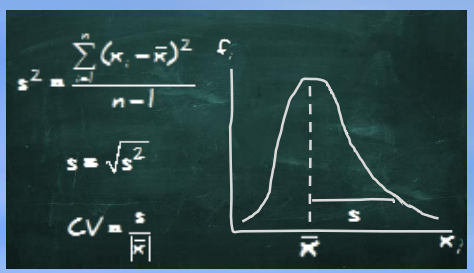
METODOLOGÍA DE LA INVESTIGACIÓN SOCIAL CUANTITATIVA

Pedro López-Roldán
Sandra Fachelli



UAB
Universitat Autònoma de Barcelona

METODOLOGÍA DE LA INVESTIGACIÓN SOCIAL CUANTITATIVA

Pedro López-Roldán
Sandra Fachelli

Bellaterra (Cerdanyola del Vallès) | Barcelona
Dipòsit Digital de Documents
Universitat Autònoma de Barcelona

Pedro López-Roldán
Centre d'Estudis Sociològics sobre la Vida Quotidiana i el Treball (http://quit.uab.cat)
Institut d'Estudis del Treball (http://iet.uab.cat/)
Departament de Sociologia. Universitat Autònoma de Barcelona
pedro.lopez.roldan@uab.cat

Sandra Fachelli
Departament de Sociologia i Anàlisi de les Organitzacions
Universitat de Barcelona
Grup de Recerca en Eduació i Treball (http://grupsderecerca.uab.cat/gret)
Departament de Sociologia. Universitat Autònoma de Barcelona
sandra.fachelli@ub.edu

Edición digital: http://ddd.uab.cat/record/129382

1ª edición, febrero de 2015

Edifici B · Campus de la UAB · 08193 Bellaterra
(Cerdanyola del Vallés) · Barcelona · España
Tel. +34 93 581 1676

Índice general

Metodología de la Investigación Social Cuantitativa

Pedro López-Roldán
Sandra Fachelli

PARTE III. ANÁLISIS

Capítulo III.10
Análisis de regresión logística

Bellaterra (Cerdanyola del Vallès) | Barcelona
Dipòsit Digital de Documents
Universitat Autònoma de Barcelona

Cómo citar este capítulo:

López-Roldán, P.; Fachelli, S. (2016). Análisis de regresión logística. En P. López-Roldán y S. Fachelli, *Metodología de la Investigación Social Cuantitativa*. Bellaterra (Cerdanyola del Vallès): Dipòsit Digital de Documents, Universitat Autònoma de Barcelona. 1ª edición. Edición digital: http://ddd.uab.cat/record/163570.

Capítulo acabado de redactar en septiembre de 2016

Índice de contenidos

PARTE III
Capítulo 10

Análisis de regresión logística

El análisis regresión logística es una técnica estadística multivariable destinada al análisis de una relación de dependencia entre una variable dependiente y un conjunto de variables independientes, de forma similar a como actúa el análisis de regresión lineal clásico. El objetivo del análisis es poder efectuar predicciones del comportamiento, esto es, estimar las probabilidades de un suceso definido por la variable dependiente en función de un conjunto de variables predictoras o de pronóstico.

En el modelo clásico de regresión lineal la variable dependiente es cuantitativa, condición que se extiende a las variables independientes si bien podemos utilizar variables cualitativas con una codificación *dummy*. En el caso de la regresión logística se trata de predecir una variable cualitativa o categórica, con la ventaja, frente al modelo de regresión clásico, de no tener que establecer la serie de condiciones de aplicación que dificultan su utilización y sus posibilidades, en particular, en el contexto de estudios por encuesta.

