DOCUWEB FABIS

Dot. Núm 0702013

Cómo hacer una Regresión Logística binaria "paso a paso" (II): análisis multivariante.

Aguayo Canela, Mariano; Lora Monge, Estrella

Hospital Universitario Virgen Macarena (Sevilla), Servicio de Medicina Interna.

Área Sanitaria de Osuna (Sevilla)

Resumen

En una primera parte de este documento se repasaron los principales aspectos teóricos de la Regresión Logística Binaria y los procedimientos que aporta el programa SPSS para llevar a cabo un análisis simple. En éste se muestra, paso a paso, la estrategia más recomendable para realizar con éxito una RL multivariante, tanto para ajustar las estimaciones y detectar interacciones como para obtener modelos predictivos.

1) Introducción

Antes de abordar este documento es muy recomendable leer detenidamente su primera parte[1] y los dos documentos[1,2] sobre "***Confusión e interacción***", ya que en el primero se explica una salida tipo de Regresión Logística con el programa SPSS con la interpretación detallada de los diferentes cuadros de diálogo, y en los segundos los conceptos básicos sobre confusión e interacción y cómo detectar estos fenómenos en los análisis estratificado y multivariante.

Para hacer práctico el documento y reproducible la tarea, vamos a trabajar con la base de datos de "**BAJO PESO AL NACER**", donde se registran 189 nacimientos en un estudio de casos y controles: 59 casos de RN de bajo peso (menos de 2.500 gramos) y 130 controles de RN de peso normal (al menos 2.500 gramos); y una serie de variables independientes, *confundentes* y de control.

En resumen, los pasos que recomendamos dar en un análisis de Regresión Logística Multivariante (REM) son:

Paso 0. **Tener claro qué se pretende en el estudio** (¡pudiera resultar una obviedad, pero no lo es!)

Paso 1. **Explorar las relaciones bivariantes (entre las predictoras y la dependiente).**

Paso 2. **Evaluar posibles interacciones y/o factores de confusión.**

Paso 3. **Construir un modelo o ecuación de RLM.**

Paso 4. **Evaluar el modelo final.**

[1] **Confusión e interacción (1): Qué son, qué suponen y cómo manejarlas en el análisis estratificado.** *Docuweb fabis nº 0702007.*

[1] **Como hacer una Regresión Logística con SPSS "paso a paso". (I).** *Docuweb fabis nº 0702012.*

[2] **Confusión e interacción (2): Su abordaje en el análisis multivariante.** *Docuweb fabis nº 0702011.*

Correspondencia: marianoaguayo@telefonica.net

Paso 0. Tener claro qué se pretende en el estudio.

Como ya hemos indicado en reiteradas ocasiones, el análisis de los datos está al servicio de los objetivos del estudio de investigación, y no al contrario. El investigador tiene que tener bien definido qué quiere obtener y, solo entonces, aplicar un procedimiento de análisis matemático.

Esto es especialmente importante cuando se llevan a cabo análisis multivariantes, en los que se introducen muchas (> 2) variables simultáneamente para evaluar sus relaciones o asociaciones, por lo que las posibilidades de encontrar dependencias espúreas (cuando no absurdas) es elevada; y por otra parte, las probabilidades de no encontrar relaciones importantes por no saber cómo explorarlas o por la imprecisión de los datos (error aleatorio) también es alta.

La Regresión Logística Multivariante tiene tres objetivos básicos:

1. **Obtener una estimación no sesgada o ajustada de la relación entre la variable dependiente (o resultado) y una variable independiente** que es la que el investigador quiere conocer, sobre la que el investigador desea averiguar su papel.

Así en nuestro ejemplo, pudiera ser que el estudio se planteara como...

"*Efecto del tabaquismo materno sobre el bajo peso al nacer: un estudio caso-control*"

En este proyecto así formulado los investigadores colocan al tabaquismo materno (nuestra variable ***TABACO***) en la independiente principal, de forma que la relación básica a explorar será ***TABACO → BAJOPESO.*** El análisis de RLM será una herramienta excelente para controlar posibles factores de confusión en la relación principal evaluada, siempre y cuando estos factores se hayan medido y registrado correctamente en los individuos del estudio.

2. **Evaluar varios factores simultáneamente que estén presumiblemente relacionados de alguna manera (o no) con la variable dependiente**, y conocer su papel (predictor, confundente, modificador de efecto) y su efecto de forma ajustada.

Así en nuestro ejemplo, el estudio se plantearía ahora como...

"Factores que influyen en el bajo peso al nacer"

En este caso no hay una variable independiente principal sino varias, que habrán sido seleccionadas por el investigador tras un profundo conocimiento del tema en cuestión y una rigurosa búsqueda bibliográfica. El análisis de RLM permitirá obtener medidas de asociación (OR) para cada variable ajustadas por las demás y detectar posibles interacciones entre ellas y el efecto estudiado (***BAJOPESO***).

3. **Construir un modelo y obtener una ecuación con fines de predicción o cálculo del riesgo**, de manera que éste pueda estimarse para un nuevo individuo con una cierta validez y precisión.

En esta ocasión, en el proyecto que hemos puesto de ejemplo, se materializaría como:

"Predicción del bajo peso al nacer: una fórmula para calcular el riesgo"

Para llevar a cabo con éxito este proyecto, el investigador debe conocer muy bien el tema en cuestión, tener información fidedigna de aquellos factores que ya se conocen de riesgo o de protección, y disponer de una amplia muestra de individuos donde medir con el menor error posible estas variables. La RLM deberá probar múltiples modelos para quedarse con el más

predictivo (menor error estándar y mayor coeficiente de determinación) y con menor número de variables (más armonioso).

Como puede entenderse, estos tres objetivos van a condicionar, además del diseño del estudio y los requisitos previos de conocimiento y madurez del tema en cuestión, los procedimientos a llevar a cabo en los análisis de RLM. Así, mientras que en el primer supuesto (una única variable principal independiente) el investigador debe emplear los métodos manuales (puesto que debe incluir necesariamente en la ecuación de regresión a la variable independiente ***TABACO*** en nuestro caso), en los supuestos 2º y 3º puede utilizar, según su pericia en el manejo del programa estadístico, una combinación de métodos manuales (**"introducir"** en el SPSS) y automáticos (**"paso a paso"** en el SPSS).

Paso 1. Explorar las asociaciones bi-variantes.

Lo primero que debería explorarse es la posible asociación entre la variable dependiente "Y" (que se desea predecir o modelizar) y las diferentes variables independientes "X" y de control, medidas como categóricas y tomadas de una en una, para hacer valoraciones bi-variantes.[3]

$$\mathbf{X_i \rightarrow Y}$$

Esto tiene como objeto tener una primera aproximación a la estimación de la medida de asociación, la OR, clasificando a las variables según el valor de esta medida y de su significación estadística en el contraste de hipótesis Chi cuadrado, aun reconociendo que pudieran tratarse de estimaciones sesgadas si existiese confusión, o de estimaciones poco informativas si existiese interacción con una tercera variable.

Un procedimiento recomendable para los principiantes (aunque más laborioso) es explorar consecutivamente las asociaciones bi-variantes mediante el análisis de las **tablas de contingencia**, seguido del análisis de **regresión logística binaria**, para comprobar cómo se distribuyen los sujetos en las diferentes categorías, marcar claramente la categoría de referencia y llegar a la conclusión de que se obtienen estimaciones idénticas por ambos métodos.

Y es que un aspecto previo muy importante a tener en cuenta en el análisis de variables categóricas es el tema de la **codificación numérica de las categorías** y la consideración que les da el programa estadístico SPSS. Tras comprobar que efectivamente nuestras variables están medidas en una escala NOMINAL (también conocida por cualitativa o categórica), conviene fijarse en los números que identifican cada categoría, pues en los procedimientos automáticos de análisis, el programa va a considerar siempre la categoría de referencia (la que tiene riesgo basal ó RR=1) aquella que tiene menor valor numérico, esto es, el "0" si las categorías codificadas El resultado que se obtiene en la ventana de resultados del SPSS es el siguiente:

Tablas de contingencia

Resumen del procesamiento de los casos

son "0" y "1"; ó el "1" si las categorías están codificadas con "1" y "2", etc. De hecho, cuando en la RL le indicamos al programa SPSS que una covariable introducida en el análisis es una

[3] Recuérdese que en la Regresión Logística Binaria la variable dependiente es categórica y dicotómica (por definición), y el modelo calcula para cada variable independiente un coeficiente de regresión que puede transformarse fácilmente en la Odds Ratio (OR), la medida fundamental que evalúa la fuerza de asociación entre dos variables categóricas dicotómicas. Ello quiere decir que este modelo matemático trabaja mejor con variables independientes categóricas dicotómicas, aunque, como veremos luego, se pueden incluir también variables numéricas o contínuas.

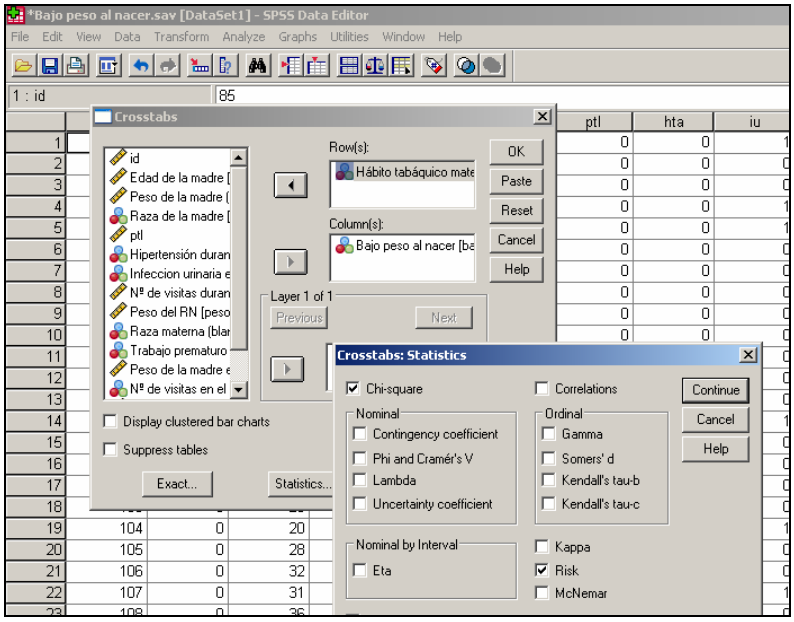
CATEGÓRICA, automáticamente convierte sus códigos iniciales en ceros (“0”) y unos (“1”) si se trata de una variable dicotómica, o las transforma en tantas variables *dummy* como categorías menos una tiene la variable inicial, de manera que, al hacer una regresión logística, el programa siempre va a evaluar variables categóricas dicotómicas codificadas con “0” y “1”, y antes de llevar a cabo el ajuste nos mostrará una tabla resumen de codificación, para que comprobemos cómo han quedado definitivamente para el análisis las diferentes variables introducidas.

Veamos de forma práctica en la base de datos BAJO PESO AL NACER qué pasa con la relación entre la variable ***TABACO*** (inicialmente recogida como dicotómica y codificada con “0” (no fuma) y “1” (fuma) y la variable dependiente ***BAJOPESO***.

TABACO BAJOPESO

Hagamos primero la evaluación mediante una tabla de contingencia (Crosstabs), colocando en las columnas la variable dependiente (***BAJOPESO***) y en las filas la variable independiente o criterio (***TABACO***).

Marquemos en Estadísticas (Statistics...) Chi-cuadrado y Riesgo, para obtener la doble evaluación: la del contraste de hipótesis y la de la medida de asociación.



La salida que obtendremos será la siguiente:

Hábito tabáquico materno * Bajo peso al nacer Crosstabulation

Count

		Bajo peso al nacer		
		>=2500 gr	<2500 gr	Total
Hábito tabáquico materno	No Fuma	86	29	115
	Fuma	44	30	74
Total		130	59	189

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	4,924[b]	1	,026		
Continuity Correction[a]	4,236	1	,040		
Likelihood Ratio	4,867	1	,027		
Fisher's Exact Test				,036	,020
Linear-by-Linear Association	4,898	1	,027		
N of Valid Cases	189				

a. Computed only for a 2x2 table

b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 23,10.

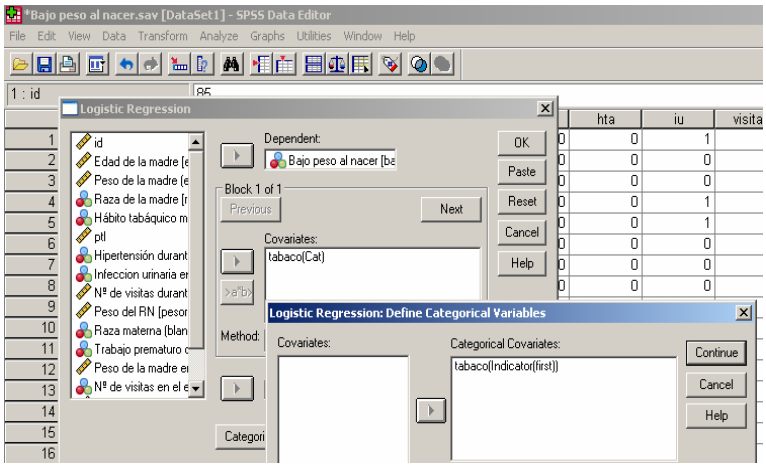
Vemos como la variable ***TABACO*** se asocia con la variable ***BAJOPESO***, de manera que la proporción de niños RN con bajo peso es mayor entre fumadoras que entre madres no fumadoras, con una Chi-cuadrado de 4,924 (***p*** asociada 0,026).

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Hábito tabáquico materno (No Fuma / Fuma)	2,022	1,081	3,783
For cohort Bajo peso al nacer = >=2500 gr	1,258	1,013	1,561
For cohort Bajo peso al nacer = <2500 gr	,622	,409	,945
N of Valid Cases	189		

Por otra parte, la fuerza de esta asociación es 2,022 que representa el riesgo que tienen las madres fumadoras frente a las que no fuman (categoría de referencia en este contraste, al tener el valor "0") de tener un RN de bajo peso. Dicho de otra manera, el hábito materno de ser fumadora hace que se incremente por dos (se duplique) el riesgo de tener un RN de bajo peso.

