REIRE
Revista d'Innovació i Recerca en Educació
www.ub.edu/ice/reire.htm

<Artículo>

Cómo obtener un Modelo de Regresión Logística Binaria con SPSS

Vanesa Berlanga-Silvente y Ruth Vilà-Baños

Fecha de presentación: 18/09/2013
Fecha de publicación: 08/07/2014

//Resumen
Los modelos de regresión logística son modelos estadísticos en los que se desea conocer la relación entre una variable dependiente cualitativa dicotómica (regresión logística binaria o binomial) y una o más variables explicativas independientes, o covariables, ya sean cualitativas o cuantitativas. También es posible analizar una variable dependiente cualitativa con más de dos valores (regresión logística multinomial), aunque en esta ficha nos centraremos en la regresión logística binaria. En cualquier caso, la ecuación inicial del modelo es de tipo exponencial, si bien su transformación logarítmica (logit) permite su uso como una función lineal. El objetivo primordial que resuelve esta técnica es el de modelar cómo influye en la probabilidad de aparición de un suceso, habitualmente dicotómico, la presencia o no de diversos factores y su valor o nivel.
Esta ficha sobre la Regresión Logística Binaria explica las opciones que tiene el programa estadístico SPSS (métodos automáticos "por pasos") y la interpretación de los principales resultados.

//Palabras clave
Clasificar, predicción, regresión logística binaria, método Forward, Wald.

// Referencia recomendada
Berlanga-Silvente, V. y Vilà-Baños, R. (2014). Cómo obtener un Modelo de Regresión Logística Binaria con SPSS. [En línea] *REIRE, Revista d'Innovació i Recerca en Educació*, 7 (2), 105-118. Accesible en: http://www.ub.edu/ice/reire.htm

// Datos de las autoras
Vanesa Berlanga-Silvente. Profesora. Universidad de Barcelona. Departamento de Métodos de Investigación y Diagnóstico en Educación (MIDE). berlanga.silvente@ub.edu

Ruth Vilà-Baños. Profesora. Universidad de Barcelona. Departamento de Métodos de Investigación y Diagnóstico en Educación (MIDE). ruth_vila@ub.edu

REIRE Revista d'Innovació i Recerca en Educació www.ub.edu/ice/reire.htm

1. Introducción

La regresión logística es una técnica multivariante predictiva de regresión. Concretamente, es un modelo que permite asignar a los individuos en una opción de respuesta según los coeficientes estimados para cada una de las variables independientes y la probabilidad de estos en la dependiente. Pretendemos encontrar el mejor modelo para explicar la relación entre una variable dependiente (binaria) y un conjunto de explicativas o covariables (no necesariamente binarias).

Considerando si la variable dependiente tiene dos categorías o más, se distingue entre la regresión logística binaria, o la regresión logística multinomial. Por ejemplo, en una investigación se trabaja con la variable dependiente binaria relativa a que el alumnado tenga éxito o fracaso académico, y queremos averiguar la probabilidad de aprobar o suspender, partiendo de variables independientes como la asistencia y la participación en clase; en este caso propondríamos un modelo de regresión logística binaria.

La regresión logística resulta útil para los casos en los que se desea predecir la presencia o ausencia de una característica o resultado (en el ejemplo, aprobar) según los valores de un conjunto de predictores (en el ejemplo, la asistencia y la participación del alumnado). Es similar a un modelo de regresión lineal, pero está adaptado para modelos en los que la variable dependiente es dicotómica. En efecto, se distingue del modelo de regresión lineal múltiple en el hecho de que en la regresión logística las variables no deben ser necesariamente cuantitativas ni tampoco cumplir supuestos de normalidad. La regresión logística no deja de ser un caso particular del análisis discriminante en el que la variable dependiente tiene dos categorías y partiendo de unos supuestos menos restrictivos, permite introducir variables categóricas como independientes en el modelo. En el análisis discriminante las variables independientes deben cumplir una serie de supuestos de normalidad y de igualdad de varianzas, que en el modelo de regresión logística no son necesarios (Torrado y Berlanga, 2013).

Por sus características, los modelos de regresión logística permiten dos finalidades:

- Cuantificar la importancia de la relación existente entre cada una de las covariables y la variable dependiente, lo que lleva implícito también clarificar la existencia de interacción y confusión entre covariables respecto a la variable dependiente, es decir, conocer la odds ratio para cada covariable.

