención nueva)	,297	,104	,849
Para la cohorte Infección = No se infecta	,727	,543	,972
Para la cohorte Infección = Se infecta	2,444	1,106	5,403
N de casos válidos	80		

Surge la pregunta si el 0R=0,297 es realmente la medida del riesgo de la atención primaria de producir una infección o es que esta inflada, o es que no es el valor correcto debido a otros factores.

Como única medida de la asociación entre la atención y la infección, se calcula el *odds-ratio* dentro de cada categoría o estrato formado por los dos grupos de edad (menores de 40 y mayores de 40). Una medida única global se obtiene como un promedio ponderado de los *odds-ratio* dentro de los estratos (*odds-ratio* de Mantel-Haenszel).

En [Estadísticos] se selecciona *Riesgo*.
En [Casillas] se selecciona *Porcentaje en columnas*.

Estimación de riesgo

Edad		Valor	Intervalo de confianza al 95% Inferior	Superior
Menor de 40	Razón de las ventajas para Atención (Atención establecida / Atención nueva)	,409	,050	3,367
	Para la cohorte Infección = No se infecta	,893	,659	1,209
	Para la cohorte Infección = Se infecta	2,182	,352	13,539
	N de casos válidos	35		
Mayor de 40	Razón de las ventajas para Atención (Atención establecida / Atención nueva)	,361	,100	1,300
	Para la cohorte Infección = No se infecta	,693	,440	1,091
	Para la cohorte Infección = Se infecta	1,920	,811	4,545
	N de casos válidos	45		

Se calcula el OR de la atención establecida respecto a la atención nueva dentro de cada estrato.

- En el estrato (< 40 años), el OR es 0,41 y no es significativo porque su intervalo de confianza cubre la unidad.
- En el estrato (≥ 40 años), el OR es 0,36 y no es significativo porque su intervalo de confianza cubre la unidad.

La signatura asintótica (p_value) vale 0,921 > 0,05, por lo tanto, no se rechaza la hipótesis nula, que establece que los OR se distribuyen de forma homogénea.

Pruebas de homogeneidad de la razón de las ventajas

	Chi-cuadrado	gl	Sig. asintótica (bilateral)
Breslow-Day	,010	1	,921
De Tarone	,010	1	,921

Si la Signatura asintótica hubiera sido menor que 0,05 no se podría haber aplicado Mantel-Haenszel, teniendo que aplicar otro método (regresión logística).

Al distribuirse los OR de forma homogénea se puede aplicar el estadístico de Mantel-Haenszel:

Estimación de la razón de las ventajas común de Mantel-Haenszel

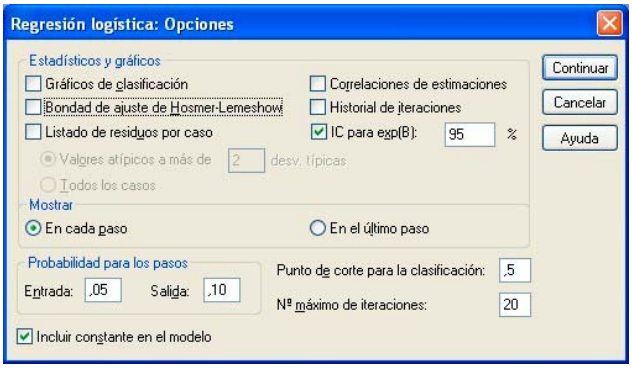
Estimación			,373
ln(estimación)			-,987
Error típ. de ln(estimación)			,558
Sig. asintótica (bilateral)			,077
Intervalo de confianza asintótico al 95%	Razón de ventajas común	Límite inferior	,125
		Límite superior	1,112
	ln(Razón de ventajas común)	Límite inferior	-2,081
		Límite superior	,107

La estimación de la razón de las ventajas común de Mantel-Haenszel se distribuye de manera asintóticamente normal bajo el supuesto de razón de las ventajas común igual a 1,000. Lo mismo ocurre con el log natural de la estimación.