Veamos qué saldría a través de una Regresión Logística Binaria. Como puede verse en la ventana correspondiente, se ha marcado ***TABACO*** como categórica y hemos indicado al programa que la categoría de referencia es la primera (*first*), esto es, la marcada con el código-valor "0" (no fumadora).



Logistic Regression

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	189	100,0
	Missing Cases	0	,0
	Total	189	100,0
Unselected Cases		0	,0
Total		189	100,0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
>=2500 gr	0
<2500 gr	1

Categorical Variables Codings

		Frequency	Parameter coding (1)
Hábito tabáquico materno	No Fuma	115	,000
	Fuma	74	1,000

Vemos como el programa SPSS nos muestra, antes de enseñarnos la salida del análisis de regresión logística propiamente dicho, unas tablas con las codificaciones de las variables que hemos introducido, codificaciones que son "internas" o propias del programa. Así, en el caso de la variable dependiente ***BAJOPESO*** ha mantenido el mismo criterio que nosotros, de forma que la categoría "1" es la de "< 2500 gramos", esto es, identifica al grupo de RN con peso bajo, que son los que queremos predecir. Y en el caso de la variable ***TABACO*** la categoría "0" es no fumadora y la categoría "1" es fumadora", también de forma análoga a lo que ya teníamos en la base de datos.[4]

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	,0% C.I.for EXP(I	
								Lower	Upper
Step 1[a]	tabaco(	,704	,320	4,852	1	,028	2,022	1,081	3,783
	Constar	-1,087	,215	25,627	1	,000	,337		

a. Variable(s) entered on step 1: tabaco.

El resultado final muestra el coeficiente B de la variable ***TABACO*** (0,704), su error estándar (S.E. 0,320), el estadístico de Wald del contraste de hipótesis (siendo H_0 : B=0), los grados de libertad (df=1) y el valor ***p*** asociado al contraste (Sig. 0,028). Seguidamente aparece EXP(B), que corresponde a la OR de la variable analizada (***TABACO***) con la dependiente predicha (***BAJOPESO***), y que -como podemos comprobar- arroja un resultado idéntico al obtenido mediante el procedimiento "Tablas de contingencia": 2,022 con un $IC_{95\%}$ entre 1,081 y 3,783.

De forma similar podríamos explorar la asociación entre otras variables presumiblemente predictoras o de control y la variable dependiente ***BAJOPESO***.

Estas son nuestras recomendaciones:

A) Si se trata de variables categóricas lo haremos a través del procedimiento Tablas de contingencia (Crosstabs).

[4] Un proceder muy recomendable es codificar en la base de datos las variables categóricas dicotómicas con ceros y unos, asignándole el valor "1" a la categoría que proponemos como *expuesta* o *a riesgo*, y la categoría "0" a la *de referencia* o *categoría basal*, con la que se va a comparar la categoría "expuesta". Así, seguiremos el mismo criterio que el programa SPSS y no tendremos problemas en la interpretación de los análisis. Igualmente, si tenemos variables categóricas policotómicas (con más de dos categorías), se recomienda codificar con "0" la categoría de menor riesgo, la *de referencia* o *basal*: será la que empleará el programa para diseñar las variables *dummys*, que construirá con el resto de categorías manteniendo la "0" como categoría fija de referencia.

1. Lo mejor es trabajar siempre con variables categóricas dicotómicas, pues en ellas se establece una categoría "de referencia" y se calcula la OR para la categoría "expuesta" en relación a dicha categoría "de referencia".[5]

2. Si tenemos variable policotómicas un procedimiento aconsejable es colapsar o agrupar categorías para transformarlas en dicotómicas, ya que en caso contrario, el programa SPSS las convertirá automáticamente en tantas variables *dummys* como categorías menos una tenga la variable inicial.

3. Si se trata de variables ordinales, podemos explorar si hay asociación lineal con la variable dependiente e introducirlas en el modelo logístico como variables contínuas (no marcaríamos "categóricas"), ofreciéndonos entonces la OR calculada un valor medio del riesgo de cada categoría frente a la inmediatamente anterior en orden decreciente.

B) Si se trata de variables contínuas podemos optar por dos soluciones:

1. Evaluar si hay diferencias en las medias de la dicha variable contínua comparando los dos grupos que se establecen por las dos categorías de la variable dependiente ***BAJOPESO***, a través de un test T de Student o de un ANOVA de una vía.

2. Intentar transformaciones de la variable contínua en categórica, preferiblemente dicotómica. El punto de corte puede establecerse arbitrariamente, aunque debe tenerse en cuenta…

   - Si existe una hipótesis teórica que pueda operativizarse en el estudio y que tenga cierto sentido explorar; así, por ejemplo, si se sospecha que el seguimiento médico, operativizado en el número de visitas durante el embarazo (***VISITAS***) puede ser un factor predictor de ***BAJOPESO***, una categorización posible de la variable independiente sería *"ninguna visita ni control médico"* (***VISITAS*** = 0) versus "*al menos una visita médica durante la gestación*" (***VISITAS*** ≥ 1).

   - Si no hay una hipótesis previa, un buen punto de corte es la mediana, que permite agrupar los individuos en dos grupos de igual tamaño; o los quantiles en general.

Aunque nos parezca que transformar una variable inicialmente recogida como contínua en categórica es perder información, a la hora del análisis se gana en eficiencia y, sobre todo, claridad en la interpretación. Debe tenerse en cuenta que si introducimos en el análisis de Regresión Logística una variable independiente contínua, la OR que se obtiene en el ajuste es el riesgo de cada valor numérico en relación al valor inmediatamente anterior; así por ejemplo, si en nuestro estudio sobre el BAJO PESO AL NACER la variable introducida es la ***EDAD*** (de la madre) en años cumplidos, la OR es un promedio del riesgo de tener un RN de bajo peso para cada edad respecto a la (edad – 1), asumiendo que ésta (la OR) no cambia para cualquier par de valores de años que consideremos en el intervalo de edad explorado.

Un resumen se muestra en el cuadro siguiente. En rojo se señalan las pruebas o test de hipótesis empleados para contrastar la H_0 (*no diferencias*, *no asociación*) y en azul la medida de asociación.

[5] Recuerde que con categóricas policotómicas el procedimiento Tablas de Contingencia (Crosstabs) no puede calcular un riesgo (Risk), puesto que no puede adoptarse una categoría de referencia.

Cuando la variable predictora (independiente) es...	La evaluación de su asociación con una variable dependiente dicotómica...		Nuestra recomendación para introducirla en la Regresión Logística binaria Multivariante
	En el análisis bivariante simple	En el análisis de Regresión Logística binaria simple	
CATEGÓRICA DICOTÓMICA	Chi cuadrado. OR	Test de Wald. OR	Déjela tal cual.
CATEGÓRICA POLICOTÓMICA	Chi cuadrado. No se calcula OR	Test de Wald. Automáticamente se crean tantas variables *dummys* como (categorías – 1) y para cada una de ellas se obtiene una OR	Intente agrupar o colapsar categorías para transformarla en dicotómica.
ORDINAL	Chi cuadrado o alternativamente otras pruebas. No se calcula OR si > 2 categorías		Intente agrupar o colapsar categorías para transformarla en dicotómica, o pruebe introducirla como continua si detecta "asociación lineal".
CONTINUA	T test o ANOVA para explorar la diferencia de medias.	Test de Wald. Se calcula una OR para cada valor en relación al (valor – 1)	Intente categorizarla (si es posible dicotomizarla).

Mostramos a continuación algunas de estas evaluaciones en nuestro estudio, a manera de ejemplo.

Veamos cómo se relaciona la variable ***RAZA*** con la variable ***BAJOPESO***. La predictora o independiente tiene en este caso tres categorías (1= "*blanca*"; 2= "*negra*"; 3= "*otras*"). Una evaluación de la asociación de esta variable policotómica con la variable dependiente del estudio arroja los siguientes resultados:

Tabla de contingencia Raza de la madre * Bajo peso al nacer

Recuento

		Bajo peso al nacer		
		>=2500 gr	<2500 gr	Total
Raza de la madre	Blanca	73	23	96
	Negra	15	11	26
	Otras	42	25	67
Total		130	59	189

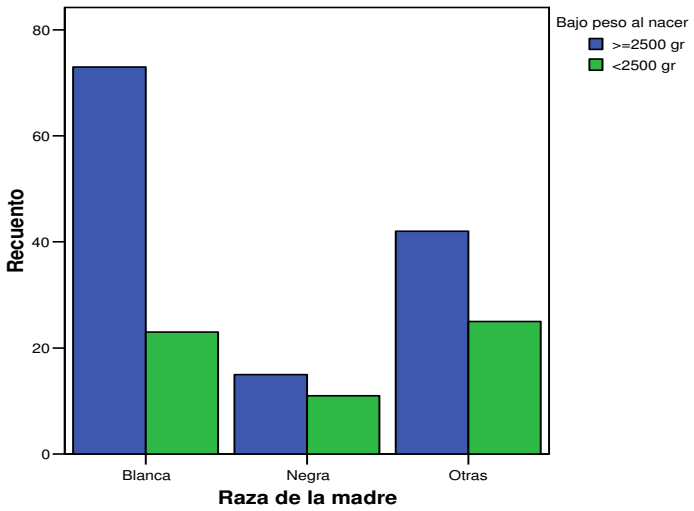
Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	5,005[a]	2	,082
Razón de verosimilitud	5,010	2	,082
Asociación lineal por lineal	3,570	1	,059
N de casos válidos	189		

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 8,12.

No se encuentra asociación porque el test Chi cuadrado no es significativo (p=0,082), aunque si evaluamos el gráfico de barras agrupadas podemos ver mejor lo que está pasando:

Gráfico de barras



A simple vista parece que la proporción de RN de bajo peso al nacer es mayor en mujeres de raza "*negra*" y menor en las madres de raza "*blanca*". Esto permite evaluar la asociación pero transformando antes la variable ***RAZA*** para convertirla en una dicotómica; tendría sentido colapsar la categoría "*Blanca*" con "*Otras razas*", o juntar "*Negras*" con "*Otras razas*", pero desde luego no se nos debe ocurrir agrupar en una sola categoría la raza "*Blanca*" con la raza "*Negra*", puesto que tienen los valores más extremos de proporción de RN de bajo peso.

Veamos que se obtiene al recodificar ***RAZA*** en una dicotómica con una categoría de raza "*Blanca*" y otra categoría con las demás razas ("*Negra*" + "*Otras*"):

Tabla de contingencia Raza materna (blanca vs otras) * Bajo peso al nacer

Recuento

		Bajo peso al nacer		
		>=2500 gr	<2500 gr	Total
Raza materna (blanca vs otras)	blanca	73	23	96
	otras	57	36	93
Total		130	59	189

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	4,787[b]	1	,029		
Corrección por continuidad[a]	4,125	1	,042		
Razón de verosimilitud	4,815	1	,028		
Estadístico exacto de Fisher				,041	,021
Asociación lineal por lineal	4,762	1	,029		
N de casos válidos	189				

a. Calculado sólo para una tabla de 2x2.

b. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 29,03.

Estimación de riesgo

	Valor	Intervalo de confianza al 95%	
		Inferior	Superior
Razón de las ventajas para Raza materna (blanca vs otras) (blanca / otras)	2,005	1,070	3,754
Para la cohorte Bajo peso al nacer = >=2500 gr	1,241	1,019	1,510
Para la cohorte Bajo peso al nacer = <2500 gr	,619	,399	,960
N de casos válidos	189		

Se encuentra asociación entre ***RAZAREC*** y ***BAJOPESO***, de manera que las madres de raza "*diferente a la blanca*" tienen el doble de riesgo (OR = 2,005) de tener un RN de bajo peso que las madres de raza "*blanca*".

Si recurrimos a la Regresión Logística e introducimos la variable ***RAZA*** como está inicialmente recogida (con tres categorías), el programa la convertirá automáticamente en dos variables dicotómicas *dummys*, para poder así calcular la OR de cada categoría frente a una de referencia. Veámoslo:

Codificaciones de variables categóricas

		Frecuencia	Codificación de	
			(1)	(2)
Raza de la madre	Blanca	96	,000	,000
	Negra	26	1,000	,000
	Otras	67	,000	1,000

Se crean dos variables nuevas: ***raza (1)*** y ***raza (2)***. La raza "*Blanca*" ha sido tomada por el programa SPSS como categoría de referencia (tiene valores ceros en ambas), ya que era la que tenía una codificación absoluta más baja en la variable original, por lo que ***raza (1)*** es una dicotómica en la que el valor "1" es "*Negra*" y ***raza (2)*** es una dicotómica en la que el valor "1" es "*Otra raza*".

Y en la ecuación de Regresión Logística:

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95,0% para EXP(B)	
								Inferior	Superior
Paso 1[a]	raza			4,922	2	,085			
	raza(1)	,845	,463	3,323	1	,068	2,328	,939	5,772
	raza(2)	,636	,348	3,345	1	,067	1,889	,955	3,736
	Constante	-1,155	,239	23,330	1	,000	,315		

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

...vemos como se introducen estas dos variables nuevas que llevan información desagregada de la antigua ***RAZA***. De hecho la variable original no tiene interpretación en la ecuación, está solo para indicarnos que de ella se han generado las dos *dummys* (aunque puede comprobarse que el estadístico de Wald tiene dos grados de libertad -hay tres categorías- y el valor ***p*** asociado es similar al obtenido en la tabla de contingencia).

Con las nuevas variables ahora sí podemos obtener su OR (puesto que son dicotómicas) de manera que concluiríamos diciendo que, comparado con ser de raza "*Blanca*", ser de raza "*Negra*" multiplica por 2,33 y ser de "*Otra raza*" multiplica por 1,89 las probabilidades de tener un RN de bajo peso (aunque en ambos casos no se alcanza la significación estadística).