- Clasificar individuos dentro de las categorías (presente/ausente) de la variable dependiente, según la probabilidad que tenga de pertenecer a una de ellas dada la presencia de determinadas covariables.

2. Condiciones de aplicación del modelo

Cuando se pretende explicar mediante un modelo de regresión el comportamiento de una variable (llamada variable endógena o dependiente) en función de los valores que tomen otras (llamadas variables exógenas o explicativas), suele utilizarse un modelo de regresión lineal múltiple (MRLM o MRLG). Ahora bien, el modelo lineal presenta ciertos problemas serios cuando la variable dependiente es binaria (y, en general, categórica), lo cual nos llevará a usar modelos de regresión no lineales, específicamente pensados para realizar regresión con variables categóricas como es el modelo de regresión logística (Logit). El modelo de regresión logística no presenta condiciones de aplicación restrictivas. Por ello se dice que la regresión logística es más robusta que el análisis discriminante, al requerir menos supuestos (Pérez y Santín, 2007).

En el modelo de regresión logística binaria la variable dependiente debe tomar exactamente dos valores (Sí-No, 0-1, Verdadero-Falso, etc.). Las variables independientes pueden estar a nivel de intervalo o ser categóricas; si son categóricas, deben ser variables *dummy* o estar codificadas como indicadores (existe una opción en el procedimiento para recodificar automáticamente las variables categóricas)[1].

En el caso de que exista alguna variable independiente cuantitativa no es necesario que sigan la Ley Normal, pese a que la solución puede ser más estable si los predictores tienen una distribución normal multivariante.

A pesar de estas condiciones de aplicación poco restrictivas, es muy recomendable que el modelo esté bien especificado teóricamente. Al igual que otras formas de regresión, la multicolinealidad entre los predictores puede sesgar las interpretaciones. Antes de aplicar el modelo, es conveniente analizar la asociación de la variable respuesta con las variables explicativas así como entre las propias variables explicativas para poder interpretar los coeficientes obtenidos de forma adecuada. Este análisis previo fundamentará la elección de las variables independientes que se añaden en el modelo (Johnson, 2000).

En resumen, el modelo debe ser el más reducido que explique los datos (principio de parsimonia), y que además sea congruente e interpretable. Hay que tener en cuenta que un mayor número de variables en el modelo implicará mayores errores estándar. Es en esta etapa cuando aparece un primer punto de debate entre docentes: ¿el conocimiento va parejo al desarrollo de la competencia? Sin despreciar el debate, la postura es clara. No se debería dar por supuesto, en las valoraciones competenciales, el nivel de conocimientos, ya que no puede ser considerado como una competencia única o fundamental al margen de las demás. Todas las competencias relevantes se complementan y definen un conjunto de necesidades formativas. El conocimiento de la materia es básico; no basta con tenerlo –saber-, sino que se debe ser competente en el mismo –hacer o saber hacer-. Esto implica actualizar los conocimientos y técnicas, así como también otros elementos que las condicionan, como los cambios sociales e institucionales en los que se desarrollan. Un segundo paso está en la medición del grado de

[1] Si la covariable cualitativa tuviera más de dos categorías, para su inclusión en el modelo debería transformarse en varias covariables cualitativas dicotómicas ficticias o de diseño (las llamadas variables dummy), de forma que una de las categorías se tomaría como categoría de referencia. Con ello cada categoría entraría en el modelo de forma individual. En general, si la covariable cualitativa posee n categorías, habrá que realizar n-1covariables ficticias.

adquisición de competencias. Existen diferencias entre "estar capacitado" para hacer algo y "ser capaz de" realizar las acciones adecuadas.

3. Los coeficientes del modelo

La regresión logística consiste en obtener una función lineal de las variables independientes que permita clasificar a los individuos en una de las dos subpoblaciones o grupos por los dos valores de la variable dependiente. Un modelo de regresión logística es un modelo que permite estudiar si dicha variable binaria depende de otra/s variable/s. Consecuentemente, la función lineal es el logaritmo de la figura 1 donde β es constante y *xk* las variables independientes, dando lugar al *modelo logístico múltiple* (Pérez, 2004).