La técnica de la regresión logística se origina en la década de los años 60 con el trabajo de Cornfield, Gordon y Smith (1961). Walter y Duncan (1967) ya la utilizan en la forma actual, siendo a partir de los años 80, con la ayuda de la informática aplicada, que se generaliza su uso. La regresión logística mezcla dos tradiciones del análisis estadístico: el análisis de tablas de contingencia con el tratamiento de modelos log-lineales, y el análisis de regresión por mínimos cuadrados ordinarios. En ambos casos nos encontramos con limitaciones que la regresión logística resuelve: en el primer caso los modelos de dependencia no podían utilizar variables continuas y en el segundo las variables categóricas no siempre funcionan como buenos predictores.

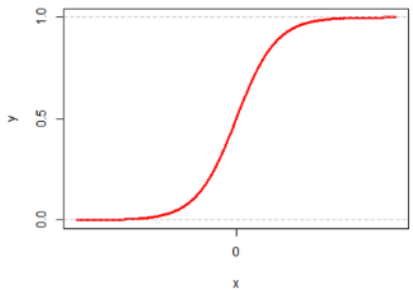
A diferencia de la regresión lineal pues, con la regresión logística el objetivo es explicar o pronosticar la pertenencia a un grupo, a partir de una variable dependiente categórica o cualitativa, en función de una o más variables independientes que pueden ser tanto cuantitativas como cualitativas. Se trata por tanto de identificar qué características o factores diferencian los grupos definidos por la variable dependiente, de forma similar a como lo hace el análisis discriminante, pero con la ventaja de poder considerar cualquier nivel de medición de las variables independientes.

Al considerar al análisis de regresión logística como técnica destinada al análisis de una relación de dependencia, nos referiremos fundamentalmente a ella como una técnica predictiva y no tanto como técnica destinada a establecer relaciones de causalidad, si bien implícitamente se razone la causalidad. Esto es, cuando diferenciamos a las variables independientes de la dependiente establecemos un modelo explicativo donde se fijan los factores que tienden a favorecer un efecto, a aumentar probabilidades de un comportamiento, que se dé un valor de la variable de dependiente.

El análisis de regresión logística tiene dos modalidades: la regresión logística binaria cuando se pretende explicar una característica o suceso dicotómico (estar desempleado o no, abstenerse en las elecciones o no), y la regresión logística multinomial en el caso más general de querer explicar una variable cualitativa politómica. Para ello se requiere convertir la variable en diversas variables dicotómicas ficticias, es decir, creando tantas variables dicotómicas (*dummy*) como categorías tenga la variable menos una, la que actuará de categoría de referencia. En este segundo caso se diferencia la situación en que la variable categórica es politómica nominal (la elección de una marca de un producto o la filiación política) o politómica ordinal (el nivel salarial o el grado de acuerdo sobre una cuestión).[1]

Por otro lado, desde el punto de vista de las variables independientes, éstas pueden ser cualitativas, tanto dicotómicas como politómicas, o cuantitativas, y se puede considerar tanto el efecto individual de cada una como el efecto de la interacción.

El modelo se formaliza a partir de de la función logística, donde se plantea una relación funcional, aquí la función es la logística, una función con forma de curva sigmoidal:



Como veremos, las relaciones entre las variables se expresan como una función exponencial y los parámetros de la ecuación se interpretan de forma multiplicativa. A través de su transformación logarítmica la relación de dependencia se puede interpretar en términos lineales, como suma de efectos.

Con estas características generales el análisis de regresión logística persigue cuantificar la importancia de la relación existente entre cada una de las variables independientes (también llamadas covariables) y la variable dependiente, y clasificar a los individuos dentro de las categorías de la variable dependiente según la probabilidad que tenga de pertenecer a una de ellas dada la influencia de las covariables.

[1] Junto a estas dos modalidades de regresión logística, y persiguiendo los mismos objetivos con el mismo tipo de variables, existen dos técnicas adicionales de interés: el análisis de regresión ordinal donde la técnica modeliza el hecho que la variable dependiente tiene sus valores ordenados (por ejemplo, quintiles de ingreso) o el análisis de regresión probit, técnica equivalente que en vez de trabajar con la función de enlace logística lo hace con la normal.

La regresión logística se puede presentar igualmente como un caso particular de un modelo lineal generalizado (MLG)[2].

