DOCUWEB FABIS

Dot. Núm 0702012

Cómo hacer una Regresión Logística con SPSS© "paso a paso". (I)

Aguayo Canela, Mariano.

Servicio de Medicina Interna. Hospital Universitario Virgen Macarena. Sevilla

Resumen

En este primer documento sobre la Regresión Logística (Binaria) se aportan los conceptos básicos teóricos para llevarla a cabo, junto con recomendaciones elementales para una correcta aplicación del análisis, y luego se explican detalladamente las opciones que tiene el programa estadístico SPSS y la interpretación de los principales resultados.

0. Introducción.

0.1. RECORDATORIO TEÓRICO.

Cuando tengamos una variable dependiente dicotómica (0/1; SI/NO; VIVO/MUERTO; CURADO/NO-CURADO, HIPERTENSIÓN/NORMOTENSIÓN, etc.) que deseemos predecir, o para la que queramos evaluar la asociación o relación con otras (más de una) variables independientes y de control, el procedimiento a realizar es una REGRESIÓN LOGÍSTICA (RL) BINARIA MULTIVARIANTE.[1]

La Regresión Logística es probablemente el tipo de análisis multivariante más empleado en Ciencias de la Vida. Las razones más poderosas son:

1. Permite introducir **como variables predictoras** de la respuesta (efecto o v. dependiente) una mezcla de **variables categóricas y cuantitativas**.

2. A partir de los coeficientes de regresión (β) de las variables independientes introducidas en el modelo **se puede obtener directamente la OR** de cada una de ellas,[2] 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).

   - Así, si la variable independiente es una variable cuantitativa, la OR que se obtiene representa la probabilidad del evento predicho que tiene un individuo con

[1] También podría llevarse a cabo un análisis discriminante, que permite –al igual que la RL- clasificar a los individuos, pero requiere el cumplimiento de dos supuestos: las p variables independientes deben seguir una distribución Normal multivariante, y las matrices de varianzas-covarianzas de las p variables independientes en cada grupo deben ser iguales. Por ello se dice que la RL es más robusta que el análisis discriminante, al requerir menos supuestos.

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

Correspondencia: marianoaguayo@telefonica.net

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 muerte, la OR será la probabilidad de muerte que tiene, por ejemplo, un individuo de 40 años en relación a la que tiene uno de 39 años.[3]

- Si la variable independiente es cualitativa, la RL sólo admite categóricas 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.

3. En la RL **la variable dependiente (la que se desea modelizar, Y) es categórica**, habitualmente dicotómica (RL binaria), lo que constituye una circunstancia muy frecuente y simple de representar fenómenos en la naturaleza y en ciencias de la vida: SI/NO, PRESENTE/AUSENTE, etc. Esto hace a este tipo de análisis el ideal para aplicar en los estudios de casos y controles, estudios en los que los casos tienen algo (habitualmente una enfermedad, un efecto o un desenlace) y los controles no.

4. Lo que se pretende mediante la RL es **expresar la probabilidad de que ocurra el evento en cuestión como función de ciertas variables**, que se presumen relevantes o influyentes. Si ese hecho que queremos modelizar o predecir lo representamos por Y (la variable dependiente), y las *k* variables explicativas (independientes y de control) se designan por X_1, X_2, X_3,…,X_k, la ecuación general (o ***función logística***) es:

$$P(Y=1) = \frac{1}{1 + exp\,(-\alpha - \beta_1 X_1 - \beta_2 X_2 - \beta_3 X_3 - \ldots - \beta_K X_k)}$$

donde α, β_1, β_2, β_3,…, β_k son los parámetros del modelo, y ***exp*** denota la función exponencial. Esta función exponencial es una expresión simplificada que corresponde a elevar el número ***e*** a la potencia contenida dentro del paréntesis, siendo ***e*** el número o constante de Euler, o base de los logaritmos neperianos (cuyo valor aproximado a la milésima es 2,718).

0.2. ANTES DE NADA…

Antes de ponerse a hacer regresión logística "*a lo loco*", es recomendable tener en cuenta ciertos detalles. Por eso recomendamos **establecer claramente lo siguiente**:

1. **Cuáles podrían ser variables realmente *predictoras* (independientes)**[4] de

[3] El modelo de RL asume que la distancia entre cada valor de la variable independiente es igual y que el cambio que se produce en la variable respuesta es constante en cada modificación unitaria de la variable independiente.

[4] Normalmente hay una (o unas pocas) variable independiente que es la que se desea evaluar, comprender o modelizar su papel sobre el efecto, analizando su relación o asociación con la variable dependiente, y debe formar parte de la hipótesis principal del estudio analítico. Por ejemplo: *¿Está relacionado el hábito materno de fumar durante el embarazo con el hecho de tener un recién nacido de bajo peso?*

la respuesta (dependiente). Esto lo da el conocimiento del tema y la revisión de la literatura.

2. **Cuáles podrían ser variables *confundentes***,[5] que será necesario ajustar o controlar, ya que, de lo contrario, la evaluación de la relación principal (X → Y) podría ser espúrea o artefactada. Esto lo da el conocimiento del tema y la revisión de la literatura, por lo que deben recogerse (e incluirse en el análisis) aquellas variables predictoras que otros estudios hayan reconocido como tales. El análisis estratificado y el análisis multivariante serán las estrategias en esta fase de análisis para corregir su efecto, procedimientos que permiten el ajuste o control.

