VIII.2. Modelos de regresión para variables dependientes cualitativas

En el caso de que la variable dependiente sea cualitativa un modelo adecuado es el de regresión logística.

Se utiliza ampliamente en investigación clínica, ya que permite estimar la probabilidad de ocurrencia de un proceso en función de ciertas variables, permitiendo evaluar la influencia de las variables independientes sobre la variable dependiente, dando como resultado una probabilidad.

La variable dependiente es siempre cualitativa pero las variables independientes pueden ser continuas, discretas, categóricas, dicotómicas o una mezcla de todas ellas.

Al igual que ocurría con la regresión para variables dependientes cuantitativas es necesario que no exista multicolinealidad entre las diferentes variables independientes y las observaciones de la muestra deben ser independientes entre sí. Sin embargo, no se requiere que los residuos presenten una distribución Normal ni la hipótesis de homocedasticidad (varianza constante de los residuos).

La variable dependiente puede ser dicotómica (0 si el hecho no ocurre y 1 si ocurre) o politómica (hay varias categorías), dando lugar a dos tipos diferentes de regresión logística, la binomial o multinomial, respectivamente.

VIII.2.1. Regresión logística binomial

La ecuación a la que responde el modelo es:

$$Y_i = \log\left(\frac{\Pi_i}{1-\Pi_i}\right)$$

donde Π_i es la probabilidad de que en el caso *i* se produzca el evento estudiado e Y_i es el valor de la variable dependiente en el caso *i*.

Expresado en forma de regresión:

$$\Pi_i = \frac{1}{1+e^{-(\alpha+\beta X_i)}}$$

donde α y β son los coeficientes de la ecuación.

En caso de existir varias variables predictoras (independientes) la regresión se transforma en:

$$\Pi_i = \frac{1}{1 + e^{\left[-(\alpha + \sum_{j=1}^{k} \beta_j X_{ji})\right]}}$$

donde Π_i es la probabilidad de que se produzca el evento estudiado en el caso *i*, *k* es el número de variables predictoras, α es un coeficiente, β_j es el coeficiente de la variable predictora *j* y X_{ji} es el valor de la variable predictora *j* en el caso *i*.

En caso de que Π_i sea mayor de 0,5 se asume (a efectos de predicción) que el evento se produce y si es menor que 0,5 que no se produce.

Un ejemplo de cálculo de regresión logística binomial con SPSS aparece en el Cuadro VIII.4.

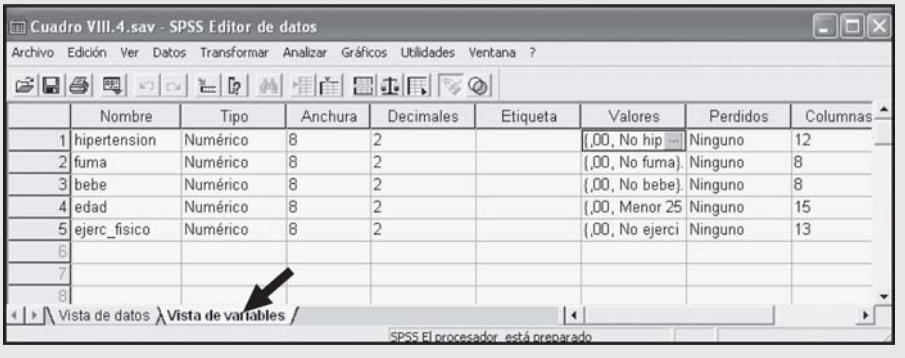
CUADRO VIII.4. Cálculo de la regresión logística binomial

EJEMPLO. Se quiere conocer si a partir de las variables fumador (1 fuma, 0 no fuma), bebedor (1 bebe más 2 copas de vino al día, 0 bebe menos de 2 copas de vino diarias), edad (0 menor de 25 años, 1 entre 25 y 50 años y 2 más de 50 años) y realización de ejercicio físico (1 más de 3 horas semanales, 0 menos de 3 horas semanales), se puede predecir la presencia de hipertensión (1 padece hipertensión, 0 no padece hipertensión) en los pacientes. Los datos aparecen en el archivo **Cuadro VIII.4.sav** y usaremos el programa SPSS.

Antes de describir los pasos del análisis es interesante indicar el proceso de etiquetado de los datos en SPSS ya que nos ayudará en gran medida a la interpretación de los resultados.

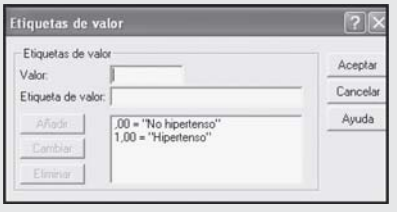
El programa SPSS permite la asignación de etiquetas a las distintas categorías (que deben introducirse con números para realizar los análisis).

Este etiquetado se realiza en la pestaña de «Vista de variables» (véase flecha).

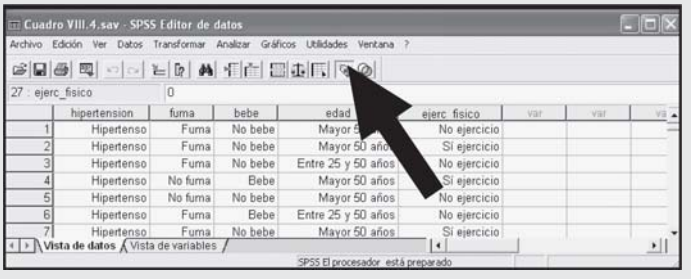


CUADRO VIII.4. *(Continuación)*

Como ejemplo vamos a hacer la variable *Hipertensión*. Hacemos clic sobre el cuadro valores de la variable y luego marcamos sobre los puntos suspensivos que aparecen (ver pantalla anterior) y nos sale una pantalla en la que debemos indicar el valor de la categoría y su etiqueta. Introducimos valor 0 y etiqueta de valor *No hipertenso* y marcamos sobre el icono «Añadir». Después escribimos el valor 1 y como etiqueta *Hipertenso*, marcando de nuevo sobre el icono «Añadir». Nos queda la siguiente pantalla.

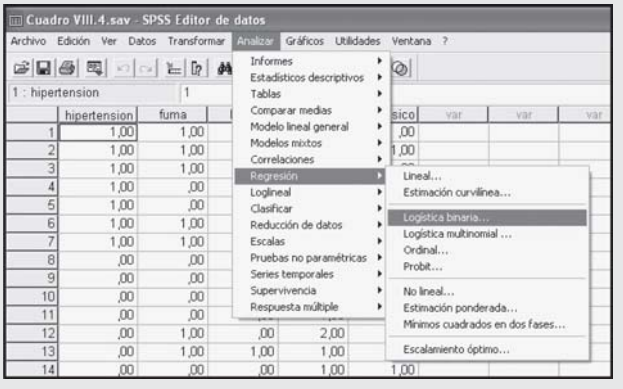


Pulsamos el icono «Aceptar» y ya tenemos los datos etiquetados. Para ver las etiquetas se pulsa sobre el icono *etiquetas* (ver flecha en la pantalla siguiente en la pestaña «Vista de datos»).