El proceso de análisis de una regresión logística se puede dividir en varias etapas o tareas:

1) Selección de las variables del modelo

Una primera tarea fundamental es que las variables se justifiquen en el contexto de unos objetivos de investigación y a partir de criterios teóricos que fundamenten la relación de dependencia. El modelo teórico puede ser más o menos elaborado. En el primer caso procedemos con una lógica más deductiva a partir de un modelo claramente definido que orienta y define las variables y la relación de dependencia. En el segundo caso también debemos disponer de criterios teóricos básicos que justifiquen la selección de las variables y una propuesta de modelo, pero sin una formulación definitiva ni cerrada que nos posibilita el utilizar criterios adicionales de tipo estadístico o empíricos para seleccionar las variables determinantes.

Con las variables seleccionadas inicialmente para el modelo se puede seguir un procedimiento como el siguiente. Hay que determinar, por un lado, el nivel de asociación entre cada variable independiente o explicativa por separado y la variable dependiente, lo que nos permitirá descartar aquellas variables que empíricamente no manifiestan una relación con la variable a explicar y nos puede sugerir asimismo posibles agrupaciones de valores de las variables. No obstante, a este resultado se puede llegar igualmente con el tratamiento conjunto de las variables iniciales sin necesidad de individualizar su relación con la variable dependiente, pero la mirada individual siempre proporciona elementos conclusivos parciales de interés que nos ayudan a construir la mirada multidimensional. Por otra parte, con las variables independientes hay que efectuar un análisis de sus interrelaciones o interacciones con el fin de constatar la existencia de multicolinealidad y determinar el nivel de asociación entre las variables no colineales y la variable dependiente. Finalmente se dispone del modelo de dependencia inicial que hay que analizar para llegar a determinar el modelo final que mejor explique la variable dependiente.

Se trata, por tanto, de un proceso en el que se estimarán varias ecuaciones de regresión logística a partir de diferentes modelos alternativos entre los que escogeremos el que mejor se ajuste a los datos para explicar la variable dependiente. Veremos que los procedimientos con el software estadístico incorporan procesos automatizados de selección del modelo de gran ayuda en esta tarea.

[2] J. Nelder y R. W. Wedderburn en el año 1972 introdujeron el concepto de modelo lineal generalizado y constituye la generalización de los modelos lineales clásicos incluyendo como casos particulares la regresión lineal, el análisis de la varianza, el análisis de la covarianza, la regresión de Poisson, la regresión logística, la regresión logit, los modelos log-lineales, los modelos de respuesta multinomial, así como ciertos modelos de análisis de la supervivencia y de series temporales. Los modelos lineales generalizados se caracterizan por tener tres componentes: uno aleatorio (referido a la variable dependiente), otro sistemático (el de las variables predictoras) y un componente de enlace (Jaccard, 2001: 3-4). En este caso la función de enlace es la logística.

2) Estimación de los coeficientes de las variables independientes

La estimación de los coeficientes o pesos de la ecuación de regresión que determinan la importancia de cada variable independiente en la explicación de la dependiente se realiza mediante un algoritmo iterativo de máxima verosimilitud propio del modelo de la regresión logística.

A lo largo del proceso de búsqueda del mejor modelo de regresión se realizan las estimaciones de los coeficientes de cada posible modelo y se valora su bondad de ajuste.

3) Clasificación de los casos

En función de la ecuación de regresión logística estimada se procede a la clasificación de los individuos según la variable dependiente pronosticada. En función del criterio de probabilidad de corte establecido un individuo es asignado a cada categoría de la variable dependiente. De este modo tenemos dos clasificaciones: la inicial que establece la variable dependiente observada, y la pronosticada en función del modelo de regresión logística. El cruce de ambas clasificaciones nos proporciona los casos que están correctamente clasificados y los que no. El porcentaje de casos bien clasificados es un indicador de la capacidad explicativa o discriminatoria del modelo.