El OR = 0,373, su intervalo de confianza cubre la unidad, por lo que no es significativo. Concluyendo que la edad es un *factor de confusión*.

Adviértase que el OR calculado inicialmente de 0,297 es muy diferente al ajustado con la edad del paciente.

Se realiza la regresión logística: Se selecciona la variable dependiente (*Infección)* y las covariables (variables independientes: Atención y Edad). Ahora tenemos que indicarle al SPSS las variables categóricas, se pulsa el botón [Categóricas].



SPSS ofrece las variables de la ecuación, los coeficientes de regresión con sus correspondientes errores estándar (ET), el valor del *estadístico de Wald* para evaluar la hipótesis nula ($p_i = 0$), la significación estadística asociada, y el valor de la OR= exp(β_i) con sus intervalos de confianza.

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95,0% para EXP(B)	
								Inferior	Superior
Paso 1[a]	Atención(1)	,985	,557	3,124	1	,077	2,678	,898	7,983
	Edad(1)	-1,364	,628	4,711	1	,030	,256	,075	,876
	Constante	-1,078	,469	5,281	1	,022	,340		

a. Variable(s) introducida(s) en el paso 1: Atención, Edad.

Es muy importante distinguir entre un *contexto explicativo* y un *contexto predictivo*. En el primer caso, el modelo para cada posible factor de riesgo o protector se ajusta con los factores que pueden ser confusores para él. Solo en los estudios *predictivos* se ajusta el mejor modelo. Debe tenerse en cuenta, en este caso, que una variable puede tener valor predictivo aunque no sea parte del mecanismo causal que produce el fenómeno en estudio.

Si el contexto es predictivo, la probabilidad del suceso para un perfil de entrada dado ha de computarse independientemente empleando los coeficientes estimados. Si se quiere saber cuál es la probabilidad de que un alumno esté insatisfecho, hay que aplicar el modelo ajustado:

$$P[\text{Infección} = 1] = \frac{1}{1 + e^{(1{,}078 \ + \ 1{,}364 . \text{Edad} \ - \ 0{,}985 . \text{Atención})}}$$

La variable *Atención*, habiendo introducido la variable de confusión *Edad*, tiene un intervalo de confianza que cubre el 1, por lo que no tienen efecto alguno sobre la variable respuesta (*Infección*).

En este sentido, se procede a volver hacer de nuevo una regresión logística binaria, quitando la variable de confusión, con la variable dependiente (*Infección)* y la variable independiente *Atención.* Se elige el Método **Introducir**.

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95,0% para EXP(B)	
								Inferior	Superior
Paso 1[a]	Atención(1)	1,213	,536	5,131	1	,024	3,364	1,178	9,608
	Constante	-1,665	,412	16,318	1	,000	,189		

a. Variable(s) introducida(s) en el paso 1: Atención.

La variable *Atención*, sin introducido la variable de confusión *Edad*, tiene un intervalo de confianza que no cubre el 1, por lo que es significativa sobre la variable respuesta (*Infección*).

El modelo ajustado resulta: $P[\text{Infección} = 1] = \frac{1}{1 + e^{(1{,}665 \ - \ 1{,}213 . \text{Atención})}}$

Ejemplo 2.- Se desea evaluar la satisfacción con la enseñanza pública de 1.027 estudiantes mediante la variable Satisfecho (Si=0, No=1) y tres variables independientes Nacionalidad (España=1, Rumania=2, Colombia=3), Género (Hombre=1, Mujer=2) y Estudios (ESO=1, Primaria=2).

Al introducir los datos en una tabla de contingencia de 4 entradas, ponderando las respectivas frecuencias, se tendrán $(2.3.2.2 = 24$ configuraciones).