Otra asociación que podemos explorar, a manera de ejemplo es la Relación ***HTA*** * ***BAJOPESO***. Se trata de dos variables cualitativas dicotómicas, por lo que en el análisis bivariante recurrimos a la tabla 2x2 y calculamos el estadístico de contraste Chi cuadrado:

Tabla de contingencia Hipertensión durante el embarazo * Bajo peso al nacer

			Bajo peso al nacer		
			>=2500 gr	<2500 gr	Total
Hipertensión durante el embarazo	No hipertensa	Recuento	125	52	177
		% de Bajo peso al nacer	96,2%	88,1%	93,7%
	Hipertensa	Recuento	5	7	12
		% de Bajo peso al nacer	3,8%	11,9%	6,3%
Total		Recuento	130	59	189
		% de Bajo peso al nacer	100,0%	100,0%	100,0%

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	4,388[b]	1	,036		
Corrección por continuidad[a]	3,143	1	,076		
Razón de verosimilitud	4,022	1	,045		
Estadístico exacto de Fisher				,052	,042
Asociación lineal por lineal	4,365	1	,037		
N de casos válidos	189				

a. Calculado sólo para una tabla de 2x2.

b. 1 casillas (25,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 3,75.

La proporción de madres hipertensas en el grupo de "*bajo peso al nacer*" es algo más del triple que en el grupo de "*peso normal al nacer*" (11,9% vs 3,8%), con una diferencia puntual de esta proporción de 8,1%.

Si calculáramos el $IC_{95\%}$ de esta proporción obtenemos:

- Límite inferior: - 0,9%
- Limite superior: 16,9%

Vemos que el intervalo de confianza contiene el valor cero (aunque… ¡por muy poquito!) Ello justifica que, aunque el contraste con X^2 sale estadísticamente significativo, con valor p=0,036, al aplicar la corrección por continuidad o el test exacto de Fisher, más conservadores, ya no se alcance el valor crítico de p (< 0,05). Por otro lado es un intervalo muy amplio, ofreciendo un resultado impreciso, como consecuencia de que hay muy pocas mujeres hipertensas (sólo 12 de 189) en la muestra estudiada, con unas proporciones de expuestas también muy pequeñas, tanto en el grupo de casos (sólo 7 de 59) como en el de controles (sólo 5 de 130).

La aproximación epidemiológica, a través del cálculo de una medida de fuerza de asociación (la OR), sería: (7 × 125) / (5 × 52) = 875/260 = 3,365. Veamos la salida de SPSS:

Estimación de riesgo

	Valor	Intervalo de confianza al 95%	
		Inferior	Superior
Razón de las ventajas para Hipertensión durante el embarazo (No hipertensa / Hipertensa)	3,365	1,021	11,088
Para la cohorte Bajo peso al nacer = >=2500 gr	1,695	,862	3,333
Para la cohorte Bajo peso al nacer = <2500 gr	,504	,296	,856
N de casos válidos	189		

El $IC_{95\%}$ de la OR calculado por el programa no contiene el valor 1, aunque lo roza por su extremo inferior (1,02), y es muy amplio (entre 1,02 y 11,09). Todo ello en el mismo sentido de lo que ya se ha comentado.

Veamos por último como se evaluaría la asociación entre una variable predictora cuantitativa y nuestra dependiente categórica dicotómica. Exploraremos la relación ***EDAD*** * ***BAJOPESO***

La edad de la madre (***EDAD***) es una variable numérica (años cumplidos en el momento de detectarse la gestación), por lo que la evaluación de su posible relación con ***BAJOPESO*** (dicotómica) es a través de una **comparación de medias**, siempre que se cumplan las premisas para aplicar una prueba paramétrica. Una opción alternativa sería categorizar la edad de la madre en una variable cualitativa, pero a priori no tenemos ningún criterio de referencia para establecer puntos de corte, por lo que optamos por la primera fórmula.

Hagamos primero una descripción de la variable ***EDAD*** en ambos grupos de comparación (establecidos por el ***BAJOPESO***), mediante el procedimiento EXPLORAR de SPSS:

Descriptivos

Bajo peso al nacer				Estadístico	Error típ.
Edad de la madre	>=2500 gr	Media		23,66	,490
		Intervalo de confianza para la media al 95%	Límite inferior	22,69	
			Límite superior	24,63	
		Media recortada al 5%		23,41	
		Mediana		23,00	
		Varianza		31,187	
		Desv. típ.		5,585	
		Mínimo		14	
		Máximo		45	
		Rango		31	
		Amplitud intercuartil		9	
		Asimetría		,754	,212
		Curtosis		,503	,422
	<2500 gr	Media		22,31	,587
		Intervalo de confianza para la media al 95%	Límite inferior	21,13	
			Límite superior	23,48	
		Media recortada al 5%		22,19	
		Mediana		22,00	
		Varianza		20,354	
		Desv. típ.		4,511	
		Mínimo		14	
		Máximo		34	
		Rango		20	
		Amplitud intercuartil		6	
		Asimetría		,300	,311
		Curtosis		-,162	,613

Con esto ya tenemos mucha información: la diferencia de medias de edad materna entre ambos grupos de RN es (23,66 – 22,31) sólo 1,35 años, y los $IC_{95\%}$ de dichas medias en cada grupo de comparación se superponen ampliamente, por lo que es muy probable que no existan diferencias estadísticamente significativas y que las variables ***EDAD*** (materna) y ***BAJOPESO*** no estén relacionadas en la población.

Por otro lado, las pruebas de normalidad detectan problemas en el grupo control, por lo que no podrán aplicarse, *sensu estricto*, los test paramétricos.

Pruebas de normalidad

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Bajo peso al nacer	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Edad de la madre	>=2500 gr	,109	130	,001	,950	130	,000
	<2500 gr	,088	59	,200*	,982	59	,521

*. Este es un límite inferior de la significación verdadera.

a. Corrección de la significación de Lilliefors

Sin embargo, como la muestra es grande (n > 100) podemos "arriesgarnos", y aplicar un test t de Student, única manera de obtener una medida de asociación (en este caso la diferencia de medias) y una estimación interválica, ya que los test no paramétricos no se llevan a cabo con estadísticos basados en momentos (media, desviación típica, etc.):

Estadísticos de grupo

	Bajo peso al nacer	N	Media	Desviación típ.	Error típ. de la media
Edad de la madre	>=2500 gr	130	23,66	5,585	,490
	<2500 gr	59	22,31	4,511	,587

Prueba de muestras independientes

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
									95% Intervalo de confianza para la diferencia	
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	Inferior	Superior
Edad de la madre	Se han asumido varianzas iguales	3,429	,066	1,638	187	,103	1,356	,828	-,277	2,990
	No se han asumido varianzas iguales			1,774	136,941	,078	1,356	,765	-,156	2,869

En efecto no hay diferencias estadísticamente significativas ("***p***" asociada al contraste 0,1) y la **diferencia de medias** (1,356 años) tiene un $IC_{95\%}$ que incluye el valor nulo (cero), oscilando entre - 0,277 y + 2,99 años, sin poder concluir que las madres del grupo "casos" ("*RN de bajo peso al nacer*") sean de menos edad que las madres del grupo "control" ("*RN de peso normal al nacer*").

Si exploramos dicha asociación mediante una regresión logística simple, introduciendo directamente la variable ***EDAD*** como independiente y no categórica, el resultado sería:

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1[a]	edad	-,051	,032	2,635	1	,105	,950
	Constante	,385	,732	,276	1	,599	1,469

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

La interpretación del coeficiente de regresión y la OR (Exp(B)) en este caso es "peculiar". De hecho el modelo ha calculado una OR para evaluar la asociación entre una variable categórica (***BAJOPESO***) y una variable contínua (***EDAD***), lo cual puede parecer un error. Y es que en realidad, lo que ha hecho el programa es calcular una OR promedio para cada cambio unitario en la variable independiente, esto es, *"0,950 es la OR o el riesgo de tener un RN de bajo peso que tiene una madre de una determinada edad comparada con una madre que tiene un año menos"*, asumiendo que este riesgo es constante para cada par de valores de edad considerados, al menos en el rango explorado en el estudio.

Por otra parte, como vemos en la salida de la RL, el contraste de hipótesis a través del estadístico de Wald no sale significativo ("***p***" asociada 0,105), una conclusión parecida a la obtenida en el test de comparación de medias antes realizado.

Y así seguiríamos explorando todas y cada una de las variables independientes o presumiblemente predictoras de la respuesta (***BAJOPESO***).[6] Al final deberíamos tener un cuadro resumen con las medidas de asociación encontradas y los valores de "***p***" en cada contraste:

Variable independiente categórica	Valor de "p" asociado al contraste	*OR*	$IC_{95\%}$ de la OR	
			INFERIOR	SUPERIOR
TABACO	**0,026**	**2,022**	1,081	3,783
HTA	**0,036** (corregida 0,076)	**3,365**	1,021	11,088
IU	**0,020**	**2,578**	1,139	5,834
RAZA (recodificada)	**0,029**	**2,005**	1,070	3,754
VISITAS (recodificada)	**0,104 (NS)**	**1,682**	0,896	3,157
TPP (recodificada)	**< 0,001**	**4,317**	1,916	9,726
Variable independiente numérica	**Valor de "p" asociado al contraste**	***Diferencia de medias***	$IC_{95\%}$ de la diferencia de medias	
			INFERIOR	SUPERIOR
EDAD	**0,103 (NS)**	**1,356 años**	-0,277	+2,99
PESOM	**0,02**	**5,1356 kg**	0,83	9,44

Debe recordarse, no obstante, que estas estimaciones puntuales e interválicas de las relaciones entre cada una de las variables independientes y la variable dependiente ***BAJOPESO*** pueden estar confundidas o modificadas por terceras variables. El análisis multivariante (analizándolas todas simultáneamente) permitirá controlar aquél (la confusión) y detectar éste (la interacción). Tener los valores de una medida de asociación como la OR "basal" para cada variable predictora nos servirá para valorar su papel y decidir si debe incluirse o no en la regresión logística multivariante.

Paso 2. Evaluar posibles interacciones o modificaciones de efecto y/o confusión.

Una de las mayores potencialidades de la RLM es la posibilidad de obtener medidas de asociación (OR) ajustadas o no confundidas, mucho más allá de lo que puede conseguirse con el análisis estratificado. De cualquier manera conviene repasar aquí los conceptos y procedimientos explicados en los documentos sobre **"Confusión e Interacción"**.

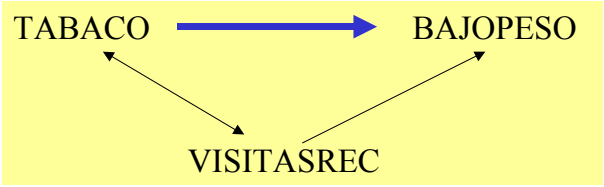
Veamos esto con algunos ejemplos.

A) Valoración de la confusión y/o interacción mediante el análisis estratificado.

En la base de datos del estudio que estamos empleando (BAJO PESO AL NACER), supongamos que se desea explorar el papel de la variable ***TABACO*** (hábito tabáquico en la gestante) sobre el desenlace ***BAJOPESO***, pero teniendo en cuenta la posible confusión que puede efectuar sobre esta asociación la variable ***VISITASREC*** (una recodificada que establece dos perfiles de seguimiento sanitario del embarazo: "*sin seguimiento*" y "*con al menos una visita médica*"). Se piensa que esta variable puede ser una predictora del efecto final (tener un RN de bajo peso), de forma que las madres sin control médico podrían tener

[6] Por supuesto ya habrá podido entenderse que no siempre será necesario recurrir a los contrastes de hipótesis bivariantes (Chi cuadrado, t de Student y ANOVA) y que cuando se tenga soltura en el manejo y en la interpretación de la Regresión Logística Binaria éste procedimiento sustituirá a los anteriores y será de elección.

mayor riesgo de dicho evento; y a su vez se hipotetiza sobre una posible asociación entre ambas en este estudio, de forma que tal vez las madres fumadoras tengan comportamientos menos saludables y menos preocupación por su embarazo.

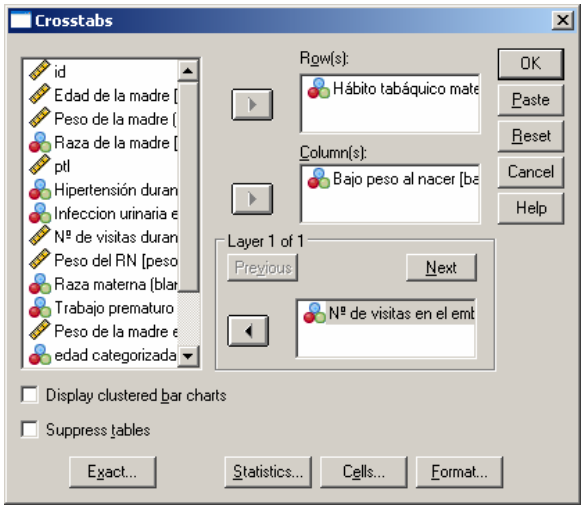


La posible relación entre las tres variables vendría expresado por el siguiente gráfico:

Si exploramos primero la relación principal ***TABACO*** → ***BAJOPESO*** mediante un análisis bivariante simple (a través de una tabla de contingencia 2x2) los resultados en SPSS son los que ya conocemos (ver antes):

$$X^2 = 4{,}924, \text{ con valor p} = 0{,}026$$
$$\text{OR}_{\text{(No Fuma/Fuma)}} = 2{,}022 \ (\text{IC}_{95\%}\ 1{,}081 \text{ a } 3{,}783)$$

Esto es, encontramos una asociación estadísticamente significativa (p=0,026) con el test Chi cuadrado y una medida de fuerza de asociación OR de 2,022 que identifica el hábito de fumar como un factor de riesgo de tener un RN de bajo peso (las madres fumadoras tendrían algo más del doble de riesgo que las no fumadoras de tener un RN de bajo peso).