Pasamos ya a la descripción de los pasos para calcular la regresión logística binaria con SPSS.

Paso 1. Entramos en la sección «Analizar», dentro de esta en «Regresión», y dentro de esta última en «Logística binaria…».



CUADRO VIII.4. *(Continuación)*

Paso 2. Nos aparecerá el siguiente cuadro en el cual es necesario, en primer lugar, definir la variable «Dependiente» y las «Covariables» (variables predictoras).

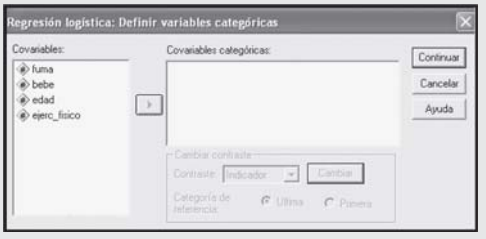
En esta pantalla cabe destacar que es posible realizar distintos modelos de regresión (que denomina bloques) al mismo tiempo, bien cambiando las variables independientes o el método de elección de variables. Para ello solamente debemos pulsar en el icono «Siguiente» y nos aparecerá la ventana que se muestra a continuación donde se deben indicar todas las opciones que se deseen para el análisis.

Como podemos observar, en la pantalla aparece el número de bloque indicando que es el Bloque 2 de 2, es decir, vamos a ejecutar 2 modelos y en esta pantalla se definirá el segundo.

Podríamos añadir todos los bloques que quisiésemos, pero con nuestros datos vamos a trabajar solo con el primer bloque, por lo que pulsamos sobre el icono «Anterior» para volver al la pantalla del Bloque 1.

CUADRO VIII.4. *(Continuación)*

Paso 3. Como nuestras variables predictoras son categóricas pulsamos sobre el icono «Categórica» para indicarlo. Nos aparece el siguiente cuadro.



Paso 4. Aunque todas nuestras covariables son categóricas, las que son dicotómicas (*fuma*, *bebe* y *ejerc_fisico)* no es necesario considerarlas como categóricas. Por lo tanto, solo pasamos al cuadro de «Covariables categóricas» la *edad*. Una vez hemos pasado la *edad* vemos como aparece un paréntesis después de la variable en el que se nos indica cómo queremos seleccionar la categoría de referencia a la hora de realizar las comparaciones, lo que es necesario para interpretar correctamente los resultados.

Paso 5. Existen diferentes comparaciones que se pueden utilizar. Para cambiarlos pulsamos sobre «Cambiar». Las posibilidades son:

-*Indicador.* Presencia o ausencia de cada categoría.

-*Simple.* Cada categoría del predictor (excepto la propia categoría de referencia) se compara con la categoría de referencia.

-*Diferencia.* Cada categoría del predictor, excepto la primera categoría, se compara con el efecto promedio de las categorías anteriores.

-*Helmert.* Cada categoría del predictor, excepto la última categoría, se compara con el efecto promedio de las categorías subsiguientes.

-*Repetidas.* Cada categoría del predictor, excepto la primera categoría, se compara con la categoría que la precede.

CUADRO VIII.4. *(Continuación)*

-*Polinómico.* Contrastes polinómicos ortogonales. Se supone que las categorías están espaciadas equidistantemente. Los contrastes polinómicos solo están disponibles para variables numéricas.

-*Desviación.* Cada categoría del predictor, excepto la categoría de referencia, se compara con el efecto global.

Para los contrastes Indicador, Simple y Desviación es posible indicar si se desea que la categoría de referencia sea la primera o última de los datos.

En nuestro caso, dejaremos Indicador, ya que es el contraste que SPSS pone por defecto y ponemos que tenga como referencia la primera categoría (valor 0). Para el cambio pulsar sobre el icono «Cambiar».

Pulsamos sobre el icono «Continuar» y volvemos a la pantalla principal en la que, como se muestra en el cuadro siguiente, ya se indica que las variables son categóricas.

Paso 6. Pulsando sobre «Método» se indica cómo queremos que se introduzcan las variables independientes en el modelo.

Existen distintos métodos según se utilicen todas las variables sin eliminar las no significativas, se seleccionen las variables hacia *adelante* (es decir, se van incluyendo en el modelo las variables más significativas hasta que todas la que no han sido seleccionadas no son significativas) o se seleccionan hacia *atrás* (es decir, se incluyen en el modelo todas las variables y se van eliminando las menos significativas y así, sucesivamente, hasta que todas las variables en el modelo sean significativas). Los métodos de introducción posibles son:

CUADRO VIII.4. *(Continuación)*

-Introducir. Procedimiento para la selección de variables en el que todas las variables de un bloque se introducen en un solo paso. Incluye todas las variables aunque no sean significativas.

-Adelante:Condicional. Contrasta la entrada basándose en la significación del estadístico de puntuación y contrasta la eliminación de acuerdo a la probabilidad de un estadístico de la razón de verosimilitud que se basa en estimaciones condicionales de los parámetros.

-Adelante:RV. Contrasta la entrada basándose en la significación del estadístico de puntuación y contrasta la eliminación en relación al estadistico de la razón de verosimilitud, que se basa en estimaciones de la máxima verosimilitud parcial.

-Adelante:Wald. Método de selección por pasos hacia adelante que contrasta la entrada basándose en la significación del estadístico de puntuación y contrasta la eliminación basándose en la probabilidad del estadístico de Wald. El estadístico de Wald permite una prueba χ^2 para contrastar la hipótesis nula de que el coeficiente de cada variable independiente es cero.

-Atrás:Condicional. Selección por pasos hacia atrás. El contraste para la eliminación se basa en la probabilidad del estadístico de la razón de verosimilitud, el cuál se basa a su vez en las estimaciones condicionales de los parámetros.

-Atrás:RV. Selección hacia atrás por pasos. El contraste para la eliminación se fundamenta en la probabilidad del estadístico de la razón de verosimilitud, el cual se basa en estimaciones de máxima verosimilitud parcial.

-Atrás:Wald. Selección por pasos hacia atrás. El contraste para la eliminación se basa en la probabilidad del estadístico de Wald.

En nuestro caso, seleccionaremos «Atrás:Wald» ya que queremos que el modelo incluya en un principio todas las variables independientes y vaya quitando variables en cada paso hasta solo quedar las variables significativas.

Paso 7. Volviendo a la pantalla y pulsando sobre el icono «Guardar» de la pantalla anterior nos aparecen las posibles opciones para guardar los resultados de la regresión logística como nuevas variables.