4) Análisis de los residuos

Con el modelo seleccionado, se puede proceder a realizar un análisis más detallado de los residuos con el fin de detectar la existencia de casos extremos, casos que difieren notablemente (más de dos unidades de desviación) entre la probabilidad observada y la probabilidad pronosticada por el modelo, y cuya eliminación puede mejorar el ajuste del modelo.

1. Análisis de regresión logística binaria simple

La regresión logística binaria se caracteriza por disponer de una variable dependiente cualitativa con dos valores (categorías o grupos) que configuran la presencia y la ausencia de una determinada característica. Por ejemplo, los ciudadanos que se abstienen en las elecciones y los que no, los que votan a un partido y los que no, los consumidores que compran un producto y los que no, las personas que están en paro y las que no, las personas que reinciden en un delito y las que no, las personas que tienen un riesgo contraer una enfermedad y las que no, las que devolverán un préstamo y las que no, etc.

La característica definida por la variable dependiente se pretende explicar en función de una serie de variables independientes o predictoras que nos determinan en qué se diferencian los dos grupos. Si consideramos tan sólo una variable independiente podemos hablar de regresión logística simple, si consideramos dos o más variables independientes el modelo de regresión logística es múltiple. En el contexto de la regresión logística estas variables se denominan también covariables. Como resultado del análisis se obtienen unos pesos o coeficientes que nos miden la importancia de cada

variable independiente para diferenciar los grupos, y en segundo término obtenemos criterios para pronosticar la clasificación de los individuos o casos.

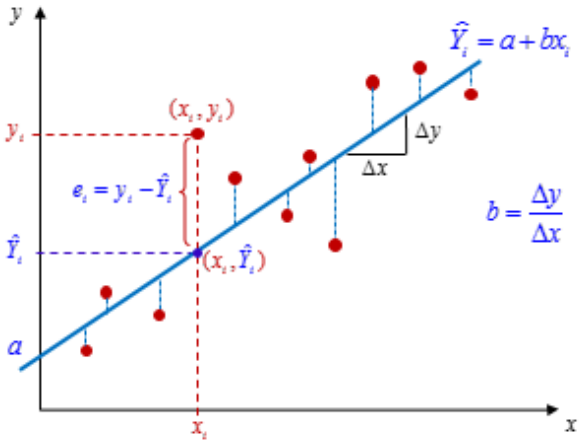
1.1. La relación logística

En el modelo de regresión lineal la relación entre las variables se expresa de forma general como:

$$y_i = a + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip} + e_i \qquad \textit{Ecuación 1}$$

En el caso particular del modelo de regresión simple, la ecuación $\hat{Y}_i = a + bx_i$, se representa gráficamente mediante una recta en el plano que se ajusta por el método de mínimos cuadrados (Gráfico III.10.1).

Gráfico III.10.1. Modelo lineal de regresión simple



La recta de regresión se extiende de forma ilimitada entre $-\infty$ i $+\infty$, si bien los valores de la recta de regresión se interpretan en el rango de valores de x observados en la muestra y tienen un sentido interpretativo, descartando valores de predicción imposibles a partir de los datos estudiados. No obstante, también puede suceder que a pesar de considerar el rango de valores de la muestra los valores pronosticados sean valores imposibles. Es el caso que se puede dar cuando consideramos en la regresión lineal variables dicotómicas de la variable dependiente, codificadas con 0 y 1, donde los valores predichos pueden ser inferiores a 0 y superiores a 1, fuera del rango definido por la variable dependiente.

La regresión logística resuelve este tipo de problema usando una función no lineal como es la función logística. Con esta función se pueden efectuar predicciones comprendidas entre un mínimo y un máximo. El modelo de regresión logística es un modelo no lineal que utiliza el método de máxima verosimilitud, un procedimiento iterativo que en fases sucesivas ajusta el modelo.