3. **Cuáles podrían ser variables *modificadoras de efecto o de interacción***,[6] que producen cambios en la relación principal evaluada (X → Y) en términos de incrementarla o disminuirla. Esto también lo da el conocimiento del tema y la revisión de la literatura, aunque en ocasiones es un descubrimiento del investigador, debiendo entonces incorporarse a las conclusiones del estudio. El análisis estratificado, evaluando la relación principal en los diferentes estratos de la variable presumiblemente *modificadora de efecto*, y el análisis multivariante, incluyendo términos multiplicativos (X•M) con la variable independiente (X) y la variable *modificadora de efecto* (M), son los procedimientos estadísticos para detectar su presencia y explicar su comportamiento.

4. **Qué sentido tiene nuestro análisis**, diferenciando dos grandes objetivos:

    a. ***Predecir una determinada respuesta a partir de las variables predictoras o independientes***, obteniendo una fórmula matemática que sirva para calcular la probabilidad del suceso estudiado en un nuevo individuo en razón de los valores que presente de las diferentes variables incluidas en el modelo. Bajo esta óptica, debemos buscar, entre todos los posibles modelos, el más parsimonioso, que es el que con el menor número de variables posibles (independientes y de control) genera una predicción más precisa y válida de la respuesta evaluada. Recuerde que introducir variables poco relevantes tiende a enmascarar el proceso de modelado y puede llevar a estimaciones no válidas. Por otra parte, intentar construir un modelo con muchas variables puede ser un problema cuando hay pocas observaciones, ocasionando estimaciones inestables y poco precisas.

    b. ***Calcular los riesgos ajustados o controlados*** (no sesgados) para

[5] Las variables de confusión son variables predictoras de la respuesta o efecto, externas a la relación principal que se analizan (no son un mero paso intermedio entre la exposición y la respuesta), y simultáneamente relacionadas con la variable independiente. Su presencia genera un sesgo o error al evaluar la relación entre la variables independiente (X) y dependiente (Y).

[6] Una variable modificadora de efecto es una característica de la relación entre el factor de estudio (exposición) y el efecto (resultado), y puede aportar datos interesantes sobre los mecanismos etiopatogénicos o causales. De detectarse este efecto, debe mostrarse, no controlarse.

cada variable independiente.[7] En este caso es importante determinar el conjunto de variables que será oportuno controlar en el análisis, incluyendo aquellas que tengan una adecuada justificación teórica. Los pasos a seguir serían:

1. **Valorar si hay interacción** (modificando el efecto) entre alguna de las variables de control y la variable independiente, con pruebas de significación estadística, dejando en el modelo los términos de interacción que sean estadísticamente significativos.

2. **Valorar si hay confusión** entre alguna de las variables de control y la relación principal evaluada, sin aplicar pruebas de significación estadística. En esta situación lo que debe analizarse es si la introducción de una variable de control en el modelo de RL produce un cambio clínicamente importante en la medida de asociación que estima el efecto de la exposición (X) sobre la respuesta (Y).[8] Si no es así dicha variable de control debe ser eliminada del modelo, pues de dejarla en él es posible que disminuya la precisión del estudio, sin aportar ajuste (sobreajuste).

3. Si al final del proceso hay más de un subconjunto de variables de control que ofrecen un similar grado de ajuste, se deberá elegir el que estime con mayor precisión el efecto principal evaluado ($X \rightarrow Y$) en la investigación.[9]

0.3. QUÉ RECOMENDAMOS...

1. Adoptar el **PRINCIPIO JERARQUICO**, que viene a decir que si en el modelo de RL se incluye un término cualquiera, todos sus términos de menor orden deben permanecer en el modelo, y que si se elimina del modelo un término cualquiera, todos los términos de mayor orden en los que intervenga también deben sacarse del modelo. Así por ejemplo, si el término de interacción $X \bullet X_1 \bullet X_2$ se incluye en el modelo logístico, sus términos de menor orden ($X \bullet X_1$, $X \bullet X_2$, $X_1 \bullet X_2$, X, X_1 y X_2) deben permanecer en la ecuación; y si se elimina del modelo, por ejemplo, la variable X_2, los términos de interacción que la contengan ($X \bullet X_1 \bullet X_2$, $X \bullet X_2$, $X_1 \bullet X_2$,...) deben sacarse de la ecuación.

---

[7] Esta es una estrategia muy común en investigación no experimental, donde la estimación del efecto (Y) debe realizarse ajustando o controlando las variables de control (o factores *confundentes*), ya que los datos se han recogido sin asignación aleatoria de los sujetos a los diferentes niveles de exposición (X).

[8] Algunos autores proponen como regla que un cambio clínicamente relevante en la OR debe ser de al menos un 10%, y preferiblemente de un 20%. Esto debe tomarse con carácter orientativo, siendo el investigador quien finalmente decida sobre este aspecto.