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

" β_0" y "β_k" son los coeficientes estimados a partir de los datos

X_k son las variables independientes

Figura 1. Función lineal del modelo de regresión logística.

A partir de los coeficientes de regresión (β) de las variables independientes introducidas en el modelo se puede obtener directamente la odds ratio[2] de cada una de ellas, que corresponde al riesgo de tener el resultado o efecto evaluado para un determinado valor (x) respecto al valor disminuido en una unidad (x-1). En otras palabras, si la variable independiente es una variable cuantitativa, la OR que se obtiene representa la probabilidad del evento predicho que tiene un individuo con un valor x frente a la probabilidad que tiene un individuo con valor (x-1). Por ejemplo, si X es la variable EDAD (en años cumplidos) y estamos prediciendo el éxito académico, la OR será la probabilidad de éxito académico que tiene un individuo de 21 años en relación con la que tiene uno de 20 años. Si la variable independiente es cualitativa, la regresión logística solo admite categorías dicotómicas, de manera que la OR es el riesgo de los sujetos con un valor frente al riesgo de los sujetos con el otro valor para esa variable.

El modelo logístico se puede escribir de otras formas equivalentes que permiten calcular directamente la probabilidad del proceso binomial para los distintos valores de las variables incluidas en el modelo (X). Se la denomina *función logística* y se presenta en la figura 2. Calcula la probabilidad de que un individuo pertenezca a una subpoblación (en el ejemplo anterior, el *aprobado*). Si la probabilidad es ≥0,5 el individuo será clasificado en la segunda categoría (aprobado), en caso contrario será clasificado en la primera (suspendido).

[2] OR = e^{β} , siendo el número "e" la base de los logaritmos neperianos, una constante cuyo valor es 2,718.

$$p = \frac{1}{1+e^{-Y}} = \frac{1}{1+e^{-(\beta_0 + X_1\beta_1 + ... + X_k\beta_k)}}$$

"Y" es la función lineal del modelo de regresión logística
"e" es la base de logaritmos neperianos (2,718)

Figura 2. Función logística para el cálculo de la probabilidad de que un sujeto pertenezca a una de las dos categorías binarias de la variable dependiente.

En la regresión logística la estimación de parámetros se lleva a cabo a través del método de máxima verosimilitud, de modo que los coeficientes que estima el modelo hacen nuestros datos "más verosímiles" (Visauta, 1998).

4. Métodos en el modelo de regresión logística

Una vez se dispone de un modelo inicial (teórico o no) debe procederse a su reducción hasta obtener el modelo más reducido que siga explicando los datos. Para ello se puede recurrir a métodos de selección paso a paso, bien mediante inclusión "hacia adelante", bien por eliminación "hacia atrás"; o a la selección de variables por mejores subconjuntos de covariables. Estos métodos se encuentran implementados en numerosos paquetes estadísticos como el programa SPSS, por lo que son muy populares. Dado que para la comprensión de los métodos de selección paso a paso se requiere un conocimiento previo acerca del ajuste del modelo, este es un aspecto que se tratará más adelante.

Hay tres opciones principales para elegir el método que nos ayude a seleccionar las variables en el modelo. La selección del método permite especificar cómo se introducen las variables independientes en el análisis. Utilizando distintos métodos se pueden construir diversos modelos de regresión a partir del mismo conjunto de variables:

1. El método "*Introducir*": permite al investigador tomar el mando, decidir qué variables se introducen o extraen del modelo.

2. El método "*Adelante*": es uno de los métodos automáticos o por pasos que deja que el programa vaya introduciendo variables en el modelo, empezando por aquellas que tienen coeficientes de regresión más grandes, estadísticamente significativos. En cada paso se reevalúan los coeficientes y su significación, con lo cual se pueden eliminar del modelo aquellos que no se consideran estadísticamente significativos.

3. El método "*Atrás*": al igual que el anterior es otro de los métodos automáticos. En este caso se parte de un modelo con todas las covariables que se hayan seleccionado en el cuadro de diálogo, y se van eliminando del modelo aquellas sin significación estadística.