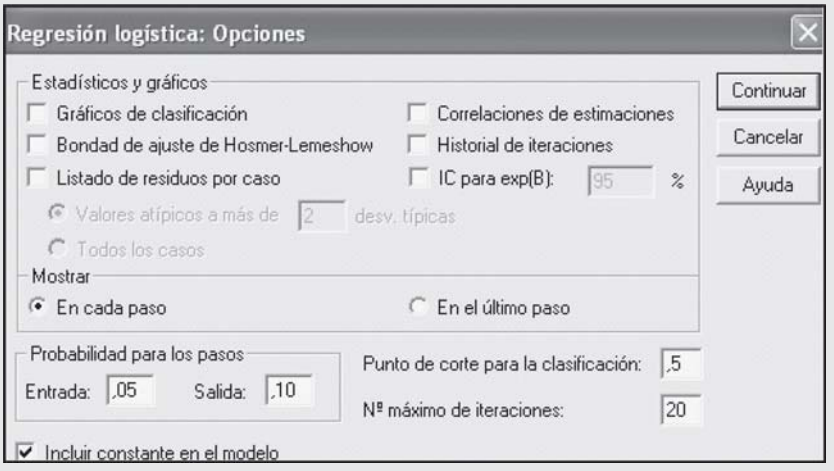
CUADRO VIII.4. *(Continuación)*

Es posible guardar la probabilidad de que se produzca el evento (en este caso hipertensión) o el grupo de pertenencia —según el modelo— de cada caso. Además es posible solicitar la influencia de cada caso particular en los valores predichos: «De Cook» mide como cambian los residuos si excluimos cada caso, «Valores de influencia» la importancia de cada dato en el ajuste del modelo y «DfBetas» mide el cambio en el coeficiente de regresión excluyendo cada caso. También es posible obtener diferentes mediciones de los residuos.

Paso 8. Pulsamos sobre el icono «Continuar» y volvemos a la pantalla principal que aparece a continuación.

Paso 9. Ya por último, pulsamos sobre el icono «Opciones» que nos permite personalizar algunas características del procedimiento.

CUADRO VIII.4. *(Continuación)*



En la sección «Estadísticos y gráficos» podemos indicar si queremos ver el «Gráfico de clasificación», así como la «Bondad de ajuste Hosmer-Lemeshow» (se basa en agrupar los casos en deciles de riesgo y comparar la probabilidad observada con la esperada dentro de cada decil). También podemos solicitar el «Listado de residuos por caso», las «Correlaciones de estimaciones», el «Historial de iteraciones» o el intervalo de confianza para los distintos β «IC para exp(B)».

En la sección «Mostrar» podemos indicar si queremos ver los resultados «En cada paso» o solo «En el último paso».

También podemos indicar la probabilidad de «Entrada» y «Salida» de las covariables en los modelos por pasos, así como el «Nº máximo de iteraciones» y el «Punto de corte para la clasificación», es decir, a partir de qué probabilidad vamos a suponer o predecir que se produce el evento.

Con nuestros datos vamos a dejar las opciones que el programa trae por defecto.

Paso 10. Se pulsa el icono «Continuar» saliendo de nuevo la pantalla principal (paso 8) y pulsamos sobre «Aceptar» para obtener el archivo de resultados.

Paso 11. Interpretación de los resultados. En los resultados nos centraremos en la sección «Bloque 1: Método = Por pasos hacia atrás (Wald)» que es donde aparecen los resultados finales. Se nos muestran los resultados para los distintos pasos que ha realizado el programa. A nosotros nos va a interesar el último paso (con nuestros datos el paso 3).

CUADRO VIII.4. *(Continuación)*

La tabla que aparece a continuación «Pruebas omnibus sobre los coeficientes del modelo», nos muestra el valor de los contrastes χ^2 entre nuestros datos y el modelo que hemos generado. Como nuestra regresión logística era por pasos, nos va a indicar los valores de χ^2 en los distintos pasos. El χ^2 de bloque no nos interesa, ya que en nuestro caso tenemos un único bloque, pero en caso de tener varios, nos indicaría si existen diferencias significativas entre los distintos bloques.

En el primer paso (introducción de las cuatro variables independientes) observamos que el χ^2 del modelo es significativo con $p < 0{,}001$, es decir, las variables independientes describen la variable dependiente de forma significativa. En este caso el χ^2 de paso (comparación entre los pasos sucesivos de la regresión) es igual al del modelo, ya que lo compara con el modelo sin considerar variables.

En el paso 2 observamos que hay una reducción en el estadístico χ^2 entre pasos, y que no hay diferencias significativas entre el modelo del paso 1 y el modelo del paso 2 con una variable menos ($p = 0{,}452$) y que el modelo sigue siendo significativo ($p < 0{,}001$).

Ya en el último paso (modelo final con solo dos variables independientes) el cambio en el valor de χ^2 tampoco es significativo frente al paso anterior ($p = 0{,}507$), y el modelo sigue siendo significativo ($p < 0{,}001$).

Pruebas omnibus sobre los coeficientes del modelo

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	30,064	5	,000
	Bloque	30,064	5	,000
	Modelo	30,064	5	,000
Paso 2[a]	Paso	-,566	1	,452
	Bloque	29,498	4	,000
	Modelo	29,498	4	,000
Paso 3[a]	Paso	-,440	1	,507
	Bloque	29,058	3	,000
	Modelo	29,058	3	,000

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

CUADRO VIII.4. *(Continuación)*

La tabla «Variables en la ecuación» nos indica las variables seleccionadas en cada paso. En el paso 1 incluye todas las variables, en el paso 2 elimina la variable *ejerc_fisico* por ser la menos significativa en el paso anterior y en el paso 3 elimina *bebe* por no ser significativa, quedándose con las variables *fuma* y *edad*, que son significativas.

En el caso de que las variables categóricas no sean dicotómicas (solo *edad*), se crea un coeficiente β para cada categoría distinta de la de referencia (que tiene valor 0). El valor positivo del coeficiente β nos indica que favorece la aparición de hipertensión (factor de riesgo). Además nos da el estadístico de Wald de los coeficientes β con su significación, siendo en el paso 3 todos significativos ($p < 0,05$).

Especial importancia tiene en esta tabla el valor «Exp(B)», que es la OR (Odds Ratio, razón impar o razón de ventajas) que representa el cociente entre la probabilidad de que ocurra el suceso que define la variable dependiente frente a la probabilidad de que no ocurra en presencia o ausencia del factor. Así, el valor 10,079 correspondiente a la variable *fuma* nos indica que entre los fumadores es 10,079 veces más grande el cociente entre la probabilidad de padecer hipertensión y la de no padecerla que entre los no fumadores, por lo que se trata de un factor de riesgo. Algo semejante ocurre con la pertenencia al grupo de *edad(2)* «> 50 años».