[9] El modelo más preciso será el que muestre un menor error estándar en el coeficiente de regresión de X, o el que aporte un intervalo de confianza de la OR asociada a la exposición X más estrecho.

MUY IMPORTANTE: Para llevar a cabo un ajuste estadístico siguiendo los tres pasos que se han descrito en el apartado anterior (punto 4.b) bajo el principio jerárquico NO SE PUEDEN UTILIZAR LOS PROCEDIMIENTOS AUTOMÁTICOS DEL PROGRAMA SPSS (hacia delante –*forward*- o hacia atrás –*backward*-), ya que estos no incorporan la norma jerárquica, y eliminan del modelo los términos no significativos, dejando los estadísticamente significativos (coeficientes de regresión no nulos). Por ello, en el análisis de regresión con objetivo de ajuste o control de la confusión debe recurrirse al PROCEDIMIENTO ***INTRODUCIR***, que permite al investigador conducir el análisis en función de los resultados que va obteniendo.

2. Puede ser oportuno llevar a cabo, antes de entrar de lleno en la RL multivariante, hacer un análisis bivariante, esto es, analizar las relaciones de la variable dependiente con cada una de las variables independientes, *modificadoras de efecto* y *confundentes*, tomadas "una a una".[10]

    a. Si la variable independiente es una categórica, el contraste será a través de una Chi cuadrado. Entonces…
        i. Evalúe la fuerza de asociación mediante la OR ó el RR
        ii. Analice la precisión del análisis mediante los intervalos de confianza de las medidas de asociación (OR ó RR)
        iii. Compruebe la significación estadística del contraste asociada al estadístico

    b. Si es una variable cuantitativa, el contraste será un ANOVA o una t de Student, para comprobar si las medias son diferentes en los grupos que establece la variable dependiente. Entonces…
        i. Evalúe la fuerza de asociación mediante la diferencia de medias
        ii. Analice la precisión del análisis mediante los intervalos de confianza de la diferencia de medias
        iii. Compruebe la significación estadística del contraste asociada al estadístico

3. Suplementariamente, se recomienda llevar a cabo análisis estratificados con aquellas variables que pensemos pueden ser *confundentes* o *modificadoras de efecto*, comparando las OR de la relación principal evaluada en cada estrato de la variable[11] presumiblemente *confundente o modificadora de efecto.*

4. Seguidamente haga una Regresión Logística Simple (o *univariante*), entrando cada vez en el modelo (con el método INTRODUCIR) una de las variables independientes o de control (COVARIABLES según SPSS), y compruebe que:

    a. Para las **variables dicotómicas** obtiene unas OR (y sus intervalos de confianza) idénticas a las que resultan de las tablas 2×2. Esto le

[10] Aquellas variables independientes que muestren asociación estadísticamente significativa con la variable dependiente deberían ser tenidas en cuenta para su inclusión en el modelo multivariante.

[11] Si las OR de cada estrato son muy parecidas o idénticas y diferentes a la OR global (sin estratificar), estamos ante una confusión (siendo la OR global un valor sesgado o confundido). Si las OR de cada estrato son muy diferentes, estamos ante una interacción, y la OR global es un promedio -sin interés- de la relación principal evaluada. Para una mejor comprensión recomendamos leer el documento ***"Confusión e Interacción"*** en esta misma colección.

ayudará a no equivocarse con el sentido de la relación, identificando el estrato de riesgo frente al estrato de referencia.

b. Para las **variables categóricas** con más de dos categorías, antes de introducirlas en el modelo, puede tomar la decisión de:
   i. Reducir sus dimensiones, agrupando categorías hasta "*dicotomizarla*".
   ii. Transformarla en un número (c-1) de ***variables dummy*** (o variables de diseño), siendo "c" el número de valores o de categorías distintas de dicha variable. Esto lo hace automáticamente el programa SPSS, en respuesta a una instrucción que se verá más tarde.

c. Para las **variables ordinales** puede adoptar la misma estrategia que hemos comentado en el punto anterior (reducirla a una dicotómica o transformarla en variables dummy), pero puede "*arriesgarse*" a introducirla como variable contínua, asumiendo que el cambio de valor entre cada categoría ordenada es lineal y proporcional.

d. Con las **variables contínuas** compruebe qué sentido clínico tiene el cambio de riesgo (OR) que comporta el incremento o decremento unitario en sus valores. Quizás llegue al convencimiento de que es mejor transformarla en una variable categórica, a costa de perder algo de información.