Si ahora hacemos el mismo análisis pero estratificando por la variable ***VISITASREC***, que creemos puede ser un factor de confusión en la relación principal que se evalúa (***TABACO*** → ***BAJOPESO***), incluyéndola en la ventana correspondiente (**Layer 1 of 1** en el programa en inglés y **Capa 1 de 1** en el programa en castellano) de la opción Crosstabs (Tabla de contingencia) de SPSS, obtendremos el siguiente resultado:

Crosstabs

Hábito tabáquico materno * Bajo peso al nacer * Nº de visitas en el embarazo categorizada Crosstabulation

Nº de visitas en el embarazo categorizada				Bajo peso al nacer >=2500 gr	Bajo peso al nacer <2500 gr	Total
Al menos una visita médica	Hábito tabáquico materno	No Fuma	Count	47	13	60
			% within Hábito tabáquico materno	78,3%	21,7%	100,0%
		Fuma	Count	18	9	27
			% within Hábito tabáquico materno	66,7%	33,3%	100,0%
	Total		Count	65	22	87
			% within Hábito tabáquico materno	74,7%	25,3%	100,0%
Sin control médico	Hábito tabáquico materno	No Fuma	Count	39	16	55
			% within Hábito tabáquico materno	70,9%	29,1%	100,0%
		Fuma	Count	26	21	47
			% within Hábito tabáquico materno	55,3%	44,7%	100,0%
	Total		Count	65	37	102
			% within Hábito tabáquico materno	63,7%	36,3%	100,0%

Vemos que en la tabla hay ahora dos estratos, establecidos por las categorías de la variable de estratificación ***VISITASREC***; y en cada estrato se nos muestran los valores para la distribución 2x2 de ***TABACO*** y ***BAJOPESO***. El procedimiento sigue mostrándonos el contraste estadístico Chi cuadrado en cada estrato y los valores de asociación OR y sus intervalos de confianza, también en cada estrato.

Chi-Square Tests

Nº de visitas en el embarazo categorizada		Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Al menos una visita médica	Pearson Chi-Square	1,342[b]	1	,247		
	Continuity Correction[a]	,795	1	,373		
	Likelihood Ratio	1,301	1	,254		
	Fisher's Exact Test				,291	,185
	Linear-by-Linear Association	1,326	1	,250		
	N of Valid Cases	87				
Sin control médico	Pearson Chi-Square	2,665[c]	1	,103		
	Continuity Correction[a]	2,033	1	,154		
	Likelihood Ratio	2,667	1	,102		
	Fisher's Exact Test				,148	,077
	Linear-by-Linear Association	2,638	1	,104		
	N of Valid Cases	102				

a. Computed only for a 2x2 table

b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 6,83.

c. 0 cells (,0%) have expected count less than 5. The minimum expected count is 17,05.

Risk Estimate

Nº de visitas en el embarazo categorizada		Value	95% Confidence Interval	
			Lower	Upper
Al menos una visita médica	Odds Ratio for Hábito tabáquico materno (No Fuma / Fuma)	1,808	,659	4,957
	For cohort Bajo peso al nacer = >=2500 gr	1,175	,872	1,583
	For cohort Bajo peso al nacer = <2500 gr	,650	,317	1,333
	N of Valid Cases	87		
Sin control médico	Odds Ratio for Hábito tabáquico materno (No Fuma / Fuma)	1,969	,869	4,462
	For cohort Bajo peso al nacer = >=2500 gr	1,282	,942	1,744
	For cohort Bajo peso al nacer = <2500 gr	,651	,387	1,096
	N of Valid Cases	102		

Recuérdese que en el análisis estratificado puede intuirse que hay <u>confusión</u> por la tercera variable cuando las OR de los estratos son parecidas y a su vez diferentes a la OR global o cruda; y que puede haber <u>interacción</u> si las OR en cada estrato son muy diferentes, siendo la OR global o cruda un promedio de ambas. En nuestro caso ambas OR son parecidas (1,808 y 1,969) y ligeramente distintas a la OR de ***TABACO*** sin estratificar por ***VISITASREC*** (2,022).

Tests of Homogeneity of the Odds Ratio

	Chi-Squared	df	Asymp. Sig. (2-sided)
Breslow-Day	,017	1	,897
Tarone's	,017	1	,898

Tests of Conditional Independence

	Chi-Squared	df	Asymp. Sig. (2-sided)
Cochran's	3,999	1	,046
Mantel-Haenszel	3,337	1	,068

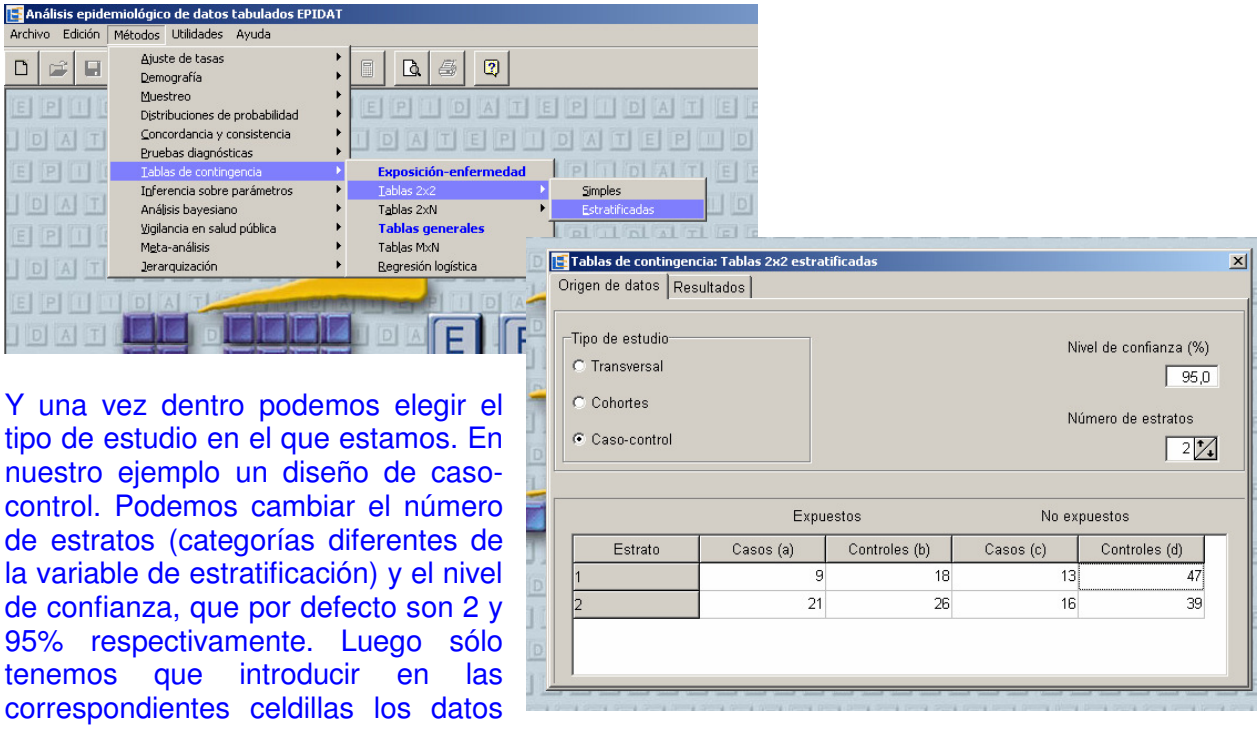
Under the conditional independence assumption, Cochran's statistic is asymptotically distributed as a 1 df chi-squared distribution, only if the number of strata is fixed, while the Mantel-Haenszel statistic is always asymptotically distributed as a 1 df chi-squared distribution. Note that the continuity correction is removed from the Mantel-Haenszel statistic when the sum of the differences between the observed and the expected is 0.

Mantel-Haenszel Common Odds Ratio Estimate

Estimate			1,905
ln(Estimate)			,644
Std. Error of ln(Estimate)			,324
Asymp. Sig. (2-sided)			,047
Asymp. 95% Confidence Interval	Common Odds Ratio	Lower Bound	1,009
		Upper Bound	3,594
	ln(Common Odds Ratio)	Lower Bound	,009
		Upper Bound	1,279

The Mantel-Haenszel common odds ratio estimate is asymptotically normally distributed under the common odds ratio of 1,000 assumption. So is the natural log of the estimate.

Si lo hacemos con el programa **EPIDAT 3.1**, (con los datos agregados) el procedimiento a seguir sería seleccionar en **Métodos** la secuencia **Tablas de contingencia** > **Tablas 2x2** > **Estratificadas**:



Y una vez dentro podemos elegir el tipo de estudio en el que estamos. En nuestro ejemplo un diseño de caso-control. Podemos cambiar el número de estratos (categorías diferentes de la variable de estratificación) y el nivel de confianza, que por defecto son 2 y 95% respectivamente. Luego sólo tenemos que introducir en las correspondientes celdillas los datos agrupados o tabulados que ya conocemos por haberlo llevado a cabo en el programa SPSS. Y finalmente oprimir la tecla de resultados en la barra superior, para que estos aparezcan en la ventana correspondiente:

```
Tipo de estudio   : Caso-control
Número de estratos: 2
Nivel de confianza: 95,0%

Tabla global          Casos      Controles        Total
---------------     ----------   ----------   ----------
Expuestos                30           44           74
No expuestos             29           86          115
---------------     ----------   ----------   ----------
Total                    59          130          189

ODDS RATIO (OR)
Estrato                  OR          IC(95,0%)
---------------     ----------   ----------   ----------
1                  1,807692      0,659253     4,956748    (Woolf)
2                  1,968750      0,868761     4,461500    (Woolf)
---------------     ----------   ----------   ----------
Cruda              2,021944      1,080660     3,783112    (Woolf)
```

```
Combinada (M-H)       1,904745     1,009406     3,594246
Ponderada             1,903196     1,008183     3,592754

Prueba de homogeneidad
                      Ji-cuadrado          gl      Valor p
---------------     ---------------     ----------   ----------
Combinada (M-H)              0,0166              1       0,8975
Ponderada                    0,0166              1       0,8975

PRUEBA DE ASOCIACIÓN DE MANTEL-HAENSZEL
    Ji-cuadrado                  gl        Valor p
---------------     ---------------     ----------
         3,9573                   1         0,0467
```

Como puede verse los resultados son muy parecidos o idénticos a los obtenidos con el programa SPSS.

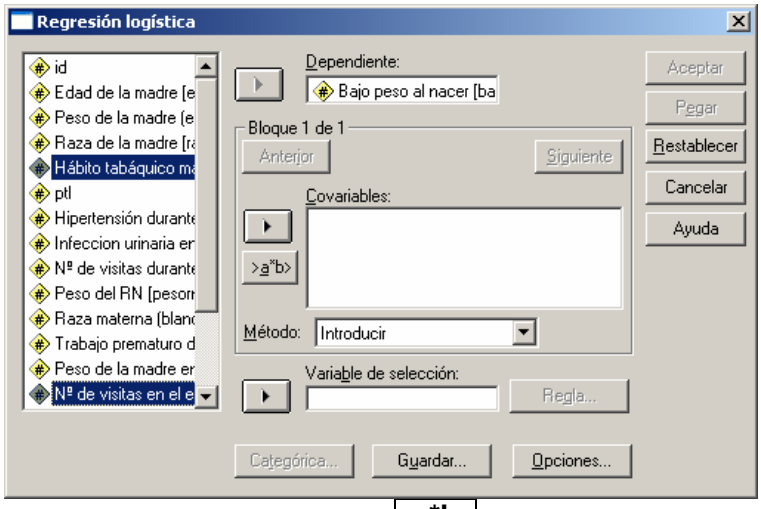
B) Valoración de la confusión y/o interacción mediante el análisis multivariante.

En el caso de evaluar esta asociación mediante un análisis Multivariante de Regresión Logística, es necesario desglosar los dos fenómenos (confusión e interacción) y comprobarlos *por separado*. Recuerde que...

- **La confusión se detecta cuando la OR** que evalúa la fuerza de asociación entre la V. Independiente y la V. Dependiente ***cambia de forma importante*** cuando se introduce en la ecuación de RLM la tercera variable.

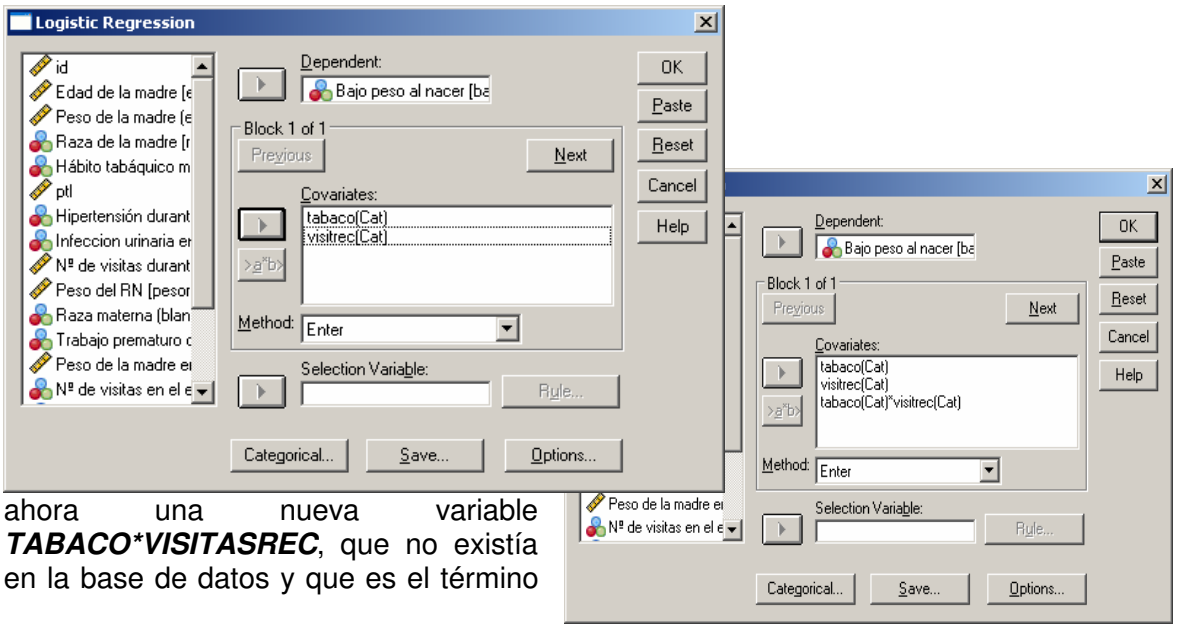
- **La interacción requiere introducir en la ecuación de RLM un término multiplicativo**, compuesto por las dos variables independientes que se presuponen interactúan en su efecto sobre la V. Dependiente; y una vez incluído ver si su coeficiente de regresión logística (B) es estadísticamente significativo (esto es, tiene un valor diferente al cero).