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1[a]	fuma	2,382	,653	13,290	1	,000	10,824
	bebe	,760	,797	,909	1	,340	2,138
	edad			10,442	2	,005	
	edad(1)	,483	1,295	,139	1	,709	1,621
	edad(2)	2,692	1,288	4,366	1	,037	14,764
	ejerc_fisico	-,531	,715	,552	1	,457	,588
	Constante	-2,976	1,315	5,125	1	,024	,051
Paso 2[a]	fuma	2,326	,629	13,668	1	,000	10,241
	bebe	,408	,618	,437	1	,509	1,504
	edad			10,776	2	,005	
	edad(1)	,550	1,279	,185	1	,668	1,732
	edad(2)	2,649	1,279	4,290	1	,038	14,147
	Constante	-3,044	1,300	5,487	1	,019	,048
Paso 3[a]	fuma	2,310	,624	13,700	1	,000	10,079
	edad			10,812	2	,004	
	edad(1)	,615	1,259	,238	1	,626	1,849
	edad(2)	2,559	1,255	4,157	1	,041	12,929
	Constante	-2,864	1,246	5,281	1	,022	,057

a. Variable(s) introducida(s) en el paso 1: fuma, bebe, edad, ejerc_fisico.

CUADRO VIII.4. *(Continuación)*

En el «Resumen de los modelos» se nos indica el valor estadístico «-2 log de la verosimilitud» (útil para hacer comparaciones) así como 2 valores de r^2 que no son equivalentes a r^2 de la regresión lineal, por lo que deben ser utilizados con cautela. El r^2 de Cox y Snell no varía entre 0 y 1 por lo que se ha modificado en lo que se llama el r^2 de Nagelkerke, que sí varía entre 0 y 1. Atendiendo a nuestro caso, tendríamos que en el paso 3 se explica un 41,5% de la variabilidad de los datos. El hecho de que en el paso 1 y 2 sea mayor es simplemente porque hay más variables en el modelo, pero no son significativas, por lo que no deben considerarse.

Resumen de los modelos

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	78,016[a]	,320	,427
2	78,582[a]	,315	,420
3	79,022[a]	,311	,415

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

La «Tabla de clasificación» nos indica el porcentaje de clasificaciones correctas con nuestro modelo. De nuevo nos fijamos en el paso 3, que nos revela que se clasifican correctamente el 71,8% de los datos. Tenemos un porcentaje de acierto similar en aquellos individuos que no tienen hipertensión (72,5%) frente a los que sí la padecen (71,1%).

Tabla de clasificación[a]

			Pronosticado		
			hipertension		Porcentaje correcto
	Observado		No hipertenso	Hipertenso	
Paso 1	hipertension	No hipertenso	29	11	72,5
		Hipertenso	17	21	55,3
	Porcentaje global				64,1
Paso 2	hipertension	No hipertenso	29	11	72,5
		Hipertenso	11	27	71,1
	Porcentaje global				71,8
Paso 3	hipertension	No hipertenso	29	11	72,5
		Hipertenso	11	27	71,1
	Porcentaje global				71,8

a. El valor de corte es ,500

CUADRO VIII.4. *(Continuación)*

En resumen, tenemos un modelo significativo (χ^2, $p < 0{,}001$) con un r^2 de Nagelkerke de 0,41 que clasifica correctamente el 71,8% de los casos.

Paso 12. Aplicación de la función de probabilidad. Si quisiésemos aplicar la función de probabilidad, tendríamos que saber cuál ha sido la codificación interna del programa (indicada por nosotros en la categoría de referencia). Para conocerla, en los resultados aparecen estas dos tablas:

Codificación de la variable dependiente

Valor original	Valor interno
No hipertenso	0
Hipertenso	1

Codificaciones de variables categóricas

		Frecuencia	Codificación de (1)	Codificación de (2)
edad	Menor 25 años	7	,000	,000
	Entre 25 y 50 años	30	1,000	,000
	Mayor 50 años	41	,000	1,000

En nuestro caso la codificación interna queda igual que la codificación de nuestros datos, con la categoría de referencia valor 0. A partir de la tabla «Variables en la ecuación» se obtienen los coeficientes de la ecuación. Así, podemos decir que la ecuación de probabilidad de que un individuo tenga hipertensión quedaría:

$$\Pi(1) = \frac{1}{1 + e^{-(-2{,}864 + 2{,}310 * fuma + 0{,}615 * edad(1) + 2{,}559 * edad(2))}}$$

CUADRO VIII.4. *(Continuación)*

En un individuo que fuma y tiene más de 50 años (según la tabla de clasificaciones *fuma* = 1, *edad(1)* = 0 y *edad(2)* = 1) la probabilidad de sufrir hipertensión es 88,1%, o lo que es lo mismo, tiene una probabilidad del 11,9% de no sufrir hipertensión.

$$\Pi(1) = \frac{1}{1 + e^{-(-2{,}864 + 2{,}310*1 + 0{,}615*0 + 2{,}559*1)}} = 0{,}881$$

En un individuo que no fuma y tiene menos de 25 años (según la tabla de clasificaciones *fuma* = 0, *edad(1)* = 0 y *edad(2)* = 0) la probabilidad de sufrir hipertensión es 5,4%, o lo que es lo mismo, tiene una probabilidad de 94,6% de no sufrir hipertensión.

$$\Pi(1) = \frac{1}{1 + e^{-(-2{,}864 + 2{,}310*0 + 0{,}615*0 + 2{,}559*0)}} = 0{,}054$$

VIII.2.2. Regresión logística multinomial

La diferencia principal con respecto a la regresión logística binomial es que la variable dependiente cualitativa no es dicotómica, sino que puede tener más de 2 categorías.

Suponiendo *k* clases o categorías, y *j* variables independientes, este modelo se puede resumir en las siguientes ecuaciones, que proporcionan las probabilidades de pertenencia a las primeras *k*-1 clases:

$$\Pi_{in} = \frac{e^{Z_{in}}}{1 + e^{Z_{i1}} + e^{Z_{i2}} + e^{Z_{i3}} + \ldots + e^{Z_{ik-1}}}$$

$$Z_{in} = \beta_{n0} + \beta_{n1}x_{i1} + \beta_{n2}x_{i2} + \ldots + \beta_{nj}x_{ij}$$

donde Π_{in} es la probabilidad de pertenencia del caso *i* al grupo *n;* Z_{in} es el valor de la variable dependiente *Z* correspondiente a la clase *n* en el caso *i*; β_{nh} es el coeficiente de la variable independiente *h* para la clase *n*; x_{ih} es el valor del predictor o variable independiente *h* para el caso *i*. La probabilidad para la última clase *k* se obtiene por diferencia a 1.

Un ejemplo de este tipo de regresión se puede ver en el Cuadro VIII.5.