En los métodos por pasos (Adelante y Atrás) el programa SPSS permite las opciones de elegir entre 3 criterios o estadísticos (razón de verosimilitud RV, condicional o Wald) para adoptar "decisiones estadísticas" con el objetivo de comprobar la significación estadística de cada uno de los coeficientes de regresión en el modelo:

Selección hacia adelante (Condicional). Método de selección por pasos que contrasta la entrada según la significación del estadístico de puntuación y contrasta la eliminación basándose en la probabilidad de un estadístico de la razón de verosimilitud que se fundamenta en estimaciones condicionales de los parámetros.

Selección hacia adelante (Razón de verosimilitud). Método de selección por pasos hacia adelante que contrasta la entrada basándose en la significación del estadístico de puntuación y contrasta la eliminación según la probabilidad del estadístico de la razón de verosimilitud, que se sustenta en estimaciones de la máxima verosimilitud parcial.

Selección hacia adelante (Wald). Método de selección por pasos hacia adelante que contrasta la entrada basándose en la significación del estadístico de puntuación y contrasta la eliminación según la probabilidad del estadístico de Wald.

Eliminación hacia atrás (Condicional). Selección hacia atrás por pasos. El contraste para la eliminación se basa en la probabilidad del estadístico de la razón de verosimilitud, el cual se fundamenta en las estimaciones condicionales de los parámetros.

5. Procedimiento de regresión logística con SPSS

Para acompañar la creación del modelo de regresión logística y concretarlo en el programa SPSS, proponemos el siguiente caso práctico:

> *Queremos predecir el éxito/fracaso académico (categorizando la variable dependiente en aprobado y suspenso), partiendo de variables como la asistencia, la participación en clase y la dedicación laboral para detectar las variables que mejor permiten predecir el éxito/fracaso académico.*

Se trata de obtener una combinación lineal de las variables independientes *Asistencia*, *Participación en clase* y *Dedicación laboral* que permita estimar la probabilidad de pertenecer a cada uno de los dos grupos establecidos por los valores de la variable dependiente *Éxito* (aprobado o suspendido). Para ello, hemos de dar respuesta a la siguiente pregunta: ¿Está relacionada la asistencia a clase, la participación o la dedicación laboral con el hecho de aprobar una asignatura?

Por tanto, las variables estudiadas son las siguientes:

- *Rendimiento académico*: calificación obtenida en el examen final de febrero en la asignatura estudiada.

- *Asistencia a clase:* valoración cualitativa que establece el profesor de la asistencia de los alumnos a clase.
- *Participación en clase:* valoración cualitativa que determina el profesor de la participación del alumnado en clase.
- *Dedicación laboral:* si trabaja o no trabaja.

Es importante que antes de empezar revisemos los requisitos y etapas de la regresión logística:

- Recodificar las variables independientes categóricas u ordinales en variables ficticias o simuladas y de las variables dependientes en 0 y 1.
- Evaluar efectos de confusión y de interacción del modelo explicativo.
- Evaluar la bondad de ajuste de los modelos.
- Analizar la fuerza, sentido y significación de los coeficientes, sus exponenciales y estadísticos de prueba.

A continuación se ejemplifican estas etapas con los datos propuestos.

Para ejecutar el análisis de regresiónlogística se seleccionan los menús: *Analizar/Regresión/Logística binaria* del SPSS (figura 3).