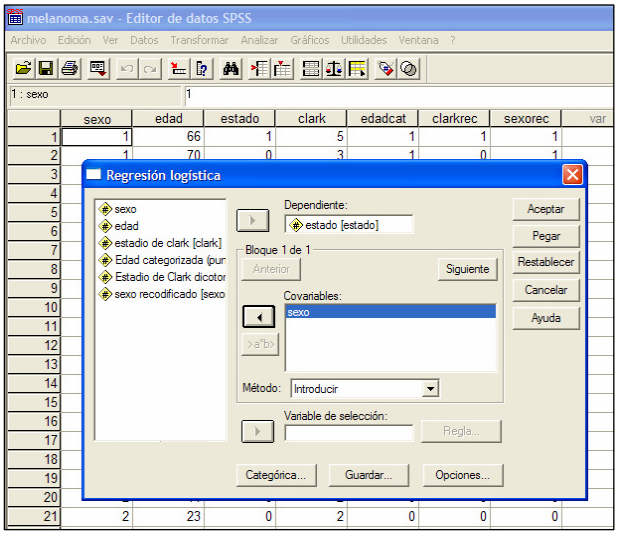
5. Finalmente pondere el número de variables a introducir en el modelo multivariante: pocas quizás no predigan mucho; muchas quizás "metan" mucho ruido (imprecisión). Debe tener en cuenta que los cálculos de la regresión se hacen a través del método de máxima verosimilitud con los datos de la muestra; una buena regla es no superar en ningún caso la relación "*una variable en el modelo por cada diez individuos en la muestra analizada*".

1. Cómo proceder en el programa SPSS.

Con la base abierta en la ventana de datos del SPSS, activamos la secuencia:

Analizar > Regresión > Logística Binaria

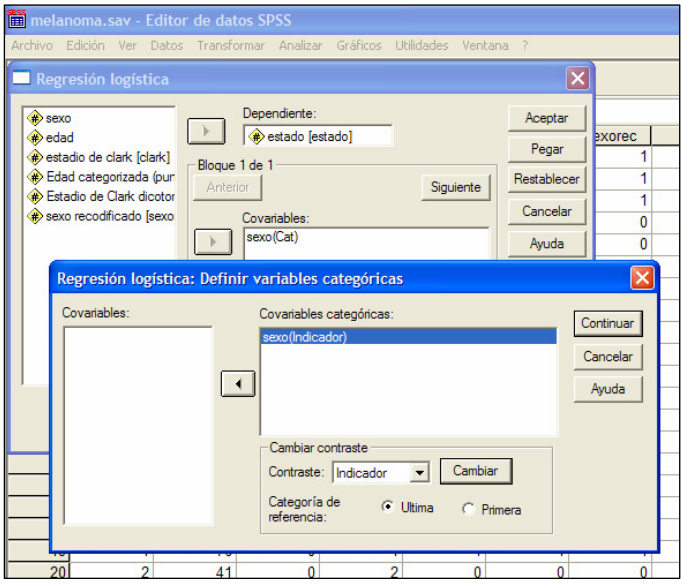
Y en el cuadro de diálogo que se abre tenemos que indicar:



- La **variable dependiente** (o resultado), la que deseamos modelizar o predecir, que será una categórica dicotómica, codificada con valores 0 y 1 (si no está así codificada el programa le asigna ese código interno).

- **La (o las) covariable (-s)**, ya sean *predictoras, confundentes*

y/o modificadoras de efecto, y que nos parecen deben ser incluidas en el modelo (por estas diferentes razones).



- **El método para seleccionar variables en el modelo.** Hay tres opciones principales:[12]

  - **El método "Introducir".** Permite al investigador tomar el mando, decidir que variables se introducen o extraen del modelo.

  - **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 reevalúa los coeficientes y su significación, pudiendo eliminar del modelo aquellos que no considera estadísticamente significativos.[13]

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

- A la hora de hacer la Regresión Logística, debemos especificar en el cuadro de diálogo principal cuál de las variables independientes o de control (co-variables) son categóricas, para lo cual se presiona en la pestaña **Categórica…** y en el siguiente cuadro de diálogo se seleccionan aquellas que cumplen este criterio de medida.

  - Una vez seleccionadas hay que especificar cuál es el método de **Contraste** (**Indicador,** por defecto) y cuál es la **Categoría de referencia** (la **Última**, por defecto). Si deseamos cambiar algunos de ellos debemos terminar aplicando la pestaña **Cambiar**, y veremos como se modifican en la ventana de variables.

- En **Opciones** podemos seleccionar tareas y resultados muy interesantes:

Estadísticos y gráficos. Las opciones disponibles son:

---

[12] El método INTRODUCIR es el adecuado cuando el objetivo del estudio es el ajuste de variables de confusión y la exploración de términos de interacción. Los métodos automáticos (ADELANTE ó ATRÁS) "por pasos" son adecuados para obtener diferentes modelos, con una finalidad predictiva, que pueden dar idea al investigador de aquellos más parsimoniosos. Como se ha comentado anteriormente debe tenerse en cuenta de que estos procedimientos automáticos en SPSS no incorporan el principio jerárquico.

[13] En los métodos por pasos (Adelante y Atrás) el programa SPSS da las opciones de elegir entre tres criterios para adoptar "*decisiones estadísticas*": razón de verosimilitud (RV), condicional y Wald. Cualquiera de ellos es correcto, aunque la mayoría de autores prefieren el criterio RV.

Gráficos de clasificación. Histograma de los valores actuales y pronosticados por el modelo para la variable dependiente.