CUADRO VIII.5. Cálculo de la regresión logística multinomial

EJEMPLO. Durante un experimento de sedimentación de materia orgánica en la columna de agua en zonas costeras, se quiere conocer si es posible determinar el organismo productor de pellets (excrementos) a partir de las características de los mismos, para inferir qué organismo tiene un mayor aporte en la sedimentación de pellets. Las variables medidas de los pellets fueron longitud, ancho, curvatura (1 significa pellet recto, 2 significa curvatura inferior al 10% y 3 significa curvatura superior al 10%) y nivel de digestión (1 significa que está poco digerido y encontramos células de fitoplancton enteras dentro del pellets, 2 significa digestión media encontrando fragmentos de células y 3 indica digestión total sin encontrar ningún fragmento). Los datos aparecen en el archivo **Cuadro VIII.5**.

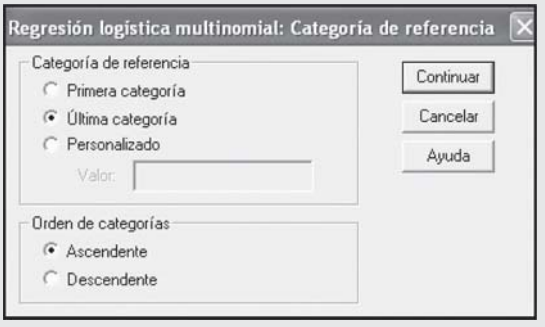
El proceso de etiquetado de variables se hizo igual que en el caso de la regresión logística binaria (Cuadro VIII.4).

Paso 1. Entramos en la sección «Analizar», dentro de esta en «Regresión», y dentro de esta última en «Logística multinomial…».

Paso 2. Nos aparecerá el siguiente cuadro en el cual es necesario, en primer lugar, definir la variable «Dependiente», los «Factores» (variables independientes categóricas) y las «Covariables» (variables dependientes continuas). En nuestro caso, seleccionamos *grupo_productor* como variable dependiente, *Curvatura* y *Nivel_Digestion* como factores y *Longitud* y *Ancho* como covariables.

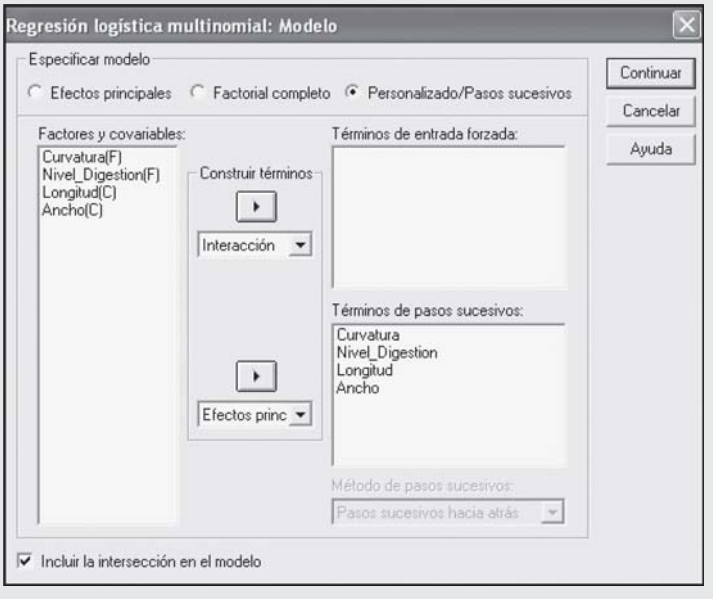
CUADRO VIII.5. *(Continuación)*

Paso 3. En la variable dependiente aparece entre paréntesis la palabra «Último». Este paréntesis indica cuál es la categoría de referencia. SPSS estimará los parámetros para las restantes. Pulsando sobre el icono «Categoría de referencia» podemos elegir la que deseamos.



Podemos elegir que la categoría de referencia sea la «Primera categoría», la «Última categoría» o «Personalizado» (indicar qué categoría queremos que sea). También es posible ordenar las categorías de forma «Ascendente» o «Descendente». En nuestro caso, vamos a elegir las opciones por defecto. Pulsamos sobre el icono «Continuar» para volver a la pantalla principal.

Paso 4. En la parte inferior de la pantalla principal de análisis (paso 2) hay diversos iconos en los que elegir diferentes opciones y parámetros del análisis. Comencemos pulsando sobre el icono «Modelo». Aparece la pantalla que vemos a continuación:



CUADRO VIII.5. *(Continuación)*

En el apartado «Especificar modelo» elegimos los términos que queremos incluir. Así tenemos tres tipos de modelo:

-*Efectos principales.* No incluye las interacciones entre las distintas variables independientes.

-*Factorial completo.* incluye todas las interacciones entre las distintas variables independientes, excepto entre las covariables.

-*Personalizado/Pasos sucesivos*. Permite especificar las interacciones que deseamos incluir, o solicitar un modelo por pasos (eligiendo solo las variables significativas).

En el caso de indicar un modelo «Personalizado/Pasos sucesivos», se activan las opciones para seleccionar el tipo de interacciones. En la columna de la izquierda aparecen las variables independientes (Factores y Covariables), éstas deben pasarse a los cuadros de la derecha donde se seleccionan las interacciones. Para cada variable independiente tenemos dos posibles opciones, ya que cada variable se puede introducir de forma forzosa en el modelo «Términos de entrada forzosa» o introducirse únicamente si son estadísticamente significativas en el modelo «Términos de pasos sucesivos». En nuestro caso introduciremos todas las variables en el cuadro de «Términos de pasos sucesivos» como se muestra en la pantalla anterior, ya que queremos quedarnos únicamente con las variables significativas.

Una vez seleccionadas las variables debemos elegir el tipo de Pasos sucesivos que queremos realizar en «Método» de pasos sucesivos. Existen cuatro posibilidades:

-*Entrada hacia delante.* En cada paso se añade al modelo el término más significativo, hasta que ninguno de los términos por pasos que quede fuera del modelo tenga una contribución estadísticamente significativa si se añadiese al modelo.

-*Eliminación hacia atrás*. Se inicia introduciendo en el modelo todos los términos especificados en la lista por pasos. En cada paso se elimina del modelo el término menos significativo, hasta que todos los términos por pasos restantes representen una contribución estadísticamente significativa para el modelo.

-*Pasos sucesivos hacia adelante.* Este método se inicia con el modelo que se seleccionaría mediante el método de entrada hacia delante. A partir de ahí, el algoritmo alterna entre la eliminación hacia atrás de los términos por pasos del modelo, y la entrada hacia delante de los términos fuera del modelo. Se sigue así hasta que no queden términos que cumplan con los criterios de entrada o exclusión.

CUADRO VIII.5. *(Continuación)*

-Pasos sucesivos hacia atrás. Este método se inicia con el modelo que se seleccionaría mediante el método de eliminación hacia atrás. A partir de ahí, el algoritmo alterna entre la entrada hacia delante de los términos fuera del modelo, y la eliminación hacia atrás de los términos por pasos del modelo. Se sigue así hasta que no queden términos que cumplan con los criterios de entrada o exclusión.