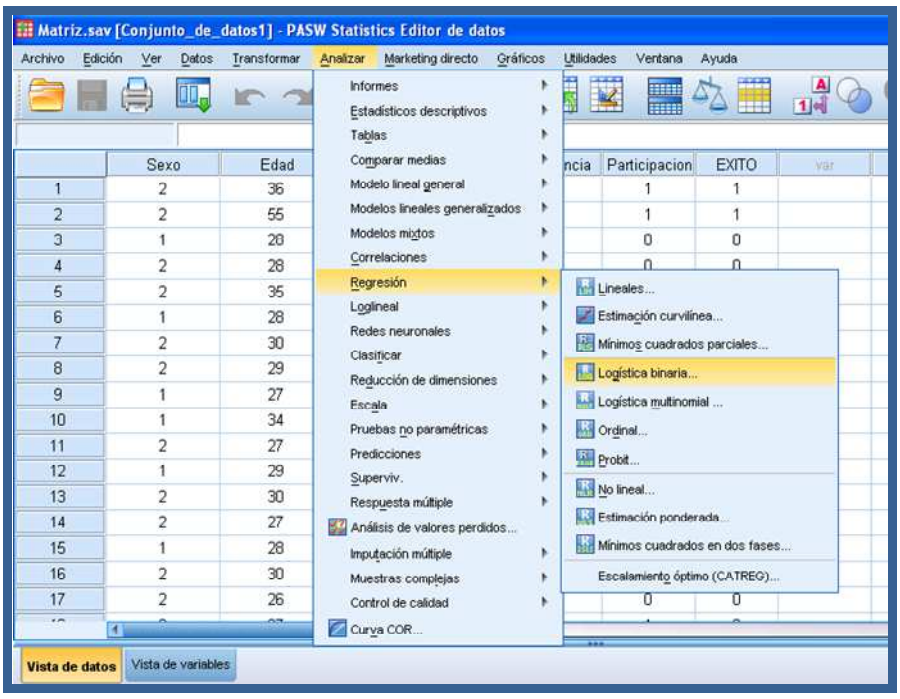

Figura 3. Regresión logística binaria en SPSS.

A continuación, aparece el cuadro de diálogo principal (figura 4) donde debemos introducir la variable dependiente (en el ejemplo es la variable dicotómica de éxito o fracaso), y las variables independientes covariables (en el ejemplo: *la asistencia, la participación y el trabajo*).

Figura 4. Cuadro de diálogo principal de la regresión logística binaria en SPSS.

En el ejemplo, 0 significa fracaso y 1 éxito. De contener otros valores distintos a 0 y 1, se recodifica automáticamente en binario (0-1) la variable dependiente. Conocer qué significa 0 y qué significa 1 es fundamental para la posterior interpretación de los resultados. Entre los resultados se obtendrá la probabilidad para cada opción, de modo que si esta tiende a 0 será más probable suspender, y si tiende a 1 la tendencia será a aprobar (según el ejemplo).

Método de selección de variables (figura 4: a)

Para realizar nuestro caso práctico, el método que utilizaremos para seleccionar el subconjunto de variables será el *FORWARD* o *Wald hacia delante*, método automático por pasos, hacia adelante que utilizará los estadísticos la *Puntuación eficiente de Raoy el estadístico de Wald* para comprobar las covariables que deben incluirse o excluirse. Uno de las cuestiones más importantes a la hora de encontrar el modelo de ajuste más adecuado para explicar la variabilidad de una característica cuantitativa es la correcta especificación del llamado modelo teórico. La ventaja del método "Adelante" es que el investigador no decide que variables se introducen o extraen del modelo, ya que se comienza por un modelo que no contiene ninguna variable explicativa.

Los métodos automáticos "por pasos" son adecuados para obtener diferentes modelos, con una finalidad predictiva, que pueden dar una idea al investigador de aquellos más parsimoniosos. Debe tenerse en cuenta que estos procedimientos automáticos en SPSS no incorporan el principio jerárquico.

Variable de selección (figura 4: b)

Una opción interesante es la de *variable de selección* del cuadro de diálogo principal. Sirve para seleccionar casos específicos para el análisis. Se elige una variable de selección y se pulsa la opción *Regla* y se incluirían en la estimación del modelo los casos definidos por la regla de

selección. En el ejemplo, podría ser importante realizar el análisis únicamente con estudiantes entre 20 y 30 años de edad.

Variables categóricas (figura 4: c)

Por defecto las variables independientes que se introducen en el modelo se definen como categóricas si están en *cadena*, y cuantitativas si están definidas como *numéricas*. En muchas ocasiones, las variables categóricas son numéricas, cuyos valores numéricos corresponden a categorías. Para ello, se dispone del botón del cuadro de diálogo *Categórica*, donde se listan las covariables identificadas como categóricas.

En el ejemplo, las tres variables son categóricas y necesariamente las introducimos en el cuadro de diálogo de la figura 5. Dejamos las opciones por defecto y automáticamente el sistema recodifica las variables en categóricas. Si son dicotómicas (como en el ejemplo) simplemente se codifican en 1 y 0 (variables *dummy*). Es decir, que para cada categoría de la variable se calcula una variable *dummy*. En caso de tener variables politómicas[3], el sistema crea tantas variables como categorías tenga cada variable (menos 1, que es la original), asignando de nuevo 0-1 a cada nueva variable.