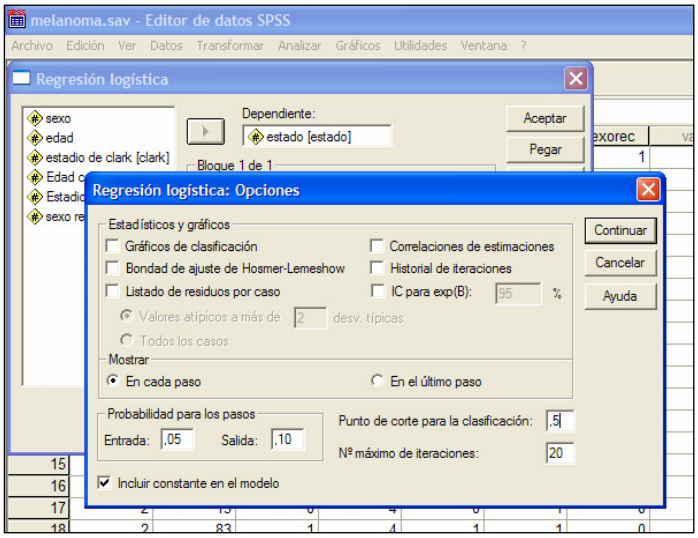
Bondad de ajuste de Hosmer-Lemeshow. Este estadístico de bondad de ajuste es un método para evaluar el ajuste global del modelo, más robusto que el estadístico de bondad de ajuste tradicionalmente utilizado en la regresión logística, especialmente para los modelos con covariables continuas y los estudios con tamaños de muestra pequeños. Se basa en agrupar los casos en deciles de riesgo y comparar la probabilidad observada con la probabilidad esperada dentro de cada decil.

Listado de residuos por caso. Muestra los residuos no estandarizados, la probabilidad pronosticada y los grupos de pertenencia observado y pronosticado.

Correlaciones de estimaciones. Muestra la matriz de correlaciones de las estimaciones de los parámetros para los términos del modelo.

Historial de iteraciones: Muestra los coeficientes y el logaritmo de la verosimilitud en cada iteración del proceso de estimación de los parámetros.

IC para la OR. Rango de valores que el N% de las veces incluye el valor e (2,718) elevado al valor del parámetro (coeficiente de regresión logística, b). Para cambiar el valor por defecto (95%), introduzca un número entre 1 y 99 (los valores habituales son 90, 95 y 99). Si el valor verdadero del parámetro poblacional es 0, los límites de confianza de Exp(B) deben incluir el valor 1 (el valor nulo de la OR).



En la opción **Mostrar** puede seleccione una de las alternativas del grupo para mostrar los estadísticos y los gráficos en cada paso o bien sólo para el modelo final (en el último paso).

En la opción **Probabilidad para los pasos**, le 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, de manera que 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. Para anular los valores por defecto, introduzca valores positivos en los cuadros Entrada y Salida. La entrada debe ser menor que la salida.

En la opción **Punto de corte para la clasificación**, se le 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 (tendrían el evento o resultado que se modeliza), mientras que aquellos con valores pronosticados menores que el punto de corte se clasifican como negativos (no tendrían el evento o resultado). Para cambiar los valores por defecto (0,5), debe introducir un valor comprendido entre 0,01 y 0,99.

En la opción **Nº máximo de iteraciones**, le permite cambiar el número máximo de veces que el modelo itera antes de finalizar.

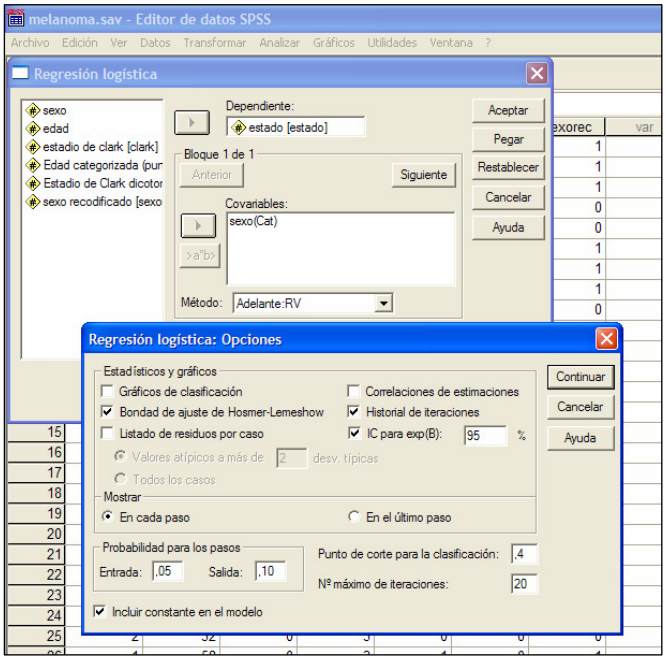
Por último, en la opción **Incluir constante en el modelo**, le permite indicar si el modelo debe incluir un término constante. Si se desactiva, el término constante será igual a 0.

2. Descripción de una salida de resultados.

Vamos a detallar una salida de resultados de una Regresión Logística Binaria realizada con el programa SPSS 13.0 en castellano.