Ya para finalizar, en «Construir términos» debemos incluir el tipo de interacciones que queremos medir entre nuestras variables: tenemos seis opciones:

-Interacción: Incluye la interacción seleccionada.

-Efectos principales: Incluye los efectos principales de las variables.

-Todas de 2: Incluye todas las interacciones de dos variables.

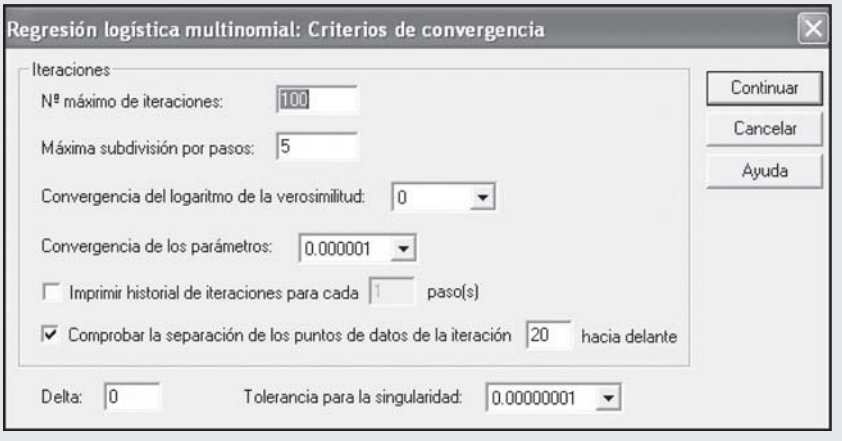
-Todas de 3: Incluye todas las interacciones de tres variables.

-Todas de 4: Incluye todas las interacciones de cuatro variables.

-Todas de 5: Incluye todas las interacciones de cinco variables.

El método de la regresión por pasos será «Pasos sucesivos hacia atrás» y en «Construir términos» indicaremos «Efectos principales» ya que supondremos que no hay interacción entre las variables independientes. Pulsamos sobre el icono «Continuar» para regresar a la pantalla principal.

Paso 5. En la pantalla principal (paso 2) pulsamos sobre el icono «Criterios» para indicar los criterios a la hora de ejecutar el modelo.



CUADRO VIII.5. *(Continuación)*

En esta ventana podemos modificar las siguientes opciones:

- *Número máximo de iteraciones* de los algoritmos del modelo.
- *El número de pasos* en la subdivisión por pasos.
- *La tolerancia de convergencia del logaritmo de la verosimilitud* (se asume convergencia si el cambio absoluto es menor que el valor especificado (no negativo), en caso de ser 0 no se aplica este criterio.
- *La tolerancia de convergencia de los parámetros* (se asume convergencia si el cambio absoluto en las estimaciones de los parámetros es menor que el valor especificado (no negativo), en caso de ser 0 no se aplica este criterio.
- Imprimir el *historial de las iteraciones* para cada paso.
- Indicar *Delta* (entre 0 y 1). Se añade a cada casilla vacía de la tabla de contingencia, ayudando a estabilizar el algoritmo y evitar sesgos en las estimaciones.
- Indicar la *tolerancia para la singularidad*. Si algún elemento del modelo tiene una tolerancia menor, se excluye.

En nuestro caso dejamos las opciones por defecto del programa. Pulsamos sobre el icono «Continuar» para volver a la pantalla principal.

Paso 6. En la pantalla principal (paso 2) pulsamos sobre el icono «Opciones» para indicar las opciones en el proceso del modelo.

CUADRO VIII.5. *(Continuación)*

Las distintas opciones que podemos indicar son:

- *Escala de dispersión*. Especifica el valor de escalamiento de la dispersión que se va a utilizar para corregir la estimación de la matriz de covarianzas de los parámetros. Existen tres tipos: *Desviación* (mediante el estadístico de la función de desviación-χ^2 de la razón de verosimilitud), *Pearson* (mediante el estadístico χ^2 de Pearson) y también se puede especificar su propio valor de escalamiento. Debe ser un valor numérico positivo.

- *Opciones de pasos sucesivos*. Sólo si se utilizan métodos por pasos para generar un modelo. Las opciones que podemos modificar son:

  · *Probabilidad de entrada.* Con mayor probabilidad especificada, más fácil resultará que una variable entre en el modelo.

  · *Prueba de entrada.* Método para introducir los términos en los métodos por pasos. Escoge entre la prueba de la razón de verosimilitud y la prueba de puntuación.

  · *Probabilidad de eliminación.* En los métodos por pasos. Cuanto mayor sea la probabilidad especificada, más fácil resultará que una variable permanezca en el modelo.

  · *Prueba de eliminación.* Éste es el método utilizado para eliminar términos en los métodos por pasos. Puede elegir entre la prueba de la razón de verosimilitud o la prueba de Wald.

  · *Efectos por pasos mínimos en el modelo* (para métodos hacia atrás). Mínimo número de términos que puede incluirse en el modelo.

  · *Efectos por pasos máximos en el modelo* (para métodos hacia adelante). Especifica el máximo número de términos que puede incluirse en el modelo.

- *Restringir jerárquicamente la entrada y la eliminación de términos.* Permite aplicar restricciones a la inclusión de términos de modelo. La jerarquía precisa que para que se incluya un término, todos los inferiores que formen parte del que se desea incluir, se encuentren antes en el modelo.

En nuestro caso elegimos las opciones que el programa tiene por defecto y pulsamos sobre el icono «Continuar» para volver a la pantalla principal.

CUADRO VIII.5. *(Continuación)*

Paso 7. En la pantalla principal (paso 2) pulsamos sobre el icono «Estadísticos» para determinar las pruebas estadísticas que queremos que salgan en los resultados.

Los estadísticos son:

- *Resumen de procesamiento de los casos.* Información sobre las variables categóricas especificadas.

- En la sección *Modelo:*

  · *Pseudo R cuadrado.* Calcula el estadístico de Cox y Snell, de Nagelkerke y el r^2 McFadden. De estos tres *pseudo-*r^2 (imitación del r^2 de Pearson) únicamente el estadístico de Nagelkerke varía de 0 a 1.

  · *Resumen de pasos.* Indica las variables introducidas o eliminadas.

  · *Información de ajuste de los modelos.* Compara el modelo generado frente a la ausencia de modelo.

  · *Criterios de información.* Indica el criterio de información de Akaike (AIC) como el criterio de información bayesiano (BIC), basados en la «Teoría de la información» ayudan a seleccionar el mejor modelo, que es el aquel que tenga un menor valor en estos criterios.

  · *Probabilidades de casillas.* Devuelve las frecuencias observadas y esperadas (con los residuos) y las proporciones por patrón en las covariables y por categoría de respuesta.