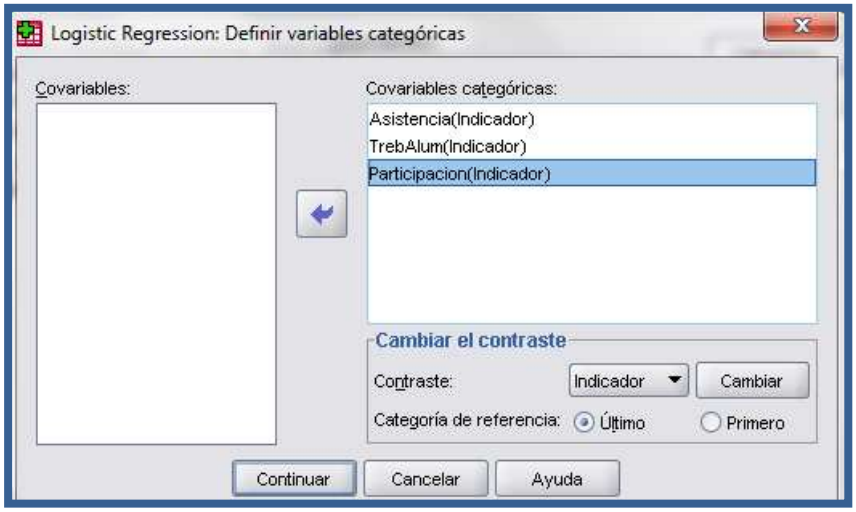

Figura 5. Subcuadro de diálogo del botón Categóricas.

La asignación de coeficientes para cada categoría se hace de acuerdo con la categoría de referencia y con el tipo de contraste (ver figura 5), que por defecto es la última y el contraste indicador, cuyo contraste indica la presencia o ausencia de la pertenencia a una categoría que se representa en la matriz de contraste como una fila de ceros.

[3] Se transforman en un número (c-1) de variables *dummy*, siendo "c" el número de valores o de categorías distintas de dicha variable, haciéndolo automáticamente el programa SPSS.

Opciones (figura 4: d)

El botón opciones del cuadro de diálogo principal ofrece las posibilidades de análisis que se reflejan en la figura 6. Recomendamos dejar las que se ofrecen por defecto y añadir las que se indican en la figura.

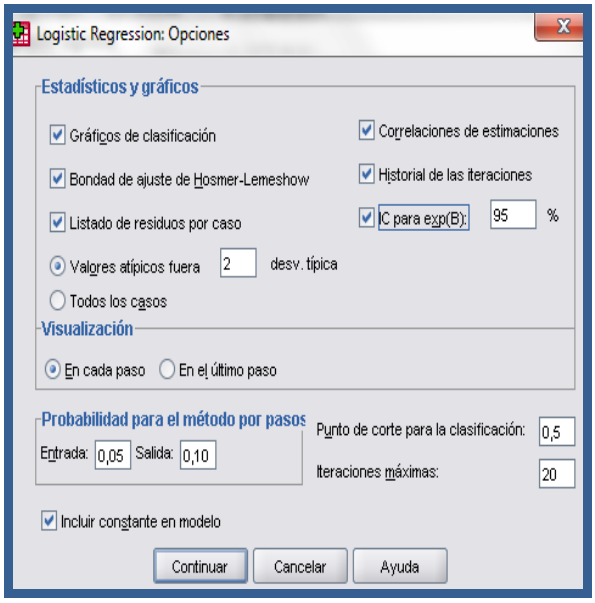

Figura 6. Subcuadro de diálogo del botón Opciones.

A modo resumen, se pueden especificar algunas opciones para el análisis de regresión logística:

- *Estadísticos y gráficos*. Permite solicitar estadísticos y gráficos. Las opciones disponibles son Gráficos de clasificación, Bondad de ajuste de Hosmer-Lemeshow, Listado de residuos por caso, Correlaciones de estimaciones, Historial de iteraciones e IC para exp(B). Asimismo, se pueden seleccionar para mostrar los estadísticos y los gráficos en cada paso o bien solo para el modelo final, en el último paso. El estadístico "Bondad de ajuste de Hosmer-Lemeshow" es más robusto que el estadístico de bondad de ajuste tradicionalmente utilizado en la regresión logística, especialmente para los modelos con covariables continuas y los estudios con tamaños de muestra pequeños.