La formulación matemática de la curva logística en el caso de la regresión logística binaria simple es:

$$y = Pr(y = 1 | x) = \frac{e^{a+bx}}{1+e^{a+bx}}$$ *Ecuación 2*

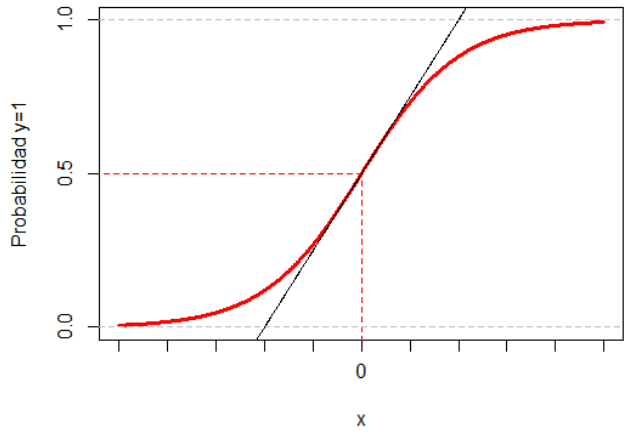
o bien, de forma equivalente:

$$y = Pr(y = 1 | x) = \frac{1}{1+e^{-(a+bx)}}$$ *Ecuación 3*

Es decir, la probabilidad de que la variable dependiente y tome el valor 1 (presencia de la característica estudiada) en función de la variable independiente x.

La representación gráfica de la función logística, de expresión general $y = f(x) = \frac{1}{1+e^{-x}}$, es una curva con forma sigmoidea (Gráfico III.10.2):

Gráfico III.10.2. La curva logística



que verifica las propiedades siguientes:

- Sus valores oscilan entre 0 y 1, $0 < f(x) < 1$, lo que permite interpretarla en términos de probabilidad.
- Su límite inferior es el valor 0 : $\lim_{x\to-\infty} \frac{1}{1+e^{-x}} = 0$
- Su límite superior es el valor 1 : $\lim_{x\to\infty} \frac{1}{1+e^{-x}} = 1$
- Cuando la x vale 0 la función vale ½ : $f(0) = \frac{1}{1+e^{-0}} = \frac{1}{2}$

La ecuación de la función logística permite asignar valores a la variable independiente para generar valores de la dependiente de la misma forma que en la regresión lineal, y su interpretación es similar. Pero en este caso los valores de predicción de la variable independiente y se situarán siempre en el intervalo $(0,1)$, lo que facilita interpretar los resultados y los parámetros de la ecuación en términos de probabilidad para pronosticar un comportamiento.

1.2. El modelo de la regresión logística

El modelo de regresión logística binaria considera dos sucesos de un fenómeno o variable Y, excluyentes y exhaustivos, que se codifican con valores 0 y 1. Si la probabilidad de que suceda uno de ellos es P, la probabilidad de que suceda la otro es igual a 1 menos la probabilidad P:

$$Pr(y=1) = P$$
$$Pr(y=0) = 1-P$$

La cuestión es considerar la información de una (o más variables en la versión múltiple) para definir un modelo que permita pronosticar la probabilidad de la variable dependiente y, es decir, se trata de encontrar una o más variables que discriminen bien entre los dos posibles valores de la variable y.

En un modelo de regresión logística binaria simple, la ecuación logística se expresa como:

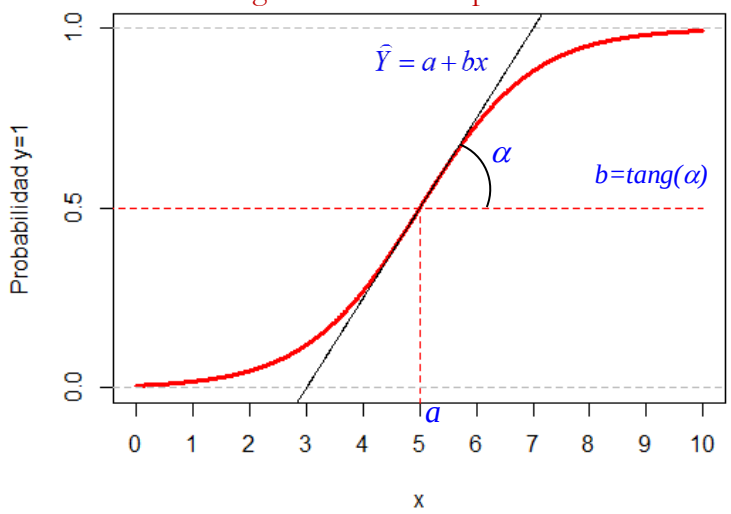
$$\Pr(y=1) = \frac{1}{1+e^{-(a+bx)}} = P \qquad \textit{Ecuación 4}$$

y, por tanto,

$$\Pr(y=0) = 1-\left(\frac{1}{1+e^{-(a+bx)}}\right) = 1-P \qquad \textit{Ecuación 5}$$

Con la representación gráfica adjunta (Gráfico III.10.3).

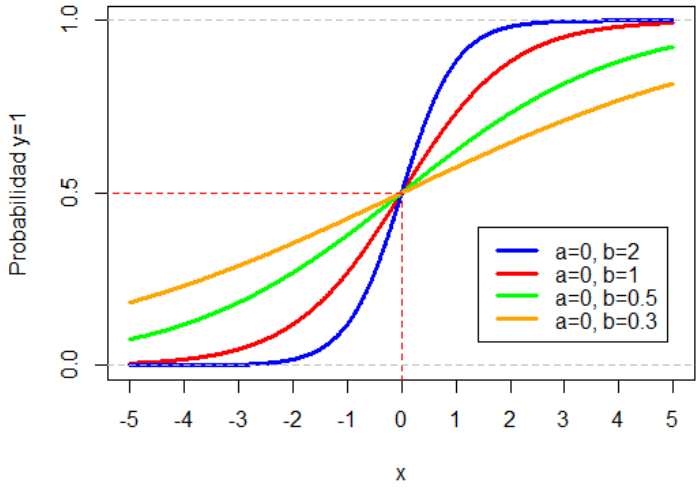
Gráfico III.10.3. Representación del modelo de regresión logística binaria simple



El coeficiente a representa la posición de la curva sobre el eje horizontal o de abscisas, y sitúa la curva más hacia la derecha o hacia la izquierda. El coeficiente b representa la pendiente de la curva en su punto de inflexión, en función de su valor más alto o más bajo tendremos una pendiente de la curva más inclinada o menos.

Por tanto, nos podemos encontrar con una familia de curvas que varían en función de los valores de *a* y de *b* (Gráfico III.10.4).

Gráfico III.10.4. Familia de curvas logísticas según valores distintos de la pendiente b



La variación de la pendiente implicará una distinta capacidad discriminatoria de los valores de *y*. Una buena variable independiente predictora es la que genera una curva con una elevada pendiente, cuando el valor absoluto de *b* es alto; si *b* se acerca al valor 0 su capacidad predictora se reduce. Por tanto, el objetivo del análisis de regresión logística consiste en encontrar las variables con el mayor coeficiente asociado.