			Satisfecho	
Estudios	Género	Nacionalidad	Sí (1)	No (0)
ESO (1)	Hombre (1)	España (1) (00)	54	109
		Rumania (2) (10)	45	90
		Colombia (3) (01)	211	84
	Mujer (2)	España (1) (00)	27	54
		Rumania (2) (10)	20	44
		Colombia (3) (01)	97	42
PRIMARIA (2)	Hombre (1)	España (1) (00)	9	19
		Rumania (2) (10)	2	8
		Colombia (3) (01)	33	6
	Mujer (2)	España (1) (00)	7	14
		Rumania (2) (10)	5	13
		Colombia (3) (01)	21	13

La variable *Nacionalidad* de tipo nominal tiene más de dos categorías, es razonable plantear que sea manejada como una variable *dummy*.

Nacionalidad	Z_1	Z_2
España	0	0
Rumania	1	0
Colombia	0	1

Se ajusta un modelo que incluya una variable nominal con 3 clases, ésta debe ser sustituida por las $(3-1)$ variables *dummy*, y a cada una de ellas corresponderá su respectivo coeficiente.

Debe recordarse que el conjunto de variables *dummy* constituye un todo indisoluble con el cual se suple a una variable nominal. Cualquier decisión que se adopte o valoración que se haga concierne al conjunto íntegro.

ensenanza_publica.sav [Conjunto_de_datos1] - Editor de datos SPSS

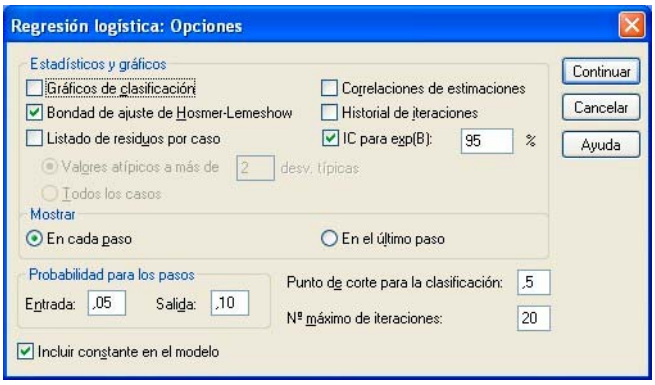
Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?

54 : Visible: 7 de 7 v

	Satisfecho	Nacionalidad	Z1	Z2	Género	Estudios	Frecuencia
1	0	1	0	0	1	1	109
2	0	1	0	0	1	2	19
3	0	1	0	0	2	1	54
4	0	1	0	0	2	2	14
5	0	2	1	0	1	1	90
6	0	2	1	0	1	2	8
7	0	2	1	0	2	1	44
8	0	2	1	0	2	2	13
9	0	3	0	1	1	1	84
10	0	3	0	1	1	2	6
11	0	3	0	1	2	1	42
12	0	3	0	1	2	2	13
13	1	1	0	0	1	1	54
14	1	1	0	0	1	2	9
15	1	1	0	0	2	1	27
16	1	1	0	0	2	2	7
17	1	2	1	0	1	1	45
18	1	2	1	0	1	2	2
19	1	2	1	0	2	1	20
20	1	2	1	0	2	2	5
21	1	3	0	1	1	1	211
22	1	3	0	1	1	2	33
23	1	3	0	1	2	1	97
24	1	3	0	1	2	2	21

Vista de datos | Vista de variables

SPSS El procesador está preparado



Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95,0% para EXP(B) Inferior	Superior
Paso 1[a]	Z1	-,061	,189	,104	1	,747	,941	,649	1,363
	Z2	1,619	,158	104,376	1	,000	5,049	3,701	6,888
	Género(1)	,129	,143	,806	1	,369	1,137	,859	1,507
	Estudios(1)	-,013	,194	,004	1	,947	,987	,676	1,443
	Constante	-,777	,216	12,934	1	,000	,460		

a. Variable(s) introducida(s) en el paso 1: Z1, Z2, Género, Estudios.