Para el ejemplo utilizaremos la base de datos "**MELANOMA.sav**", donde la variable ***ESTADO*** es la dependiente (o resultado) que deseamos modelizar (codificada con valores 0 = vivo; 1 = muerto), y disponemos de las siguientes covariables:

- ***SEXO*** (codificada inicialmente como 0 = hombre y 1 = mujer)
- ***EDAD*** (recogida como variable contínua, representa los años cumplidos en el momento del diagnóstico de melanoma)
- ***NIVEL DE CLARK*** (variable ordinal con cinco categorías –del 1 al 5- especifica el nivel de *invasividad tumoral*)



En este primer documento sobre RL vamos a comenzar con una sola variable independiente (regresión logística simple), seleccionando en la ventana de variables "SEXO" y, pulsando la pestaña correspondiente, la introducimos en la ventana de covariables.

Seguidamente hay que señalar al programa que la variable SEXO es categórica, y seleccionadas las opciones más habituales, se obtiene la siguiente salida, con 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).

Regresión logística

Resumen del procesamiento de los casos

Casos no ponderados [a]		N	Porcentaje
Casos seleccionados	Incluidos en el análisis	122	100,0
	Casos perdidos	0	,0
	Total	122	100,0
Casos no seleccionados		0	,0
Total		122	100,0

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

Primero aparece un cuadro resumen con el número de casos (n) introducidos, los seleccionados para el análisis y los excluídos (casos perdidos, por tener algún valor faltante).

Codificación de la variable dependiente

Valor original	Valor interno
vivo	0
muerto	1

Inmediatamente aparece una tabla que 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.

En este caso coincide con la codificación empleada en la base de datos. Es importante que el valor 1 identifique a la categoría de la variable dependiente que resulte ser el resultado evaluado (en nuestro caso "muerto"), ya que ello permite comprender mejor el coeficiente b_i de las variables independientes y de control: un coeficiente de regresión positivo indicará que la probabilidad de morir por melanoma (valor interno 1) se incrementa con la exposición X.

Codificaciones de variables categóricas

		Frecuencia	Codificación de (1)
sexo	hombre	56	1,000
	mujer	66	,000

Esta segunda tabla muestra la codificación empleada en las variables independientes y de control (covariables). En nuestro ejemplo sólo es una variable la que hemos seleccionado (SEXO) y nos indica que la categoría codificada como 1 es hombre y la codificada como 0 es mujer. Además nos señala la frecuencia absoluta de cada valor. Si en el cuadro de Definir Variables Categóricas hemos 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. Por tanto, vamos a obtener la probabilidad de morir de los hombres (categoría 1) frente a las mujeres (categoría 0).

Bloque 0: Bloque inicial

En este bloque inicial se calcula la verosimilitud de un modelo que sólo tiene el término constante (a ó b_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 el ***menos dos veces el logaritmo neperiano de la verosimilitud*** (-2LL), que es un número positivo.

Historial de iteraciones [a,b,c]

Iteración		-2 log de la verosimilitud	Coeficientes Constante
Paso 0	1	157,105	-,623
	2	157,093	-,644
	3	157,093	-,644

a. En el modelo se incluye una constante.

b. -2 log de la verosimilitud inicial: 157,093

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

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. En este primer paso sólo se ha introducido el término constante en el modelo.[14] Como habíamos solicitado en **Opciones** el historial de iteraciones, la salida del ordenador nos muestra un resumen del proceso iterativo de estimación del primer parámetro (b_0). El proceso ha necesitado tres ciclos para estimar correctamente el término constante, porque la variación de -2LL entre el segundo y tercer bucle ha cambiado en menos del criterio fijado por el programa (0,001). También nos muestra el valor del parámetro calculado (b_0 = -0.644).

[14] Un modelo si poder predictivo alguno asigna a cualquier sujeto la probabilidad 0.5. Si n es el número de observaciones, $L = 0.5^n$, y LL = n × ln 0.5. En nuestro caso sería LL = 122 × (-0,6931472) = -84,563958; y el estadístico -2LL valdría 169,12791.

Tabla de clasificación[a,b]

Observado			Pronosticado		
			estado		Porcentaje correcto
			vivo	muerto	
Paso 0	estado	vivo	80	0	100,0
		muerto	42	0	,0
	Porcentaje global				65,6

a. En el modelo se incluye una constante.

b. El valor de corte es ,500

Esta tabla, que es muy parecida a la empleada para valorar una prueba diagnóstica, es la que 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. Por defecto se ha empleado un punto de corte de la probabilidad de Y para clasificar a los individuos de 0,5: esto significa que aquellos sujetos para los que la ecuación –con éste único término- calcula una probabilidad < 0,5 se clasifican como ESTADO=0 (vivo), mientras que si la probabilidad resultante es ≥ 0,5 se clasifican como ESTADO=1 (muerto). En este primer paso el modelo ha clasificado correctamente a un 65,6% de los casos, y ningún sujeto "muerto" ha sido clasificado correctamente.