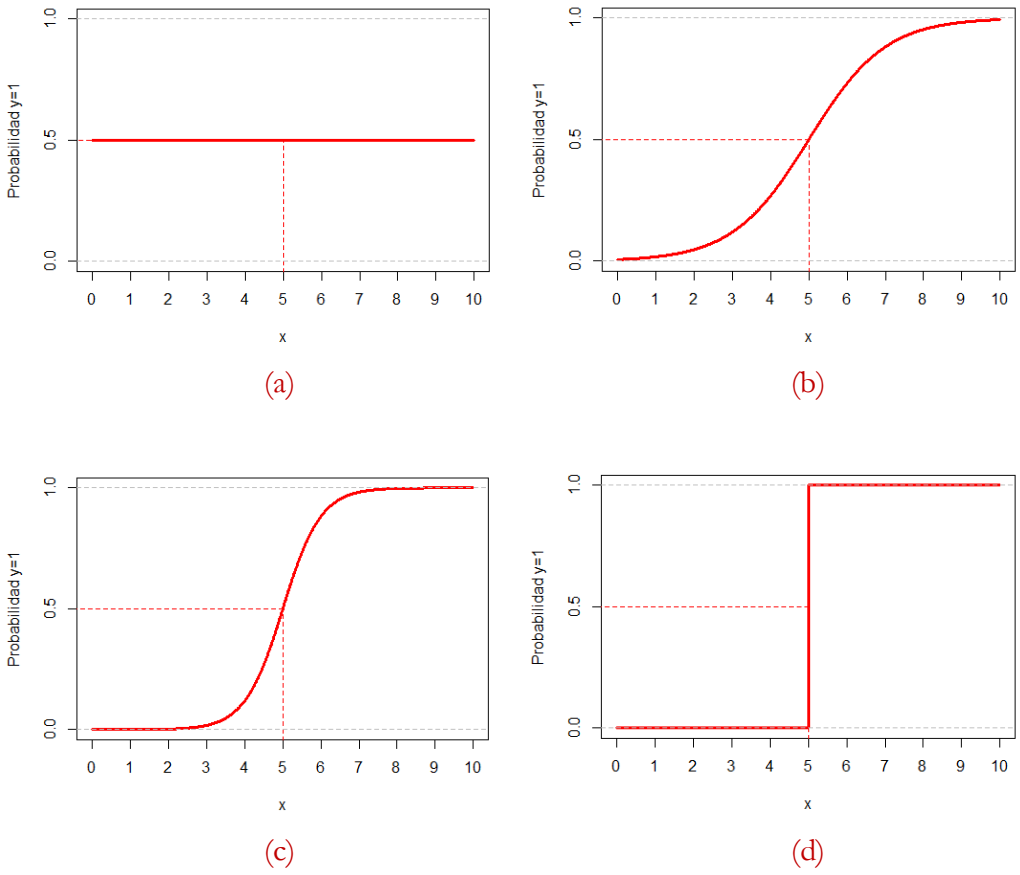
En las cuatro representaciones presentadas en el Gráfico III.10.5 podemos observar cuatro casos de curvas logísticas ordenadas desde la mínima capacidad discriminatoria (figura a) hasta la máxima capacidad de discriminar los valores de la variable dependiente (figura d).

La interpretación de los coeficientes de la regresión logística difiere del caso de la regresión lineal. Aquí el coeficiente no es la medida de cuánto variará *y* ante una variación en una unidad de *x*, sino el cambio producido por una variación de una unidad de *x* en el logaritmo neperiano (*log*) del cociente de probabilidades de los dos sucesos, la denominada transformación logit.

La transformación logit surge de considerar la relación o el cociente de probabilidad entre dos sucesos, llamada ventaja o razón (como traducción de la expresión inglesa *odds*). La razón de un suceso es el cociente entre la probabilidad de que éste suceda y la probabilidad de que no suceda:

$$odds = \frac{P}{1-P} = \frac{Probabilidad\ de\ que\ ocurra\ un\ suceso}{Probabilidad\ de\ que\ no\ ocurra\ un\ suceso} \qquad \textit{Ecuación 6}$$

Gráfico III.10.5. Representación de curvas logísticas con diferente capacidad explicativa