En ambos fenómenos deben emplearse en el programa SPSS los procedimientos manuales (no automáticos) de RLM (***Enter*** o ***Introducir***). Y **en la interacción deben seleccionarse juntas las dos variables predictoras o independientes**, para lo cual una vez señalada la primera (no importa el orden) se pulsa la tecla ***CONTROL*** y sin dejar de pulsarla se selecciona con el ratón la segunda variable en la ventana correspondiente. Cuando las dos aparecen sombreadas (están preseleccionadas) veremos que se activa un botón que hasta ese momento estaba desdibujado, justo debajo de la flecha de selección de covariables, y que tiene el siguiente rótulo:



>a*b>

En ese momento es cuando puede "*cliquearse*" con el ratón dicha pestaña, para llevar el nuevo término (de interacción) a la ventana de covariables. En dicha ventana aparecerá



ahora una nueva variable ***TABACO*VISITASREC***, que no existía en la base de datos y que es el término de interacción que se desea evaluar.

Por otra parte, para cumplir con el "**principio jerárquico**", los componentes elementales del término de interacción (en nuestro caso las variables ***TABACO*** y ***VISITASREC***) deben entrar a formar parte de la ecuación de RLM, por lo que deben seleccionarse en la ventana de variables y traspasarlas a la ventana de covariables. Recuerde que éstas deben ser categóricas (y así debe señalarse) y nuestra recomendación es que sean dicotómicas, ya que los términos de interacción con variables policotómicas (al igual que los términos de interacción de más de dos variables) son difíciles de interpretar.

Cuando el investigador quiera explorar tanto confusión como modificación de efecto (interacción) en una asociación –y podrían darse ambos fenómenos simultáneamente- debe proceder primero evaluando si existe interacción entre las covariables y, detectada o descartada ésta, valorar si había confusión.

Así, en nuestro ejemplo, si se seleccionan ***TABACO*** y ***VISITASREC***, introduciéndolas juntas en el análisis (***TABACO*VISITASREC***) y sus términos simples (por el principio jerárquico) ***TABACO*** y ***VISITASREC***, y se señala que ambas son categóricas y que la categoría de referencia debe cambiarse (*change*) por la primera (*first*), el resultado final de las variables incluidas en la ecuación de RL es:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	tabaco(1)	,592	,515	1,323	1	,250	1,808	,659	4,957
	visitrec(1)	,394	,432	,834	1	,361	1,483	,636	3,457
	tabaco(1) by visitrec(1)	,085	,663	,017	1	,898	1,089	,297	3,991
	Constant	-1,285	,313	16,820	1	,000	,277		

a. Variable(s) entered on step 1: tabaco, visitrec, tabaco * visitrec .

Como vemos, el término de interacción no es significativo (p=0,898), por lo que podemos descartar[7] que exista modificación de efecto. En este caso tiene sentido entonces **explorar en segundo lugar si hay confusión**, sacando de la ventana de covariables el término de interacción y dejando las variables ***TABACO*** y ***VISITASREC***:

[7] Se dice que hay interacción cuando el término de interacción tiene un coeficiente de regresión cuyo contraste de hipótesis tiene un valor ***p*** significativo, en general < 0,1

El resultado final es:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	tabaco(1)	,644	,324	3,951	1	,047	1,903	1,009	3,590
	visitrec(1)	,431	,327	1,730	1	,188	1,538	,810	2,921
	Constant	-1,304	,277	22,16	1	,000	,271		

a. Variable(s) entered on step 1: tabaco, visitrec.

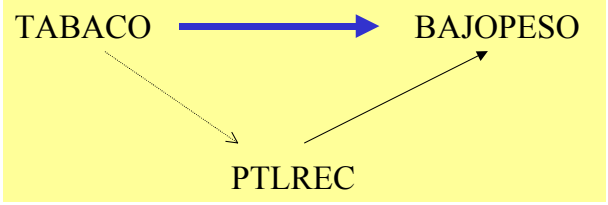
Como puede apreciarse el coeficiente de regresión de la variable ***TABACO*** es estadísticamente significativo (p=0,047) y su OR es 1,903. Esta es la OR ajustada por la variable ***VISITASREC***, dando un valor idéntico a la OR ponderada (y sus $IC_{95\%}$) que se obtuvo en el análisis estratificado (*vide supra*).

La decisión de dejar esta variable en la ecuación y obtener un estimador diferente y menos sesgado de la asociación entre ***TABACO*** y ***BAJOPESO*** depende del investigador y no hay test estadístico para ayudarnos. Sin embargo podemos dar algunas recomendaciones para apoyar su inclusión. ***En general, debe quedarse aquella variable que***:

- ***al introducirse cambia la OR cruda de la variable principal de forma "importante" (la mayoría de los autores dice que en al menos un 10%)***
- ***no modifica sustancialmente los intervalos de confianza de la OR de la variable principal, en el sentido de aumentarlos (lo que haría más imprecisa la estimación)***
- ***no cambia la significación estadística del contraste de Wald para el coeficiente de regresión de la variable principal***
- ***ella misma tiene un coeficiente de regresión significativo y una OR con un $IC_{95\%}$ que no contiene el valor nulo (=1), ya que en ese caso sería predictora del desenlace evaluado***

En este ejemplo el cambio de la OR de ***TABACO*** que se ha generado al introducir la variable ***VISITASREC*** no ha cambiado mucho (ha pasado de 2,022 a 1,903 y esto es un 5% aproximadamente) y su coeficiente de regresión (B=0,431) no es significativo, por lo que bajo estas solas premisas no hay argumentos para dejarla en la ecuación.

Veamos ahora qué ocurre con otra variable que está reconocida como un factor de riesgo de bajo peso al nacer (***PTLREC***, que es una categórica que distingue si hay antecedentes o no de trabajo prematuro de parto) y que podría estar asociada con el hábito tabáquico materno durante la gestación (medido en nuestro estudio con la variable ***TABACO***); es más, pudiera tratarse de una variable intermedia en la cadena etiológica principal que estamos evaluando...



Mostramos la salida de la Regresión Logística Binaria directamente, introduciendo las dos covariables (y tras comprobar que no existe interacción entre ellas):

Variables in the Equation

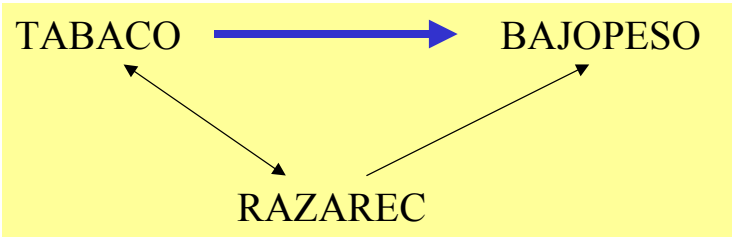
		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	ptlrecod(1)	1,352	,421	10,284	1	,001	3,864	1,691	8,827
	tabaco(1)	,536	,334	2,588	1	,108	1,710	,889	3,288
	Constant	-1,262	,228	30,611	1	,000	,283		

a. Variable(s) entered on step 1: ptlrecod, tabaco.

Vemos como en este caso se modifica la OR de ***TABACO*** hasta un valor de 1,71 (un descenso del 15%) pero además el coeficiente de regresión de dicha variable se hace no significativo (p=0,108) y los $IC_{95\%}$ de la OR contienen el valor nulo (van de 0,889 a 3,288). Desde luego hemos descubierto a la vez que la variable ***PTLREC*** es una predictora de ***BAJOPESO***, puesto que la asocia a ella una OR -ajustada por la variable ***TABACO-*** de 3,864 y con un coeficiente de regresión estadísticamente significativo, y por ello, una variable que debería quedar en la ecuación de regresión multivariante.

Diríamos entonces que la verdadera OR de ***TABACO*** para el desenlace predicho (***BAJOPESO***=1) es de 1,7 y este sería el riesgo que tienen las madres que fuman frente a las que no fuman de tener un RN de peso bajo, independientemente de que tengan o no trabajo prematuro de parto.

Exploremos ahora otra variable que podría causar confusión en la relación principal que se evalúa, el factor racial que en nuestro estudio se ha medido con la variable ***RAZAREC*** (una categórica dicotómica que agrupa por una parte a las madres de raza "*blanca*" –categoría "0" y por otro a las madres que "*no son de raza blanca*" –categoría "1"-). El esquema sería:



Veamos la salida del procedimientos Crosstabs de SPSS para evaluar la relación simple entre ***RAZAREC*** y ***BAJOPESO***, a fin de comprobar si es una variable asociada a la dependiente:

Crosstabs

Raza materna (blanca vs otras) * Bajo peso al nacer Crosstabulation

			Bajo peso al nacer		
			>=2500 gr	<2500 gr	Total
Raza materna (blanca vs otras)	blanca	Count	73	23	96
		% within Raza materna (blanca vs otras)	76,0%	24,0%	100,0%
	otras	Count	57	36	93
		% within Raza materna (blanca vs otras)	61,3%	38,7%	100,0%
Total		Count	130	59	189
		% within Raza materna (blanca vs otras)	68,8%	31,2%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	4,787[b]	1	,029		
Continuity Correction [a]	4,125	1	,042		
Likelihood Ratio	4,815	1	,028		
Fisher's Exact Test				,041	,021
Linear-by-Linear Association	4,762	1	,029		
N of Valid Cases	189				

a. Computed only for a 2x2 table

b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 29,03.

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Raza materna (blanca vs otras) (blanca / otras)	2,005	1,070	3,754
For cohort Bajo peso al nacer = >=2500 gr	1,241	1,019	1,510
For cohort Bajo peso al nacer = <2500 gr	,619	,399	,960
N of Valid Cases	189		

Todo hace indicar que sí: el contraste Chi cuadrado es significativo (p=0,029) y la OR es 2,005 que traduce que las mujeres de "*otras razas*", en comparación con las madres de raza "*blanca*", tienen un riesgo de tener un RN de bajo peso del doble. Veamos que pasa cuando ambas se introducen juntas en la RLM:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	tabaco(1)	1,113	,364	9,337	1	,002	3,044	1,490	6,215
	razarec(1)	1,100	,364	9,114	1	,003	3,005	1,471	6,139
	Constant	-1,839	,350	27,546	1	,000	,159		

a. Variable(s) entered on step 1: tabaco, razarec.

El resultado es sorprendente: crecen ambas OR sobre sus estimaciones crudas (recuérdense que eran 2,022 para ***TABACO*** y 2,005 para la variable ***RAZAREC***), en proporciones importantes (casi un 30% en ambos casos), permanecen ambas con coeficientes de regresión estadísticamente significativos y, la única "pega", es que aumentan los $IC_{95\%}$ de las OR, que traducen una estimaciones más imprecisas. De cualquier manera estas OR son mejores que las crudas y, desde luego, podemos decir que la variable ***RAZAREC*** confunde la relación entre ***TABACO*** y ***BAJOPESO***: de hecho, las mujeres que fuman tienen el triple de riesgo (y no el doble) de las que no fuman de tener un RN de bajo peso, independientemente de la raza materna.

Cuando esto ocurre nunca debemos de dejar de explorar tanto la asociación entre las variables predictoras como una posible interacción entre ellas en su acción sobre el efecto que se evalúa (***BAJOPESO***).

Veamos primero si existe asociación entre ***TABACO*** y ***RAZAREC*** mediante un análisis simple de Tabla de Contingencia 2x2. Da igual cuál de ellas se coloca en la columnas y cual en las filas:

Hábito tabáquico materno * Raza materna (blanca vs otras) Crosstabulation

			Raza materna (blanca vs otras)		
			blanca	otras	Total
Hábito tabáquico materno	No Fuma	Count	44	71	115
		% within Hábito tabáquico materno	38,3%	61,7%	100,0%
	Fuma	Count	52	22	74
		% within Hábito tabáquico materno	70,3%	29,7%	100,0%
Total		Count	96	93	189
		% within Hábito tabáquico materno	50,8%	49,2%	100,0%

Como puede comprobarse a simple vista, en la muestra analizada existe una clara desproporción entre el porcentaje de madres fumadoras entre las de raza "*blanca*" (70,3%) y las de "*otras razas*" (29,7%); es decir, la mayoría de las madres fumadoras son "*blancas*" y la mayoría de las no fumadoras son de "*otras razas*", siendo estas diferentes proporciones estadísticamente significativas (p < 0,001), según el contraste Chi cuadrado:

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	18,458[b]	1	,000		
Continuity Correction[a]	17,200	1	,000		
Likelihood Ratio	18,870	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	18,361	1	,000		
N of Valid Cases	189				

a. Computed only for a 2x2 table

b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 36,41.

Y la medida de asociación OR es de 0,26 (o alternativamente su inverso 3,85).

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Hábito tabáquico materno (No Fuma / Fuma)	,262	,140	,490
For cohort Raza materna (blanca vs otras) = blanca	,544	,413	,717
For cohort Raza materna (blanca vs otras) = otras	2,077	1,422	3,033
N of Valid Cases	189		

A la vista de todo ello se debe pensar que ambas covariables producen confusión en este estudio, sin poder descartar que además generen modificación de efecto (interacción) sobre la relación principal que se evalúa.