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 0	Constante	-,644	,191	11,435	1	,001	,525

Variables que no están en la ecuación

			Puntuación	gl	Sig.
Paso 0	Variables	sexo(1)	4,786	1	,029
	Estadísticos globales		4,786	1	,029

Finalmente se presenta el parámetro estimado (B), su error estándar (E.T.) y su 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 (Exp(B)). En la ecuación de regresión sólo aparece, en este primer bloque, la constante, habiendo quedado fuera la variable SEXO. Sin embargo, como vemos en la subtabla inferior, como tiene una significación estadística asociada al índice de Wald de 0,029, el proceso automático por pasos continuará, incorporándola a la ecuación.[15]

[15] En versiones de SPSS anteriores a la 10.0 seguidamente se realiza una prueba estadística que permite comprobar la hipótesis nula del conjunto de coeficientes β de las variables NO-INCLUIDAS en la ecuación de regresión. Se denomina **Chi Cuadrado "Residual"**, y tiene un nº de grados de libertad igual al nº de variables no incluidas en el modelo (en nuestro caso sería 1). Si esta prueba es estadísticamente significativa se rechaza la H_0 de nulidad del conjunto de coeficientes de regresión, y por tanto el proceso debe continuar para incorporarlas, puesto que aportarán información al modelo.

Bloque 1: Método = Por pasos hacia adelante (Razón de verosimilitud)

Como puede apreciarse en el encabezamiento, se inicia de forma automática (POR PASOS) un segundo paso (BLOQUE 1), especificándose que se hace con el método hacia delante (ADELANTE) y empleando el criterio de la razón de la verosimilitud (RV) para contrastar las nuevas variables a introducir o sacar del modelo.

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

Iteración		-2 log de la verosimilitud	Coeficientes	
			Constante	sexo(1)
Paso 1	1	152,396	-,970	,755
	2	152,295	-1,057	,842
	3	152,295	-1,059	,843
	4	152,295	-1,059	,843

a. Método: Por pasos hacia adelante (Razón de verosimilitud)

b. En el modelo se incluye una constante.

c. -2 log de la verosimilitud inicial: 157,093

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

En la primera tabla se muestra el proceso de iteración, que ahora se realiza para dos coeficientes, la constante (ya incluida en el anterior paso) y la variable SEXO. Vemos como disminuye el -2LL respecto al paso anterior (el modelo sólo con la constante tenía un valor de este estadístico de 157,093, mientras que ahora se reduce a 152,295), y el proceso termina con cuatro bucles. Los coeficientes calculados son para la constante $b_0 = -1.059$, y para la variable SEXO $b_1 = 0{,}843$.

Seguidamente se nos aporta información sobre el ajuste del modelo con estas estimaciones. La probabilidad de los resultados observados en el estudio, dadas las estimaciones de los parámetros, es lo que se conoce por verosimilitud; pero como éste es un número pequeño (habitualmente menor de uno) se emplea el -2LL ("*menos dos veces el logaritmo neperiano de la verosimilitud"*). En la siguiente tabla (**PRUEBA OMNIBUS SOBRE LOS COEFICIENTES DEL MODELO**) se muestra una prueba Chi Cuadrado que evalúa la hipótesis nula de que los coeficientes (β) de todos los términos (excepto la constante) incluidos en el modelo son cero.[16] El estadístico Chi Cuadrado para este contraste es la diferencia entre el valor de -2LL para el modelo sólo con la constante y el valor de -2LL para el modelo actual:

$$\text{Chi cuadrado} = (-2LL_{\text{MODELO 0}}) - (-2LL_{\text{MODELO 1}}) = 157{,}093 - 152{,}295 = 4{,}798$$

[16] Esta prueba de bondad de ajuste es comparable al **test F global** que en la Tabla ANOVA se realiza para evaluar el modelo de Regresión Lineal.

Pruebas omnibus sobre los coeficientes del modelo

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	4,797	1	,029
	Bloque	4,797	1	,029
	Modelo	4,797	1	,029

Como puede verse en la tabla de la Prueba Omnibus, el programa nos 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 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, el Chi Cuadrado del Bloque es el mismo que el 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.

En nuestro ejemplo, al haber sólo una covariable introducida en el modelo (además de la constante), un único bloque y un único paso, coinciden los tres valores. La significación estadística (0,029) nos indica que el modelo con la nueva variable introducida (SEXO) mejora el ajuste de forma significativa con respecto a lo que teníamos.

Seguidamente se aportan tres medidas **RESUMEN DE LOS MODELOS**, complementarias a la anterior, para evaluar de forma global su validez: la primera es el valor del -2LL y las otras dos son Coeficientes de Determinación (R^2), parecidos al que se obtiene en Regresión Lineal, que expresan la proporción (en tanto por uno) de la variación explicada por el modelo. Un modelo perfecto tendría un valor de -2LL muy pequeño (idealmente cero) y un R^2 cercano a uno (idealmente uno).

Resumen de los modelos

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	152,295[a]	,039	,053

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

- -2 log 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 log de la verosimilitud (LL) para el modelo respecto al log de la verosimilitud (LL) para un modelo de línea base. Sus valores oscilan entre 0 y 1. En nuestro caso es un valor muy discreto (0,039) que indica que sólo el 3,9% de la variación de la variable dependiente es explicada por la variable incluida en el modelo.

- 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.

A continuación, como se lo hemos indicado en **Opciones**, se muestra una prueba de ajuste global del modelo que se conoce como PRUEBA DE HOSMER Y LEMESHOW.

Prueba de Hosmer y Lemeshow

Paso	Chi-cuadrado	gl	Sig.
1	,000	0	.