- *Probabilidad para el método por pasos*. Permite controlar los criterios por los cuales las variables se introducen y se eliminan de la ecuación. Puede especificar criterios para la entrada o para la salida de variables. Una variable se introduce en el modelo si la probabilidad de su estadístico de puntuación es menor que el valor de entrada, y se elimina si la probabilidad es mayor que el valor de salida.

- *Punto de corte para la clasificación.* Permite determinar el punto de corte para la clasificación de los casos. Los casos con valores pronosticados que han sobrepasado el punto de corte para la clasificación se clasifican como positivos, mientras que aquellos con valores pronosticados menores que el punto de corte se clasifican como negativos.

- *Número máximo de iteraciones*. Permite cambiar el número máximo de veces que el modelo itera antes de finalizar.

- *Incluir constante en el modelo*. Permite indicar si el modelo debe incluir un término constante. Si se desactiva, el término constante será igual a 0.

Generar nuevas variables (figura 4: e)

Una opción interesante se encuentra en el botón *Guardar* del cuadro de diálogo principal. Una vez estimado el modelo de regresión logística resulta de gran interés analizar diversos tipos de residuales, estadísticos de influencia, etc. como herramientas útiles para identificar aquellos puntos en los que el modelo no ajusta correctamente. Estos estadísticos se encuentran en el cuadro de diálogo de la opción *Guardar* tal como se muestra en la figura 7.

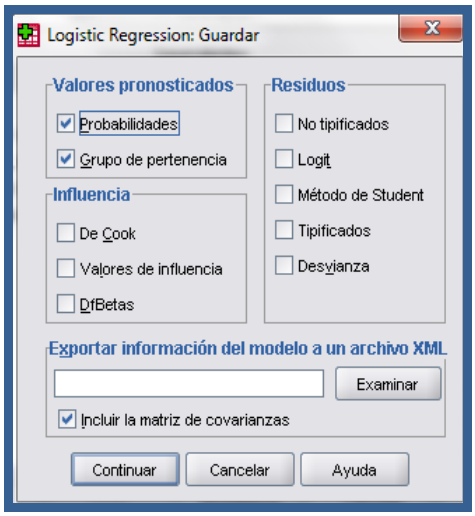

Figura 7. Subcuadro de diálogo del botón Guardar.

Concretamente recomendamos señalar las opciones *Probabilidades* y *Grupo de pertenencia* para que se generen dos nuevas variables en la base de datos: una con la probabilidad de ocurrencia de cada caso predicha por el modelo y otra con el grupo al que se asigna cada sujeto de acuerdo con su probabilidad, según pronostica el modelo.

Otras opciones del mismo cuadro de diálogo son los *Residuos* y los *Estadísticos de influencia*. Los primeros generan la diferencia entre las probabilidades observadas y las predicciones (en escala logit, con la estimación de la desviación estándar), los segundos miden la influencia de cada caso en los residuales y en las predicciones respectivamente (Visauta, 1998), es decir, guarda los valores de los estadísticos que miden la influencias de los casos sobre los valores pronosticados.

Todas estas opciones generan nuevas variables con las que pueden realizarse todo tipo de análisis y gráficos.

6. Cálculos e interpretación del modelo de regresión logística

Entre los *outputs* que se generan es importante fijarse en los parámetros estimados por el modelo de regresión logística. En la figura 8 se presentan los coeficientes de regresión con sus correspondientes errores estándar (E.T.), el valor del estadístico de Wald para evaluar la hipótesis nula (Bi=0), la significación estadística asociada y el valor de la OR (Exp(B)).

Bloque 1: Método = Por pasos hacia adelante (Wald)

Variables en la ecuación[c]

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1[a]	Asistencia	2.626	.254	106.511	1	.000	13.812
	Constante	-.754	.162	21.636	1	.000	.471
Paso 2[b]	Asistencia	21.382	,110	.000	1	,000	1.932E9
	Participacion	20.907	,125	.000	1	,000	1.202E9
	Constante	-20.688	,140	,000	1	,000	,000

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

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

c. Se ha detenido un procedimiento por pasos ya que al eliminar la variable menos significativa se obtuvo un modelo previamente ajustado.

Figura 8. Parámetros estimados por el modelo de regresión logística.