A continuación se muestra la salida de SPSS con el término de interacción ***TABACO*RAZAREC***:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	tabaco(1)	1,751	,598	8,561	1	,003	5,758	1,782	18,599
	razarec(1)	1,693	,580	8,510	1	,004	5,435	1,743	16,948
	razarec(1) by tabaco(1	-1,141	,776	2,163	1	,141	,320	,070	1,461
	Constant	-2,303	,524	19,280	1	,000	,100		

a. Variable(s) entered on step 1: tabaco, razarec, razarec * tabaco .

Puede apreciarse cómo, a pesar de no ser estadísticamente significativo, el término de interacción introduce nueva información en la estimación de la OR para las variables

TABACO y ***RAZAREC*** en relación a ***BAJOPESO***, que aún se incrementan más (pasan a ser de 5,758 y de 5,435 respectivamente), aunque lógicamente a costa de que dichas estimaciones se hagan con más imprecisión, traducido por unos mayores errores estándar (S.E.) de los coeficientes de regresión y unos más amplios $IC_{95\%}$ de las OR. Dejar o no el término de interacción en la ecuación de RLM es una decisión que deben tomar los investigadores y dependerá del objetivo del estudio y de la consistencia del marco teórico que soporta dichas relaciones. Si la decisión la dejamos al arbitrio "estadístico", el término de interacción debería ser removido, puesto que no es estadísticamente significativo (p=0,141).

A efectos puramente académicos resumimos la interpretación del riesgo en caso de que exista interacción entre dos variables sobre un resultado esperado. Si hubiésemos considerado que hay interacción entre ***TABACO*** y ***RAZAREC*** en su relación o asociación sobre ***BAJOPESO***, y que los coeficientes de regresión B y sus correspondientes OR son los obtenidos en la ecuación de regresión logística anterior, las estimaciones ajustadas serían las que se recogen en el siguiente cuadro:

Riesgo de tener un RN de bajo peso (BAJOPESO = "1")		***TABACO***	
		Sí ("1")	NO ("0")
RAZAREC	NO BLANCA ("1")	**0,32**	**5,435**
	BLANCA ("0")	**5,758**	**1**

En efecto, el riesgo basal o de referencia (la unidad) sería el que tienen las madres "*no fumadoras*" de "*raza blanca*" y 0,32 (factor protector) sería el riesgo de tener un RN de bajo peso al nacer cuando coinciden el factor raza ("*no ser de raza blanca*") y el factor tabaco ("*fumar*").

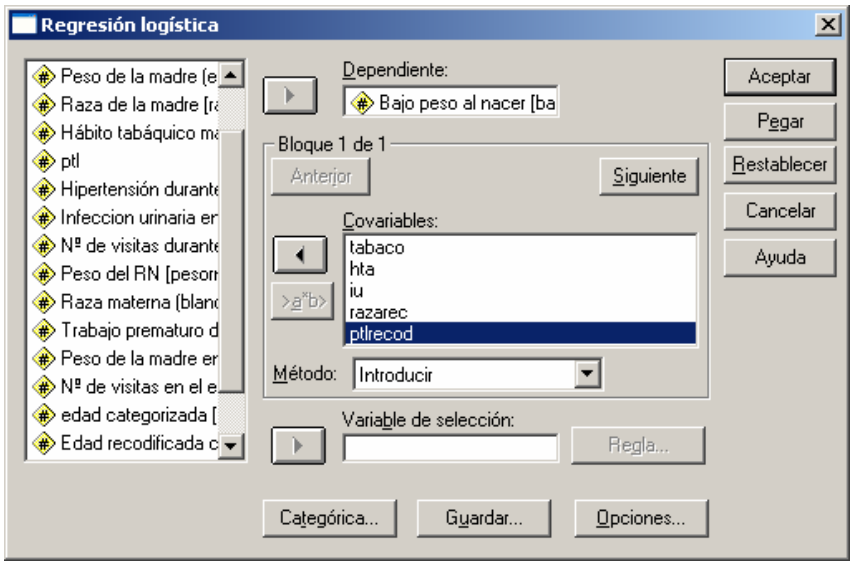
Paso 3. Construir un modelo o ecuación de regresión logística multivariante.

En este paso es especialmente importante saber qué deseamos obtener bajo la óptica del objetivo de nuestro estudio, y aconsejamos releer el "**Paso 0**" dentro de este mismo documento, puesto que el procedimiento que utilizaremos en el análisis podrá variar en función de dicho objetivo.

Si queremos "llevar la batuta" y evaluar factores confundentes y términos de interacción tenemos que recurrir obligatoriamente a un análisis multivariante con el procedimiento ENTER (INTRODUCIR en la versión en castellano del SPSS) incluyendo en el modelo la variable independiente principal que quiere controlarse y las posibles confundentes; o las variables que mostraron significación estadística en su asociación univariante con la variable resultado ***BAJOPESO***, seleccionándolas de la ventana de VARIABLES e introduciéndolas en la ventana de COVARIABLES. Recordemos:

TABACO: hábito tabáquico materno durante el embarazo. Categoría expuesta "SI".
HTA: hipertensión arterial durante el embarazo. Categoría expuesta "SI".
IU: Infección urinaria durante el embarazo. Categoría expuesta "SI".
RAZA (recodif.): Ser de raza blanca o de otra raza. Categoría expuesta "OTRAS".
PTL (recodif.): Tener historia o no de trabajo prematuro de parto. Categoría expuesta "SI".
PESOMKG: Peso materno en kg al comienzo de la gestación. Variable cuantitativa.

Incluso a la hora de elegir las covariables no debemos seguir un criterio "purista" desde el punto de vista estadístico, incluyendo sólo aquellas que en el contraste univariante tenían significación estadística (p<0,05). Los investigadores pueden decidir cuáles de entre ellas tienen sentido clínico o epidemiológico que justifiquen su selección; e incluso pueden probar a incluir variables con valor "***p***" asociado al contraste univariante no significativo (0,05 < p < 0,25) siempre que otros estudios las consideren predictoras, confundentes o modificadoras de efecto.

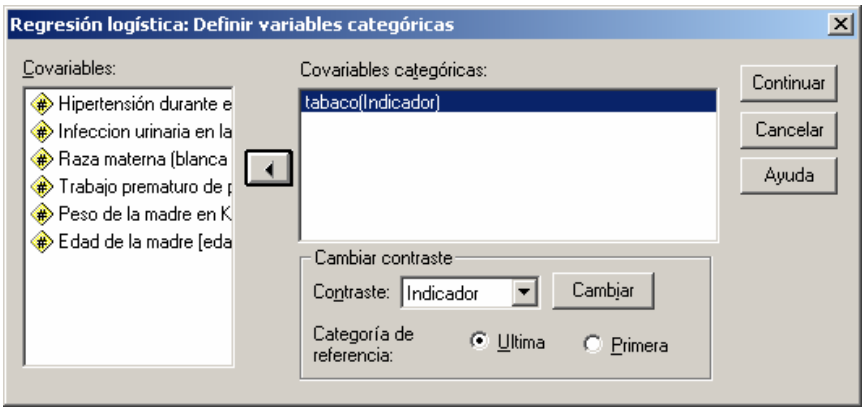


En cuanto al número de variables seleccionadas existen límites que vienen impuestos fundamentalmente por el tamaño muestral y por la existencia de un mínimo de valores en las categorías a riesgo de cada covariable introducida.

Así, en lo que respecta a la primera restricción, hay una regla que aconseja "*no seleccionar más que una variable por cada diez individuos estudiados con el desenlace que se quiere modelizar*" (en nuestro ejemplo, sólo hay 59 individuos con "*bajo peso al nacer*", por lo que parece razonable no incluir en el modelo final más de seis variables independientes. Una regla alternativa es la del "20", que dice "*no se deben seleccionar más de una variable por cada veinte individuos analizados en la base de datos*", asumiendo que están balanceados los que tienen y los que no tienen el desenlace o efecto (en nuestro ejemplo son 189 individuos pero no están "balanceados", ya que hay 59 casos y 130 controles).

Igualmente, no deberían incluirse en los modelos variables con pocos valores en la categoría en riesgo o expuesta (< 10) y nunca si en dicha categoría o estrato no hay ningún valor. Si la variable tuviese más de dos categorías, una solución sería colapsar o agrupar categorías.

Una vez seleccionado un plantel de variables, aquellas que son categóricas deben identificarse como tales en el programa, seleccionándolas tras pulsar sobre la pestaña "**Categórica...**" y cumplimentando para cada una la CATEGORÍA DE REFERENCIA



(recomendamos que sea la primera, a la que habitualmente hemos codificado como "cero" y es la que tiene menor riesgo) y el procedimiento de contraste (INDICATOR por defecto). Recuérdese que para cambiar la categoría de referencia debe finalizarse oprimiendo el botón **Cambiar**.

En la salida de SPSS de la Regresión logística binaria no condicional comprobamos que el programa ha identificado correctamente las categorías expuestas (1) y las de referencia (0) al introducir las variables categóricas en el modelo:

Codificaciones de variables categóricas

		Frecuencia	Codificación de (1)
Trabajo prematuro de parto	NO	159	,000
	SI	30	**1,000**
raza materna (blanca vs otras)	blanca	96	,000
	otras	93	**1,000**
Hipertensión durante el embarazo	No hipertensa	177	,000
	Hipertensa	12	**1,000**
Infeccion urinaria	No	161	,000
	Si	28	**1,000**
Hábito tabáquico materno	No Fuma	115	,000
	Fuma	74	**1,000**

Y el resultado de la regresión logística es:

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95,0% para EXP(B) Inferior	Superior
Paso 1[a]	tabaco(1)	,928	,399	5,421	1	,020	2,530	1,158	5,527
	razarec(1)	1,013	,396	6,542	1	,011	2,754	1,267	5,985
	hta(1)	1,852	,706	6,886	1	,009	6,374	1,598	25,423
	iu(1)	,735	,462	2,537	1	,111	2,086	,844	5,157
	pesomkg	-,031	,014	4,684	1	,030	,969	,943	,997
	ptlrecod(1)	1,120	,451	6,171	1	,013	3,065	1,267	7,417
	Constante	-,372	,916	,165	1	,684	,689		

a. Variable(s) introducida(s) en el paso 1: tabaco, razarec, hta, iu, pesomkg, ptlrecod.

La mayoría de las covariables incluidas en el modelo mantienen la significación estadística ($p<0,05$) en el contraste de hipótesis que las relaciona con la variable dependiente ***BAJOPESO***, salvo la variable ***IU*** (infección urinaria durante la gestación) que arroja un valor "**p**" de 0,11. Un procedimiento automático ***hacia atrás*** habría excluído dicha variable del modelo final.

Sobre los valores de OR obtenidos podemos comprobar que son similares a los calculados en el análisis bivariante, aunque algo mayores para las variables ***TABACO*** (pasa de 2,022 a 2,53), ***RAZAREC*** (pasa de 2,05 a 2,75), ***HTA*** (pasa de 3,36 a 6,37), y algo menores para las variables ***IU*** (pasa de 2,58 a 2,086) y ***PTLRECOD*** (pasa de 4,32 a 3,065). Estos valores de OR están ajustados para cada variable y representan una estimación de su fuerza de asociación con ***BAJOPESO*** controladas todas las demás variables incluidas en el modelo. Sin embargo su cálculo conlleva una mayor imprecisión en las estimaciones, como puede comprobarse viendo cómo los intervalos de confianza de las OR se han "separado" y los errores estándar de los coeficientes de regresión han aumentado.

Con estas seis variables, el modelo tiene una capacidad de clasificar correctamente al 75% de los casos analizados, como puede verse en la tabla siguiente, aunque clasifica "mejor" el peso normal (≥ 2.500 g) que el peso bajo (< 2.500 g):

Tabla de clasificación[a]

	Observado		Pronosticado: Bajo peso al nacer >=2500 gr	Pronosticado: Bajo peso al nacer <2500 gr	Porcentaje correcto
Paso 1	Bajo peso al nacer	>=2500 gr	117	13	90,0
		<2500 gr	35	24	40,7
	Porcentaje global				74,6

a. El valor de corte es ,500

Y la proporción de la variabilidad de ***BAJOPESO*** que es explicado por este modelo no es demasiado buena (entre un 17,4% -ver R cuadrado de Cox y Snell- y un 24,5% -según el R cuadrado de Nagelkerke-), esto es, sigue existiendo un porcentaje importante de "influencia" sobre el hecho de tener un RN de bajo peso que no depende de las variables analizadas:

Resumen de los modelos

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	198,516[a]	,174	,245

a. La estimación ha finalizado en el número de iteración 5 porque las estimaciones de los parámetros han cambiado en menos de ,001.

Con estos resultados sería "arriesgado" proponer un MODELO DE PREDICCIÓN como conclusión en este estudio, aunque con carácter puramente académico puede elaborarse la ecuación de regresión logística siguiente:

$$Logit\ (p) = -0{,}372 + 0{,}928*(TABACO) + 1{,}013*(RAZAREC) + 1{,}852*(HTA) + 0{,}735*(IU) + 1{,}120*(PTLRECOD) - 0{,}031*(PESOMKG)$$

Siendo Logit (p) = ln (p / (1-p) = ln (odds)

Y donde p = P (Y = 1), esto es, la probabilidad de que un individuo tenga la característica evaluada; en nuestro ejercicio que sea un "*RN de bajo peso*", puesto que el valor "1" se asigna en la base de datos a los RN con peso inferior a 2.500 g.

O alternativamente:

$$P(BAJOPESO=1) = \frac{1}{1+e^{-(-0{,}372 + 0{,}928*TABACO + 1{,}013*RAZAREC + 1{,}852*HTA + 0{,}735*IU + 1{,}120*PTLREC - 0{,}031*PESOKG)}}$$

Si hubiésemos dejado al programa SPSS que hiciera automáticamente la regresión, mediante el procedimiento PASOS HACIA DELANTE (criterio E. Wald), tras introducir todas las variables independientes medidas en el estudio, el resultado final hubiera sido en el tercer paso:

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95,0% para EXP(B)	
								Inferior	Superior
Paso 3[c]	hta(1)	1,894	,721	6,899	1	,009	6,646	1,617	27,311
	pesomkg	-,038	,015	6,482	1	,011	,963	,936	,991
	ptlrecod(1)	1,407	,429	10,778	1	,001	4,083	1,763	9,456
	Constante	1,017	,853	1,421	1	,233	2,766		

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

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

c. Variable(s) introducida(s) en el paso 3: pesomkg.