Tabla de contingencias para la prueba de Hosmer y Lemeshow

		estado = vivo		estado = muerto		
		Observado	Esperado	Observado	Esperado	Total
Paso 1	1	49	49,000	17	17,000	66
	2	31	31,000	25	25,000	56

Esta es otra prueba para evaluar la bondad del ajuste de un modelo de regresión logística. 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. Se trata de calcular, para cada observación del conjunto de datos, las probabilidades de la variable dependiente que predice el modelo, ordenarlas, agruparlas[17] y calcular, a partir de ellas, las frecuencias esperadas, y compararlas con las observadas mediante una prueba X^2.

Debe decirse aquí que esta prueba de bondad de ajuste tiene algunas "pegas": el estadígrafo de Hosmer y 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 eso, muchos autores proponen simplemente cotejar valores observados y esperados mediante simple inpección 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%.

Con nuestro modelo (con una sola variable, SEXO), la tabla de clasificación obtenida es la siguiente:

[17] Habitualmente en cuartiles, deciles o otra segmentación similar. Si son deciles el primer grupo contendría todos los sujetos con una p(Y) calculada menor de 0,1, el segundo grupo con aquellos cuyos valores están entre 0,1 y 0,2, y así sucesivamente.

Tabla de clasificación[a]

			Pronosticado		
			estado		Porcentaje correcto
	Observado		vivo	muerto	
Paso 1	estado	vivo	80	0	100,0
		muerto	42	0	,0
	Porcentaje global				65,6

a. El valor de corte es ,500

En la tabla de clasificación podemos comprobar que nuestro modelo tiene una especificidad alta (100%) y una sensibilidad nula (0%). Con la constante y una única variable predictora (SEXO), clasifica mal a los individuos que murieron por melanoma (ESTADO = 1) cuando el punto de corte de la probabilidad de Y calculada se establece (por defecto) en 50% (0,5). Si solicitamos en **Opciones** **Gráficos de clasificación** podremos obtener una representación de lo que está sucediendo:

```
Observed Groups and Predicted Probabilities

         ⇔                                                              ⇔
F        ⇔               m                                              ⇔
R     60 ⇳               m                                              ⇳
E        ⇔               m          m                                   ⇔
Q        ⇔               v          m                                   ⇔
U        ⇔               v          m                                   ⇔
E     40 ⇳               v          m                                   ⇳
N        ⇔               v          m                                   ⇔
C        ⇔               v          v                                   ⇔
Y        ⇔               v          v                                   ⇔
      20 ⇳               v          v                                   ⇳
         ⇔               v          v                                   ⇔
         ⇔               v          v                                   ⇔
         ⇔               v          v                                   ⇔
Predicted
⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇳⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇳⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇳⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩⇩
⇩⇩⇩⇩
  Prob:   0           ,25           ,5           ,75           1
  Group:  vvvvvvvvvvvvvvvvvvvvvvvvvvvvvvmmmmmmmmmmmmmmmmmmmmmmmmmmmmmm

          Predicted Probability is of Membership for muerto
          The Cut Value is 0,50
          Symbols: v - vivo
                   m - muerto
          Each Symbol Represents 5 Cases.
```

Podemos comprobar como nuestro modelo calcula unas probabilidades de Y menores de 0,5 para todos los casos, por los que los clasifica como "vivos" (ESTADO *PREDICHO* = 0). Esto concuerda con la escasa capacidad explicativa que se ha detectado con los coeficientes de determinación, y debe mejorar cuando se vayan incluyendo variables más explicativas del resultado o términos de interacción.

Por último, el programa nos 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 (β_i=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]	sexo(1)	,843	,389	4,697	1	,030	2,324	1,084	4,985
	Constante	-1,059	,281	14,144	1	,000	,347		

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

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

$$P(ESTADO=muerto) = \frac{1}{1 + \exp(1{,}059 - 0{,}843 \times SEXO)}$$

…y que podría servirnos para predecir la probabilidad de tener el resultado (ESTADO) de "muerto" de un individuo con melanoma en función de su género (SEXO). Así, un individuo hombre (SEXO = 1) tendría, según esta ecuación logística, una probabilidad de muerte…

$$P(ESTADO=muerto) = \frac{1}{1 + \exp(1{,}059 - 0{,}843 \times 1)} = \frac{1}{1 + 2{,}718^{(0{,}216)}} = 0{,}446$$

… con esta probabilidad predicha -como es menor que 0,50- se clasificaría como "ESTADO=vivo".

Por último, la salida de ordenador nos 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 siguiente presenta, para cada variable del modelo (en nuestra caso sólo una, SEXO), los cambios en la verosimilitud si dicha variable se elimina; si la significación estadística asociada (**Sig. del cambio**) fuese mayor que el criterio de exclusión (***POUT***) establecido, la variable se eliminaría del modelo en el paso siguiente. En nuestro ejemplo, como el cambio de verosimilitud es estadisticamente significativo (p = 0,29), la variable en cuestión queda en el modelo.

Modelo si se elimina el término

Variable		Log verosimilitud del modelo	Cambio en -2 log de la verosimilitud	gl	Sig. del cambio
Paso 1	sexo	-78,546	4,797	1	,029