Por lo tanto, hemos de interpretar que en nuestra ecuación de regresión solo aparecen las variables *Asistencia* y *Participación en clase* más la constante, y queda fuera la variable *Dedicación laboral*.

El método Forward o Wald en un primer paso ha seleccionado la variable *Asistencia* mediante el estadístico "Puntuación eficiente de Rao". Posteriormente, continuando con el proceso de selección de variables, de entre las restantes variables independientes, la candidata a ser seleccionada ha sido *Participación*, seguida de *Dedicación laboral*. En el siguiente paso mediante el estadístico Wald la candidata a ser eliminada de las tres variables seleccionadas en el modelo ha sido *Dedicación Laboral*.

En resumen, se han realizado tres etapas a través de las cuales se han ido seleccionando las 3 variables y eliminando aquella que no era significativa, es decir, en la última etapa se ha repetido todo el proceso hasta que ninguna variable no seleccionada ha satisfecho el criterio de selección y ninguna de las seleccionadas ha satisfecho el de eliminación.

Con estos datos (figura 8) podemos construir la ecuación de regresión logística, que en nuestro ejemplo sería:

$$Y = -20{,}68 + 21{,}38\ (asistencia) + 20{,}90\ (participación)$$

Esta ecuación logística nos permite calcular la probabilidad de éxito o fracaso para un alumno que asiste a clase y que participa:

$$Y = -20{,}68 + 21{,}38\,(1) + 20{,}90\,(1) = 21{,}6$$

$$p(aprobar) = \frac{1}{1+e^{-Y}} = \frac{1}{1+e^{-21{,}6}} = 0{,}99$$

Tal como hemos visto antes, 0 es fracaso y 1 éxito académico, la probabilidad de un estudiante que asista a clase y participe, el modelo lo clasificará como aprobado (éxito), dado que el punto de corte está en 0,5. Estos dos datos los genera SPSS como variables nuevas para cada uno de los sujetos. El número de aciertos en esta clasificación es fundamental para la bondad de ajuste del modelo.

A partir de los datos de la tabla de clasificación y con un riesgo α=0,05 podemos concluir que, en términos generales, de un total de 400 estudiantes, 325 han sido clasificados correctamente, o en otras palabras, el 81,3% (figura 9) ha sido correctamente clasificado como aprobado o suspendido. Para valorar la capacidad predictiva del modelo se calcularon los valores de sensibilidad y especificidad (Ferrán, 2001). Podemos comprobar que nuestro modelo tiene una especificidad alta (100%) y una sensibilidad baja (49,7%), por lo que el modelo clasifica adecuadamente a los estudiantes que aprueban y deficientemente a las estudiantes que suspenden, esto pudiera estar relacionado con la distribución de la muestra para esta variable.

Tabla de clasificación[a]

Observado			Pronosticado		
			Éxito o Fracaso académico		Porcentaje correcto
			SUSPENDIDO	APROBADO	
Paso 1	Éxito o Fracaso académico	SUSPENDIDO	119	30	79,9
		APROBADO	56	195	77,7
	Porcentaje global				78,5
Paso 2	Éxito o Fracaso académico	SUSPENDIDO	74	75	49,7
		APROBADO	0	251	100,0
	Porcentaje global				81,3

a. El valor de corte es ,500

Figura 9. Tabla de clasificación para el Modelo de Regresión Logística.

<Referencias bibliográficas>

Ferrán, A., M. (2001). *Spss para windows – Análisis estadístico*. España: McGraw Hill/Interamericana.

Johnson, D., E. (2000). *Métodos multivariados aplicados al análisis de datos*. México: Thompson.

Pérez, C. (2004). *Técnicas de análisis multivariante de datos. Aplicaciones con SPSS*. Madrid: Pearson educación.

Pérez, C. y Santín, D. (2007). *Minería de Datos: Técnicas y Herramientas.* Madrid: Ediciones Paraninfo, S.A.

Torrado, M. y Berlanga, V. (2013). Análisis Discriminante mediante SPSS. [En línea] *REIRE, Revista d'Innovació i Recerca en Educació*, 6 (2), 150-166. Consultado el 22 de febrero de 2014en http://www.ub.edu/ice/reire.htm .

Visauta, B. (1998). *Análisis estadístico con SPSS para Windows. Estadística multivariante*. Madrid: McGrawHill.