Si el contexto es predictivo, la probabilidad del suceso para un perfil de entrada dado ha de computarse independientemente empleando los coeficientes estimados. Si se quiere saber cuál es la probabilidad de que un alumno esté insatisfecho, hay que aplicar el modelo ajustado:

$$P[\text{Insatisfacción}] = \frac{1}{1+e^{(0{,}777\ +\ 0{,}061.Z_1\ -\ 1{,}619.Z_2\ -\ 0{,}129.\text{Género}\ +\ 0{,}013.\text{Estudios})}}$$

- Para una alumna colombiana de primaria, los valores de las variables son: Género=2, Nacionalidad (Z_1=0, Z_2=1), Estudios=2:

$$P[\text{Insatisfacción}] = \frac{1}{1+e^{(0{,}777\ -\ 1{,}619.1\ -\ 0{,}129.2\ +\ 0{,}013.2)}} = 0{,}745$$

- Para un alumno rumano de primaria, los valores de las variables son: Género=1, Nacionalidad (Z_1=1, Z_2=0), Estudios=2:

$$P[\text{Insatisfacción}] = \frac{1}{1+e^{(0{,}777\ +\ 0{,}061.1\ -\ 0{,}129.1\ +\ 0{,}013.2)}} = 0{,}324$$

- Para una alumna española de primaria, los valores de las variables son: Género=2, Nacionalidad (Z_1=0, Z_2=0), Estudios=2:

$$P[\text{Insatisfacción}] = \frac{1}{1+e^{(0{,}777\ -\ 0{,}129.2\ +\ 0{,}013.2)}} = 0{,}367$$

📖 Adviértase que en las variables de la ecuación, por el Método ***Introducir*** (entran todas las variables en el análisis), no se ha analizado el intervalo de confianza (IC) de los coeficientes.

De haberlo hecho, los coeficientes de las variables (Z1, Género y Estudios), respectivamente, tienen un intervalo de confianza que cubre el 1, es decir, hay un riesgo de 1, por lo que debían salir estas variables de la ecuación y volver a realizar el análisis.

En el caso de haber utilizado el Método ***Adelante RV*** (método automático por pasos, hacia delante, que utiliza la prueba de la Razón de Verosimilitud para comprobar las covariables a incluir o excluir), éstas variables hubieran salido de la ecuación:

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95,0% para EXP(B)	
								Inferior	Superior
Paso 1[a]	Z2	1,646	,136	147,008	1	,000	5,185	3,974	6,766
	Constante	-,731	,094	60,938	1	,000	,481		

a. Variable(s) introducida(s) en el paso 1: Z2.

Variables que no están en la ecuación

			Puntuación	gl	Sig.
Paso 1	Variables	Z1	,112	1	,738
		Género(1)	,810	1	,368
		Estudios(1)	,001	1	,977
	Estadísticos globales		,921	3	,820

Se tendría que sacar las variables del análisis y volverlo a realizar con el Método **Introducir**.

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95,0% para EXP(B)	
								Inferior	Superior
Paso 1[a]	Z2	1,646	,136	147,008	1	,000	5,185	3,974	6,766
	Constante	-,731	,094	60,938	1	,000	,481		

a. Variable(s) introducida(s) en el paso 1: Z2.

El modelo ajustado: $P[\text{Insatisfacción}] = \frac{1}{1+e^{(0,731\ -\ 1,646.Z_2)}}$

UAM
Universidad Autónoma de Madrid

Asignatura ..... **Grupo** .....

Apellidos ..... **Nombre** .....

Ejercicio del día .....

UAM
Universidad Autónoma de Madrid

Asignatura ..... Grupo .....

Apellidos ..... Nombre .....

Ejercicio del día .....

Instrumentos Estadísticos Avanzados
Facultad Ciencias Económicas y Empresariales
Departamento de Economía Aplicada
Profesor: Santiago de la Fuente Fernández