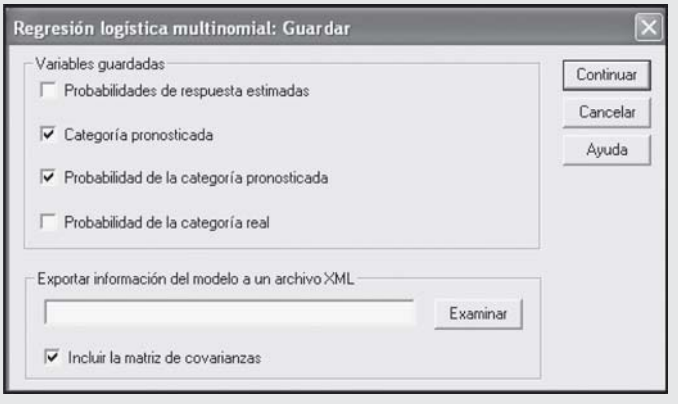
CUADRO VIII.5. *(Continuación)*

  - · *Tabla de clasificación.* Tabla de las respuestas observadas respecto a las respuestas pronosticadas. Un alto porcentaje de acierto indica que la regresión es adecuada.
  - · *Bondad del ajuste.* Indica los estadísticos de χ^2 de Pearson y de χ^2 de la razón de verosimilitud.

- *Parámetros.* Estadísticos relativos a los parámetros del modelo.
  - · *Estimaciones.* Parámetros del modelo con un nivel de confianza especificado.
  - · *Contraste de la razón de verosimilitud.* Razón de verosimilitud para los efectos parciales del modelo.
  - · *Correlaciones asintóticas.* Matriz de las correlaciones entre las estimaciones de los parámetros.
  - · *Covarianzas asintóticas*. Matriz de las covarianzas de las estimaciones de los parámetros.
- *Definir subpoblaciones.* Permite seleccionar un subconjunto de factores y covariables de manera que pueda definir los patrones en las covariables utilizados por las probabilidades de casilla y las pruebas de bondad de ajuste.

Con nuestros datos seleccionamos las opciones que se muestran en la ventana anterior. Pulsamos sobre «Continuar» para volver a la pantalla principal.

Paso 8. En la pantalla principal (paso 2) pulsamos sobre el icono «Guardar» y seleccionamos los datos que queremos que el programa guarde como nuevas variables.



CUADRO VIII.5. *(Continuación)*

En nuestro caso seleccionamos «Categoría pronosticada» y «Probabilidad de la categoría pronosticada» y pulsamos sobre el icono «Continuar», para volver a la pantalla principal, en la que pulsamos en el icono «Aceptar» para ver el archivo de los resultados del modelo.

Paso 9. Interpretación de los resultados. La interpretación de resultados es similar a la descrita para la regresión logística binaria. La tabla «Resumen de los pasos» nos indica las variables introducidas en el modelo. En este caso vemos que solo hay un paso porque todas las variables eran significativas. Además, en caso de haber introducido las interacciones entre las variables independientes, si hubiera prescindido de algún término los χ^2 correspondientes aparecerían en esta tabla (en nuestro caso habíamos seleccionado solo efectos principales y no sale ningún contraste).

Resumen de los pasos

Modelo	Acción	Efecto(s)	Criterio de ajuste del modelo: -2 log verosimilitud	Contrastes de selección de efectos: Chi-cuadrado[a,b]
Paso 0 0	Introducido	<todos>[c,d]	23,480	.

Método por pasos: Por pasos hacia atrás

En la tabla «Información del ajuste del modelo» podemos ver el grado de significación del modelo cuando se compara con el que no tiene ninguna variable independiente (solo la constante). Lo hace a partir de un estadístico χ^2, determinando la probabilidad de obtener un χ^2 de ese valor o mayor si los coeficientes fuesen nulos. En este ejemplo, un $p < 0{,}001$ indica que el modelo final es significativamente distinto del que solo tiene la constante.

Información del ajuste del modelo

Modelo	Criterio de ajuste del modelo: -2 log verosimilitud	Contrastes de la razón de verosimilitud: Chi-cuadrado	gl	Sig.
Sólo la intersección	218,896			
Final	23,480	195,416	18	,000

CUADRO VIII.5. *(Continuación)*

En la tabla «Bondad de ajuste» realiza dos χ^2 diferentes con las desviaciones entre los valores observados y la predicción. Como $p > 0,05$ las desviaciones son pequeñas, es decir, el modelo se ajusta a los datos.

Bondad de ajuste

	Chi-cuadrado	gl	Sig.
Pearson	25,345	114	1,000
Desviación	23,480	114	1,000

Los valores de «Pseudo R-cuadrado» son altos, indicando que un alto porcentaje de la variabilidad de los datos está explicada por el modelo. Según el r^2 de Nagelkerke explica un 97,7% de la variabilidad.

Pseudo R-cuadrado

Cox y Snell	,916
Nagelkerke	,977
McFadden	,893

Los «Contrastes de la razón de verosimilitud» del modelo final nos indican cuáles son las variables independientes significativas en el modelo. Como podemos comprobar, el modelo acepta todas las variables introducidas con $p < 0,05$.

Contrastes de la razón de verosimilitud

	Criterio de ajuste del modelo	Contrastes de la razón de verosimilitud		
Efecto	-2 log verosimilitud del modelo reducido	Chi-cuadrado	gl	Sig.
Intersección	23,480[a]	,000	0	.
Curvatura	45,875	22,395	6	,001
Nivel_Digestion	65,950	42,470	6	,000
Longitud	150,876	127,395	3	,000
Ancho	84,036	60,556	3	,000

El estadístico de chi-cuadrado es la diferencia en las -2 log verosimilitudes entre el modelo final y el modelo reducido. El modelo reducido se forma omitiendo un efecto del modelo final. La hipótesis nula es que todos los parámetros de ese efecto son 0.

a. Este modelo reducido es equivalente al modelo final ya que la omisión del efecto no incrementa los grados de libertad.

CUADRO VIII.5. *(Continuación)*

El programa estima una función polinómica para cada clase (grupo productor) excepto para la última (*Poliquetos*), que es la categoría de referencia. En la tabla «Estimaciones de los parámetros» (columna B) se muestran las estimaciones de todos los parámetros o coeficientes que multiplican a las variables independientes (de acuerdo con la función expresada al principio de este apartado VIII.2.2.), incluidas las que crea el programa para cada nivel (excepto uno) de las variables categóricas. El número total de parámetros estimados es muy elevado, y el número de datos resulta insuficiente, por lo que no es sorprendente que muchos de ellos aparezcan como no significativos. Un valor de Exp(B) alejado de la unidad (intervalo de confianza que no incluye el 1) identifica a los factores que mejor determinan la pertenencia a cada grupo. Observando los intervalos de confianza (solamente aquellos que no incluyen el valor 1), vemos como la *Curvatura=1* multiplica por 68,97 la probabilidad de pertenencia al grupo productor *Mejillones*, y el *Nivel_Digestión=2* multiplica por un número mucho mayor (295614,05) la probabilidad de pertenencia al mismo grupo productor, siendo por lo tanto estas dos características las que identifican a la clase *Mejillones,* mientras que en el segundo grupo *Copépodos* la característica más relevante es *Curvatura=1*, aunque en sentido excluyente: en este caso la probabilidad de pertenencia al grupo se multiplica por cero (Exp(B) prácticamente 0) cuando *Curvatura=1*. Estas características permiten identificar correctamente al 100% de los elementos de la muestra pertenecientes a las clases *Mejillones* y *Copépodos*.