Variables que no están en la ecuación

			Puntuación	gl	Sig.
Paso 3	Variables	tabaco(1)	2,203	1	,138
		razarec(1)	3,393	1	,065
		iu(1)	2,710	1	,100
		visitrec(1)	2,351	1	,125
		edad	2,588	1	,108
	Estadísticos globales		12,308	5	,031

Sólo se quedaría con tres variables explicativas (***HTA***, ***PTLREC*** y ***PESOKG***), y el modelo final no es mejor (en cuanto a valor predictivo) que el obtenido por el procedimiento manual INTRODUCIR.

Si hubiésemos elegido el procedimiento automático **HACIA ATRÁS** (BACKWARD), método **E. Wald**, con todas las variables independientes medidas en el estudio, el resultado en el programa RL de SPSS hubiese sido:

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

Primero un resumen de las covariables categóricas y sus codificaciones:

Codificaciones de variables categóricas

		Frecuencia	Codificación de parámetros (1)
Nº de visitas en el embarazo categorizada	Al menos una visita médica	87	,000
	Sin control médico	102	**1,000**
Hipertensión durante el embarazo	No hipertensa	177	,000
	Hipertensa	12	**1,000**
Infeccion urinaria en la gestación	No	161	,000
	Si	28	**1,000**
Raza materna (blanca vs otras)	blanca	96	,000
	otras	93	**1,000**
Trabajo prematuro de parto	NO	159	,000
	SI	30	**1,000**
Hábito tabáquico materno	No Fuma	115	,000
	Fuma	74	**1,000**

Pruebas omnibus sobre los coeficientes del modelo

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	37,489	8	,000
	Bloque	37,489	8	,000
	Modelo	37,489	8	,000
Paso 2[a]	Paso	-,208	1	,648
	Bloque	37,281	7	,000
	Modelo	37,281	7	,000
Paso 3[a]	Paso	-1,125	1	,289
	Bloque	36,156	6	,000
	Modelo	36,156	6	,000
Paso 4[a]	Paso	-2,498	1	,114
	Bloque	33,658	5	,000
	Modelo	33,658	5	,000

a. Un valor de chi-cuadrado negativo indica que ha disminuido el valor de chi-cuadrado con respecto al paso anterior.

El proceso iterativo termina en el cuarto paso, con un modelo que es significativamente mejor que el modelo con todas las variables (véase como han ido disminuyendo los grados de libertad (gl) del contraste Chi cuadrado de la Prueba Omnibus sobre el modelo, desde 8 en el primer paso (9 coeficientes, el de la constante y el de las ocho variables iniciales) hasta 5 en el 4º paso (se habrá quedado con 5 variables más la constante en la ecuación final de regresión). Seguidamente nos aporta información sobre los ajustes de los diferentes modelos que ha ido generando en cada paso, tanto a través del estadístico [–2 log verosimilitud] como el R cuadrado.

Resumen de los modelos

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	197,183[a]	,180	,253
2	197,391[a]	,179	,252
3	198,516[a]	,174	,245
4	201,014[a]	,163	,229

a. La estimación ha finalizado en el número de iteración 5 porque las estimaciones de los parámetros han cambiado en menos de ,001.

Luego nos muestra un resumen de las tablas de clasificación que se han ido obteniendo con los valores predichos por el modelo en cada paso frente a los valores realmente observados. La mejoría es escasa en el % de predicción correcto:

Tabla de clasificación[a]

			Pronosticado		
			Bajo peso al nacer		Porcentaje correcto
	Observado		>=2500 gr	<2500 gr	
Paso 1	Bajo peso al nacer	>=2500 gr	115	15	88,5
		<2500 gr	36	23	39,0
	Porcentaje global				73,0
Paso 2	Bajo peso al nacer	>=2500 gr	116	14	89,2
		<2500 gr	39	20	33,9
	Porcentaje global				72,0
Paso 3	Bajo peso al nacer	>=2500 gr	117	13	90,0
		<2500 gr	35	24	40,7
	Porcentaje global				74,6
Paso 4	Bajo peso al nacer	>=2500 gr	119	11	91,5
		<2500 gr	35	24	40,7
	Porcentaje global				75,7

a. El valor de corte es ,500

Y por último un resumen de los coeficientes y su significación estadística de las diferentes covariables que ha ido considerando en cada paso, extrayéndo en cada uno la menos signicativa, para volver a ajustar el modelo sin ella. Así puede comprobarse como tras el primer paso elimina la covariable ***VISITREC*** que es la que tiene una "***p***" asociada al contraste de Wald no significativo y de mayor valor (0,649).

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp (B)	I.C. 95,0% para EXP(B)	
								Inferior	Superior
Paso 1[a]	edad	-,037	,038	,951	1	,330	,964	,894	1,038
	tabaco(1)	,860	,414	4,312	1	,038	2,363	1,049	5,320
	hta(1)	1,798	,704	6,526	1	,011	6,037	1,520	23,983
	iu(1)	,681	,468	2,118	1	,146	1,976	,790	4,944
	razarec(1)	,919	,410	5,035	1	,025	2,507	1,123	5,594
	ptlrecod(1)	1,244	,468	7,055	1	,008	3,469	1,385	8,685
	pesomkg	-,029	,014	3,875	1	,049	,972	,945	1,000
	visitrec(1)	,170	,373	,208	1	,649	1,185	,571	2,460
	Constante	,303	1,228	,061	1	,805	1,353		
Paso 2[a]	edad	-,039	,038	1,099	1	,295	,961	,893	1,035
	tabaco(1)	,904	,402	5,048	1	,025	2,469	1,122	5,432
	hta(1)	1,819	,702	6,718	1	,010	6,163	1,558	24,379
	iu(1)	,693	,466	2,210	1	,137	1,999	,802	4,985
	razarec(1)	,951	,403	5,570	1	,018	2,589	1,175	5,705
	ptlrecod(1)	1,219	,464	6,904	1	,009	3,384	1,363	8,403
	pesomkg	-,029	,014	3,989	1	,046	,972	,944	,999
	Constante	,435	1,191	,134	1	,715	1,546		
Paso 3[a]	tabaco(1)	,928	,399	5,421	1	,020	2,530	1,158	5,527
	hta(1)	1,852	,706	6,886	1	,009	6,374	1,598	25,423
	iu(1)	,735	,462	2,537	1	,111	2,086	,844	5,157
	razarec(1)	1,013	,396	6,542	1	,011	2,754	1,267	5,985
	ptlrecod(1)	1,120	,451	6,171	1	,013	3,065	1,267	7,417
	pesomkg	-,031	,014	4,684	1	,030	,969	,943	,997
	Constante	-,372	,916	,165	1	,684	,689		
Paso 4[a]	**tabaco(1)**	,930	,395	5,542	1	,019	2,534	1,168	5,495
	hta(1)	1,763	,708	6,202	1	,013	5,830	1,456	23,351
	razarec(1)	1,003	,392	6,545	1	,011	2,725	1,264	5,875
	ptlrecod(1)	1,222	,446	7,500	1	,006	3,393	1,415	8,133
	pesomkg	-,033	,014	5,359	1	,021	,967	,940	,995
	Constante	-,120	,907	,017	1	,895	,887		

a. Variable(s) introducida(s) en el paso 1: edad, tabaco, hta, iu, razarec, ptlrecod, pesomkg, visitrec.

Al final se queda con cinco variables predictoras, todas con significación estadística en el contraste de su coeficiente de regresión (***TABACO***, ***HTA***, ***RAZAREC***, ***PTLRECOD*** y ***PESOMKG***), siendo este resultado diferente al obtenido por el procedimiento automático HACIA DELANTE (recuérdese que sólo había incluido tres variables en el modelo, ***HTA***, ***PTLRECOD*** y ***PESOMKG***).

Con estos métodos automáticos (HACIA DELANTE y HACIA ATRÁS) el investigador puede tener una orientación para llevar a cabo un procedimiento manual conocido como ***"REGRESIÓN paso a paso"***, y en el que puede ir construyendo modelos introduciendo y

eliminando variables en cada paso, lo que le permitirá evaluar los cambios que se van produciendo en el los coeficientes y sus errores estándar, las modificaciones en el estadístico de ajuste [–2 log verosimilitud] y en las tablas de clasificación de cada modelo, incluso optar por dejar covariables que, aún no siendo "*estadísticamente significativas*" en el contraste de su coeficiente, aportan información al modelo o permiten obtener estimaciones ajustadas de otras variables predictoras. Por supuesto también podrá introducir términos de interacción, imposibles en los procedimientos automáticos, como ya hemos comentado en varias ocasiones.

Paso 4. Evaluar el modelo final.

Cualquier ecuación de regresión logística que se obtenga debe ser considerada como ***provisional***, ya que debería someterse a una evaluación de cómo el modelo se ajusta a los datos.

Puede parecer una incongruencia hablar de "*ajuste a los datos*" al modelo cuando estos han sido empleados precisamente para obtener la ecuación de regresión; sin embargo el modelado matemático no es perfecto por muchas razones (la estimación se hace por un proceso iterativo de cálculo a través del proceso de máxima verosimilitud) y los valores pronosticados no siempre coinciden con los verdaderamente observados. Téngase en cuenta, además, que en el caso de la ecuación de RL, lo que se obtiene para cada combinación de valores de las diferentes variables predictoras incluídas en el modelo es una probabilidad, un valor que oscila entre 0 y 1, y con él un individuo debe ser clasificado en una de las dos posibilidades o categorías que establece la variable resultado o dependiente. En general si el valor de probabilidad predicho está entre 0 y 0,5 el individuo se clasifica como Y=0, mientras que si la probabilidad calculada es > 0,5 el sujeto se clasifica como Y=1. Este punto de corte (0,5) es el que el programa SPSS asume por defecto, pero puede modificarse en la ventana de **OPCIONES**, donde pone "**Punto de corte para la clasificación**".

Una de las formas de evaluar el ajuste del modelo es, precisamente, mediante una especie de "*valoración de una prueba diagnóstica*", comprobando cómo clasifica el nuevo test (el modelo obtenido) a los individuos de la muestra en comparación con el *gold estándar* (la realidad, lo observado). De hecho, el programa SPSS de RL analiza automáticamente, tras seleccionar las variables del modelo, cuál sería la clasificación de los individuos del estudio tras aplicar la ecuación obtenida, y crea una tabla 2x2 con los valores pronosticados y los realmente observados, como ya hemos visto antes.

En el ejemplo que estamos trabajando, con seis variables predictoras en el modelo, la tabla de clasificación es:

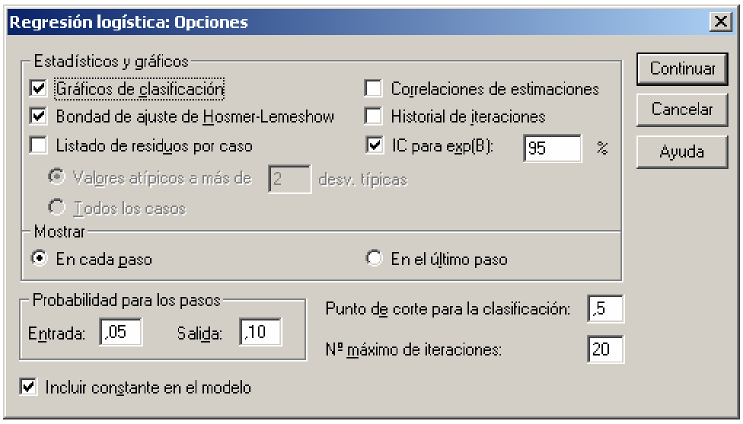
Tabla de clasificación[a]

			Pronosticado		
			Bajo peso al nacer		Porcentaje correcto
	Observado		>=2500 gr	<2500 gr	
Paso 1	Bajo peso al nacer	>=2500 gr	117	13	90,0
		<2500 gr	35	24	40,7
	Porcentaje global				74,6

a. El valor de corte es ,500

...donde podemos apreciar como el modelo obtenido clasificaría correctamente a sólo 24 (de los 59) RN de "*bajo peso al nacer*" (Y=1), por lo que su **sensibilidad** sería del 40,7% (24/59); por el contrario clasifica correctamente a la mayoría de los que tienen verdaderamente un "*peso normal al nacer*" (Y=0), en concreto a 117 de los 130 sujetos "controles" de nuestro estudio, por lo que la **especificidad** del modelo es del 90% (117/130). Y de forma global diríamos que ha clasificado correctamente al 74,6% de los individuos ([24+117]/189).

Si deseamos tener una comprobación visual de cómo se distribuyen los diferentes individuos en función de los valores de probabilidad predichos, no tenemos más que marcar “**gráficos de clasificación**” en la ventana de **OPCIONES**:



E incluso podríamos obtener diferentes gráficos y tablas de clasificación modificando manualmente el “**Punto de corte para la clasificación**”. Este es el fundamento para construir curvas ROC que, luego veremos, el programa SPSS puede hacer a través del procedimiento **GRAFICOS**.

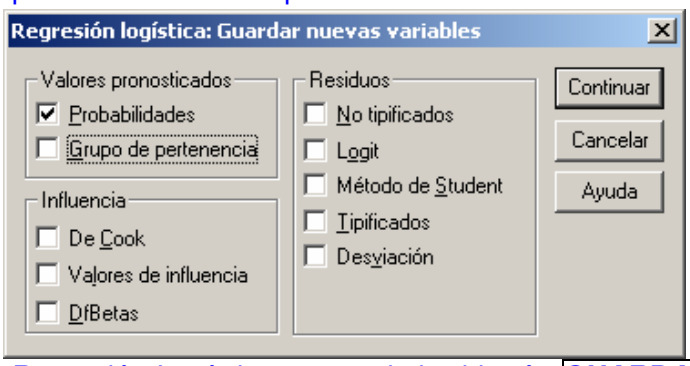
Los gráficos de clasificación que brinda el programa SPSS no son muy buenos. Veamos la salida para el punto de corte 0,5:

```
              Observed Groups and Predicted Probabilities

      16 □
         □
         □              *
F        □              *
R     12 □    *         *
E        □    >        ***
Q        □    >       >>*
U        □    >      *>>>
E      8 □    >      >>>>*
N        □    >**   * >>>>**
C        □   >>>>   * *>>>>>*          *
Y        □  >>>>>   >*>>>>>>**         >
       4 □ >>>>>>   >*>>>>>>>*  *     >*   **             *
         □ >>>>>>*  >*>>>>>>>>  *    *>*   ** *      *    *
         □ >>>>>>>>>>>>>>>>>>  > *>>>  *>>** >   *  *   > *
         □ >>>>>>>>>>>>>>>>>> >> *>>>>*>>>**>****>> >>>*  ***
*
Predicted
□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□
  Prob:    0            0,25           0,5            0,75
1
  Group:
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>******************************

          Predicted Probability is of Membership for *2500 gr
          The Cut Value is 0,50
          Symbols: > - >= 2500 gr (peso normal al nacer)
                   * - < 2500 gr (bajo peso al nacer)
          Each Symbol Represents 1 Case.
```

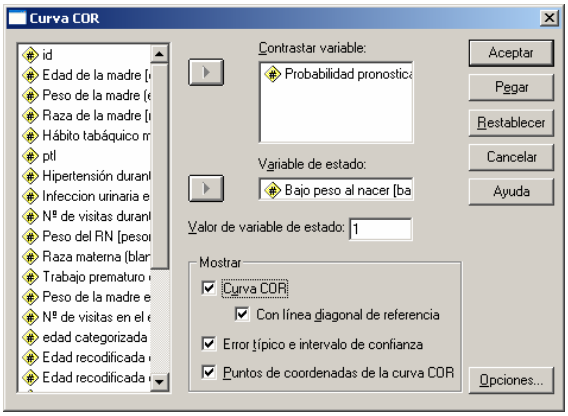
Los asteriscos (*) corresponden a los individuos que eran los casos del estudio, esto es, las gestaciones con RN de bajo peso al nacer, y podemos comprobar visualmente que a pesar

de ello, el modelo de RL obtenido les calculó un valor de probabilidad (Y=1) mayor a 0,5 (y por tanto los clasificó como "valor predicho = Bajo peso al nacer") en algo menos de la mitad de las ocasiones, y el modelo es por tanto, poco sensible.

Para obtener la CURVA ROC es necesario primero guardar los valores de probabilidad pronosticados por el modelo de RL para cada individuo de la muestra, lo que se consigue en



la ventana de Regresión Logística tras oprimir el botón **GUARDAR** y marcar en la ventana siguiente la opción "**Probabilidades**". Con ello se genera una nueva variable de forma automática, que aparecerá en nuestra ventana "vista de datos" con el nombre por defecto ***PRE_1*** (etiqueta: "Probabilidad pronosticada") Luego hay que ir al procedimiento **GRAFICOS** y seleccionar **CURVA COR**:



La variable a contrastar será la que se ha creado en el paso anterior (***PRE_1***) y la variable de estado será la dependiente del estudio (***BAJOPESO***), debiendo señalarse el valor de la variable estado que se pronostica por el modelo (en nuestro caso "1", el valor correspondiente a "*RN de bajo peso*"). El resultado sería:

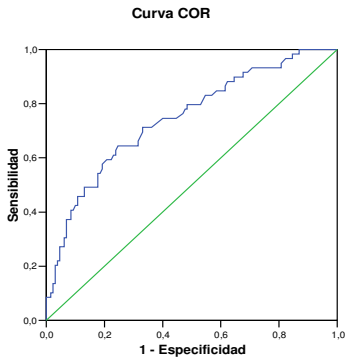
Curva COR

lesumen del proceso de casos

Bajo peso al nace	N válido (según lista)
Positivo[a]	59
Negativo	130

Los valores mayores en la variable de resultado de cc indican una mayor evidencia de un estado real positiv

a. El estado real positivo es <2500 gr.



Los segmentos diagonales son producidos por los empates.

Área bajo la curva

Variables resultado de contraste: Probabilidad pronosticada

Área	Error típ.[a]	Sig. asintótica[b]	Intervalo de confianza asintótico al 95%	
			Límite inferior	Límite superior
,745	,039	,000	,669	,822

La variable (o variables) de resultado de contraste: Probabilidad pronosticada tiene al menos un empate entre el grupo de estado real positivo y el grupo de estado real negativo. Los estadísticos pueden estar sesgados .

a. Bajo el supuesto no paramétrico

b. Hipótesis nula: área verdadera = 0,5

Vemos como la sensibilidad y la especificidad del modelo calculado varían según se establezca un punto de corte u otro para clasificar a los individuos en un grupo de predicción y que el área bajo la curva es 0,746. Este es el poder de discriminación del modelo construído, un 75% del máximo posible. Evidentemente es estadísticamente significativo (la hipótesis nula es la no discriminación, que en la gráfica ROC corresponde a los puntos que caen sobre la diagonal).

Una segunda opción para evaluar el ajuste del modelo construido es a través del test de Hosmer-Lemeshow, una de las diferentes técnicas empleadas para evaluar la bondad del ajuste, muy recomendable cuando se incluyen en el modelo al menos una variable cuantitativa.

Para obtenerlo hay que marcar la correspondiente casilla en la ventana **OPCIONES** de la RL ("**Bondad de ajuste de Hosmer-Lemeshow**"). El test consiste en establecer los deciles de riesgo o probabilidad predicha por el modelo de presentar el evento, y en cada una de estas diez categorías se comparan los valores observados y los predichos, tanto para los que tienen el resultado explorado como para los que no lo tienen. Si hay una elevada coincidencia entre observados y esperados (un buen ajuste), el test Chi cuadrado que contrastará ambas distribuciones (con 8 grados de libertad) no mostrará significación estadística.

Tabla de contingencias para la prueba de Hosmer y Lemeshow

		Bajo peso al nacer = >=2500 gr		Bajo peso al nacer = <2500 gr		
		Observado	Esperado	Observado	Esperado	Total
Paso 1	1	18	17,756	1	1,244	19
	2	18	16,993	1	2,007	19
	3	12	15,898	7	3,102	19
	4	17	14,793	2	4,207	19
	5	16	14,993	4	5,007	20
	6	15	13,719	4	5,281	19
	7	11	12,763	8	6,237	19
	8	10	10,465	9	8,535	19
	9	7	8,350	12	10,650	19
	10	6	4,272	11	12,728	17

Prueba de Hosmer y Lemeshow

Paso	Chi-cuadrado	gl	Sig.
1	10,768	8	,215

Cuando el test Chi cuadrado de la prueba no es significativo quiere indicarnos que no hay motivos para pensar que los resultados predichos sean diferentes de los observados (o que si hay diferencias pueden explicarse razonablemente por el azar o error del muestreo) y que el modelo puede considerarse aceptable. Por otra parte, la tabla de contingencia para la prueba de Hosmer Lemeshow nos da información adicional sobre cada categoría de riesgo,

de manera que podemos comprobar en qué regiones de la predicción el modelo se ajusta peor a los datos.

Una tercera opción para valorar el ajuste del modelo en su conjunto es a través de la llamadas medidas globales de bondad del ajuste, entre las que se encuentran:

- **La devianza del modelo** (que corresponde a menos dos veces el logaritmo neperiano de la verosimilitud). Esta debe ser menor en el modelo más ajustado de todos los probados con el subconjunto de covariables.

- **La razón de verosimilitud (estadístico G)**, que se calcula comparando la devianza de dos modelos, y que sigue una distribución Chi cuadrado con tantos grados de libertad como la diferencia en el número de parámetros entre los dos modelos que se comparan. La significación estadística del contraste (***X*²** del modelo en el programa SPSS) indica un buen ajuste.

- **La R^2** (y otras **pseudo-R^2**) del modelo, estadísticos que pretenden cumplir la misión del coeficiente de determinación de la regresión lineal y, por tanto, expresar la variabilidad de la variable dependiente que es explicada por el modelo. Valores próximos a 1 serán indicadores de muy alto ajuste y capacidad predictiva, mientras que valores cercanos a cero justo de lo contrario.

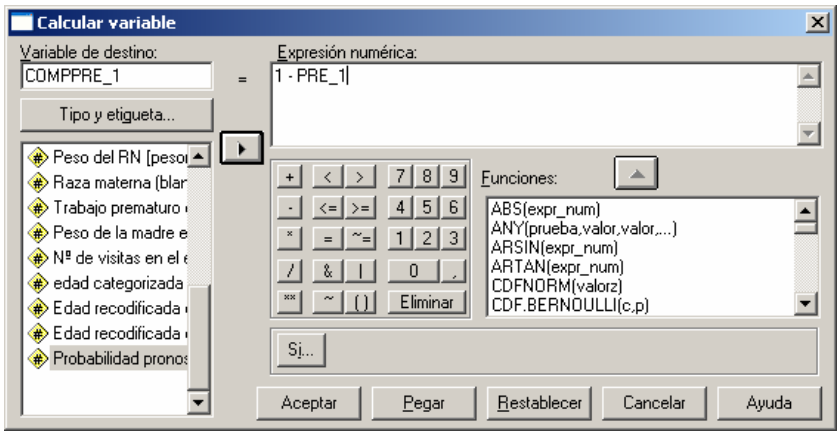
Finalmente, todo análisis de regresión logística múltiple en el que se hayan incluido variables cuantitativas debe completarse con un estudio de linealidad en el logit.

Para ello es necesario obtener (y guardar como una variable) para cada individuo de la base de datos el logit (p), que recordemos se calcula por la siguiente función:

$$Logit\ (p) = \ln \frac{p}{1-p} = \ln \frac{P(Y=1)}{1-P(Y=1)}$$

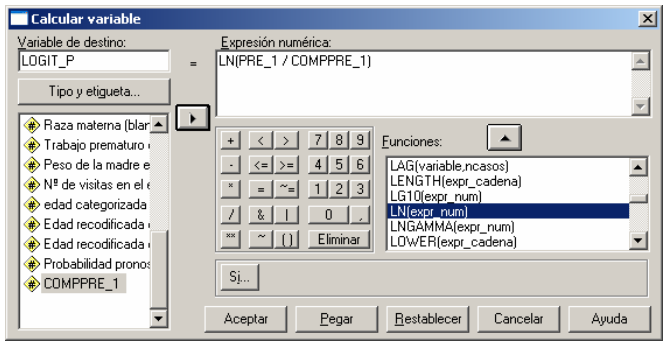
La probabilidad predicha por el modelo, esto es, P(Y=1), se obtiene con el programa SPSS como ya se ha visto antes, marcando en la ventana **GUARDAR** la casilla **Probabilidades**. Automáticamente, tras hacer la regresión logística, se crea una variable nueva que el programa llama ***PRE_1***.

Para obtener su valor complementario [1 – p(Y=1)] hay que obtener manualmente una nueva variable calculada (la llamaremos ***COMPPRE_1***, de complementario de ***PRE_1***) a través de la opción **TRANSFORMAR** y luego **CALCULAR** en la barra de herramientas superior, indicando en la ventana de "**Expresión numérica**" la correspondiente fórmula:



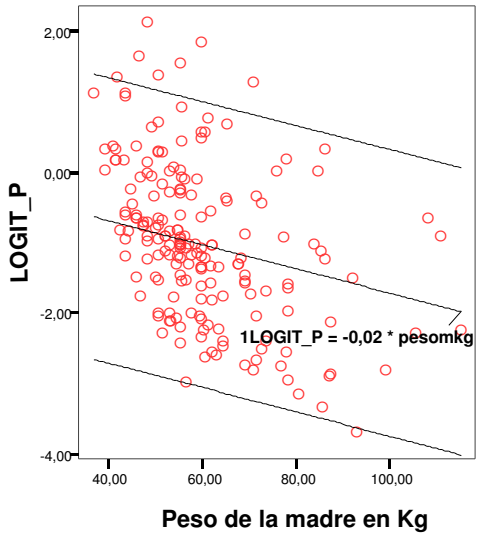
Una vez obtenidas las dos variables, que pueden comprobarse estarán ya en la vista de variables, se tiene que calcular el logit de p (la llamaremos ***LOGIT_P***) esto es, se tiene que

obtener una nueva variable calculada que será el logaritmo neperiano del cociente entre las dos anteriores:



Para comprobar la linealidad del logit podemos recurrir a procedimientos gráficos (**DISPERSIÓN**) o a procedimientos matemáticos (**CORRELACIÓN**, **REGRESIÓN LINEAL**), enfrentando la variable cuantitativa que deseemos explorar frente a la variable ***LOGIT_P*** que hemos obtenido por los pasos anteriores.

Veamos qué ocurre con la variable ***PESOMKG*** que hemos dejado como covariable en el modelo que hemos ajustado para el desenlace ***BAJOPESO***:



Regresión lineal por el origen con
Intervalo de predicción individual al 95,00%

Puede verse como hay cierta relación lineal aunque con una importante dispersión individual. De hecho, el análisis de correlación bivariante entre estas dos variables detecta una correlación lineal negativa con un coeficiente de Pearson de – 0,435, que es estadísticamente significativo (p<0,001).

Correlaciones

		LOGIT_P	Peso de la madre en Kg
LOGIT_P	Correlación de Pearson	1	-,435**
	Sig. (bilateral)		,000
	Suma de cuadrados y productos cruzados	229,532	-1272,100
	Covarianza	1,221	-6,766
	N	189	189
Peso de la madre en Kg	Correlación de Pearson	-,435**	1
	Sig. (bilateral)	,000	
	Suma de cuadrados y productos cruzados	-1272,100	37198,968
	Covarianza	-6,766	197,867
	N	189	189

**. La correlación es significativa al nivel 0,01 (bilateral).