Estimaciones de los parámetros

grupo_productor[a]		B	Error típ.	Wald	gl	Sig.	Exp(B)	Intervalo de confianza al 95% para Exp(B)	
								Límite inferior	Límite superior
Mejillones	Intersección	-91,288	70796,793	,000	1	,999			
	[Curvatura=1,00]	4,234	,000	.	1	.	68,970	68,970	68,970
	[Curvatura=2,00]	6,616	17966,770	,000	1	1,000	747,256	,000	.[b]
	[Curvatura=3,00]	0[c]	.	.	0	.	.	.	.
	[Nivel_Digestion=1,00]	15,799	14638,925	,000	1	,999	7269524	,000	.[b]
	[Nivel_Digestion=2,00]	12,597	,000	.	1	.	295614,1	295614,050	295614,1
	[Nivel_Digestion=3,00]	0[c]	.	.	0	.	.	.	.
	Longitud	,164	55,952	,000	1	,998	1,178	2,78E-048	5,0E+047
	Ancho	,335	878,972	,000	1	1,000	1,398	,000	.[b]
Copépodos	Intersección	962,101	67854,383	,000	1	,989			
	[Curvatura=1,00]	-174,680	,000	.	1	.	1,37E-076	1,37E-076	1,37E-076
	[Curvatura=2,00]	-111,480	18632,379	,000	1	,995	3,84E-049	,000	.[b]
	[Curvatura=3,00]	0[c]	.	.	0	.	.	.	.
	[Nivel_Digestion=1,00]	-234,962	25279,185	,000	1	,993	9,07E-103	,000	.[b]
	[Nivel_Digestion=2,00]	-118,177	15564,271	,000	1	,994	4,75E-052	,000	.[b]
	[Nivel_Digestion=3,00]	0[c]	.	.	0	.	.	.	.
	Longitud	-2,244	190,855	,000	1	,991	,106	3,71E-164	3,0E+161
	Ancho	-7,840	696,348	,000	1	,991	,000	,000	.[b]
Larvas crustáceos	Intersección	2,665	9,038	,087	1	,768			
	[Curvatura=1,00]	-2,065	3,270	,399	1	,528	,127	,000	77,038
	[Curvatura=2,00]	,843	2,330	,131	1	,717	2,324	,024	223,406
	[Curvatura=3,00]	0[c]	.	.	0	.	.	.	.
	[Nivel_Digestion=1,00]	-,501	3,142	,025	1	,873	,606	,001	286,275
	[Nivel_Digestion=2,00]	-,871	2,117	,169	1	,681	,419	,007	26,524
	[Nivel_Digestion=3,00]	0[c]	.	.	0	.	.	.	.
	Longitud	,029	,027	1,196	1	,274	1,030	,977	1,085
	Ancho	-,141	,073	3,757	1	,053	,869	,753	1,002

a. La categoría de referencia es: Poliquetos.

b. Se ha producido un desbordamiento de punto flotante al calcular este estadístico. Por lo tanto, el valor asignado ha sido el valor perdido del sistema.

c. Este parámetro se ha establecido a cero porque es redundante.

CUADRO VIII.5. *(Continuación)*

El grupo *Larvas crustáceos* no tiene ninguna variable independiente que permita su identificación con claridad (los intervalos para exp(B) incluyen el valor 1 en todos los casos), y para la cuarta categoría, o grupo de referencia *Poliquetos* no disponemos de estimaciones de los parámetros (la probabilidad de asignación se calcula por diferencia a 1 para este grupo), aunque podrían obtenerse si se repite el análisis cambiando la categoría de referencia. En estas dos últimas categorías, el porcentaje de acierto es menor, como se aprecia en la tabla siguiente.

En la tabla «Clasificación» se observa que el modelo clasifica correctamente un 89,9%. Clasifica perfectamente los pellets producidos por *Mejillones* y *Copépodos* y comete un error de alrededor de un 20% en las otras dos categorías.

Clasificación

Observado	Pronosticado				
	Mejillones	Copépodos	Larvas crustáceos	Poliquetos	Porcentaje correcto
Mejillones	19	0	0	0	100,0%
Copépodos	0	21	0	0	100,0%
Larvas crustáceos	0	0	16	4	80,0%
Poliquetos	0	0	4	15	78,9%
Porcentaje global	24,1%	26,6%	25,3%	24,1%	89,9%

En el caso de la regresión logística multinomial existe una función de probabilidad para cada grupo productor (*Mejillones*, *Poliquetos*, *Copépodos* y *Larvas de crustáceos*). Por lo tanto, para poder estimar el grupo de pertenencia pronosticado es más fácil a través de los cálculos que realiza el propio programa y que se explicó como guardarlos en el paso 8. Volviendo a la matriz de datos vemos que se han creado dos nuevas variables «PRE 1» (Categoría pronosticada) y «PCP_1» (Probabilidad de la categoría pronosticada). Para calcular la «Categoría pronosticada» y «Probabilidad de la categoría pronosticada» de nuevos casos, estos se introducen en la matriz de datos y se vuelve a ejecutar el programa.

	grupo_productor	Longitud	Ancho	Curvatura	Nivel_Digestion	PRE_1	PCP_1
1	Mejillones	713,00	60,00	Hasta 10%	Poco digerido	Mejillones	1,00
2	Mejillones	825,00	43,00	Hasta 10%	Poco digerido	Mejillones	1,00
3	Mejillones	609,00	61,00	Recto	Digestion media	Mejillones	1,00
4	Mejillones	781,00	83,00	Más de un	Poco digerido	Mejillones	1,00
5	Mejillones	598,00	73,00	Hasta 10%	Muy digerido	Mejillones	1,00
6	Mejillones	433,00	69,00	Recto	Poco digerido	Mejillones	1,00
7	Mejillones	825,00	43,00	Hasta 10%	Poco digerido	Mejillones	1,00
8	Mejillones	615,00	58,00	Hasta 10%	Digestion media	Mejillones	1,00
9	Mejillones	781,00	68,00	Recto	Poco digerido	Mejillones	1,00
10	Mejillones	598,00	73,00	Hasta 10%	Muy digerido	Mejillones	1,00