Así por ejemplo, si el 75% de la población vota en unas elecciones (la probabilidad es del 0,75), el 25% se abstiene y el *odds* será 3: $\frac{P}{1-P}=\frac{0,75}{0,25}=3$. De la misma forma que pasamos de las probabilidades a las razones, podemos pasar de las razones a las probabilidades:

$$P=\frac{odds}{odds+1} \qquad \textit{Ecuación 7}$$

Si el odds es 3 la probabilidad es: $\frac{3}{3+1}=0,75$. En ambos casos se cuantifica qué tan probable es un suceso, su "riesgo". El riesgo relativo es el cociente de probabilidades de un suceso en dos condiciones distintas. El *odds ratio* (o razón de razones de probabilidad) es el cociente de dos *odds*. Si en el municipio A vota el 80% y en el municipio B el 50%, el *odds ratio* será 4: $odds\ ratio=\frac{odds\ A}{odds\ B}=\frac{0,8/0,2}{0,5/0,5}=\frac{4}{1}=4$.[3]

[3] El análisis de razones se presentó en los capítulos III.6 y III.7 sobre tablas de contingencia y análisis log-lineal.

A partir de las expresiones de la *Ecuación 4* y la *Ecuación 5*, obtenemos:

$$\frac{\Pr(y=1)}{\Pr(y=0)} = \frac{\frac{1}{1+e^{-(a+bx)}}}{1-\left(\frac{1}{1+e^{-(a+bx)}}\right)} = \frac{P}{1-P} \qquad \textit{Ecuación 8}$$

La expresión se puede simplificar para obtener:

$$\frac{\Pr(y=1)}{\Pr(y=0)} = \frac{P}{1-P} = e^{a+bx} \qquad \textit{Ecuación 9}$$

La representación gráfica de esta expresión es la del Gráfico III.10.6.

Gráfico III.10.6. Representación de $\frac{P}{1-P} = e^{a+bx}$

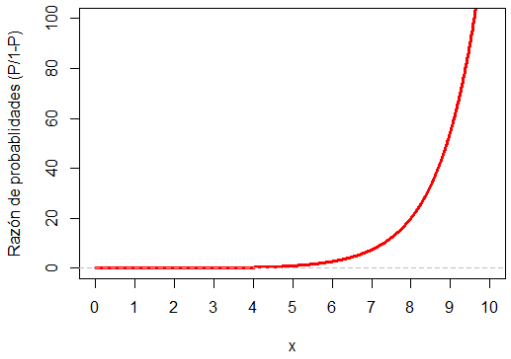
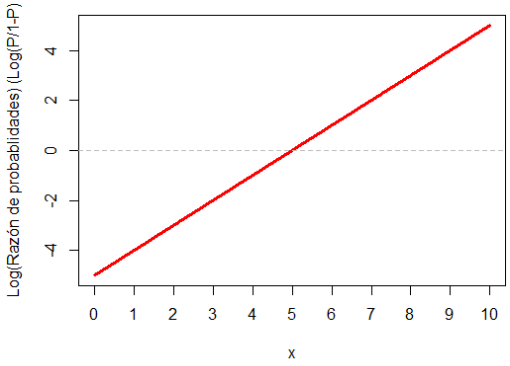


Gráfico III.10.7. Representación de $\log\left(\frac{P}{1-P}\right) = a+bx$



Finalmente, si a la expresión se le aplica la transformación logarítmica nos queda la transformación logit que permite identificar el modelo en forma lineal y aditiva (Gráfico III.10.7):

$$\log\left(\frac{P}{1-P}\right) = a+bx \qquad \textit{Ecuación 10}$$

En consecuencia, el coeficiente de regresión logística b interpreta como el cambio que se produce en la transformación logit, en el logaritmo de la razón de un suceso (del cociente de probabilidades), por cada cambio unitario que se produce en la variable independiente.

Como en el modelo de regresión lineal, realizamos estimaciones de parámetros poblacionales y éstos están afectados por un error de estimación. En el modelo de regresión lineal se asume que los errores estándar de cada coeficiente siguen una distribución normal de media 0 y varianza constante (supuesto de homoscedasticidad). En el caso de la regresión logística el error sigue una distribución binomial, con media y varianza, proporcionales al tamaño muestral y a $Pr(y=1|x)$.

Para obtener los coeficientes de la ecuación de regresión logística y sus correspondientes errores se realizan estimaciones de máxima verosimilitud que maximizan la probabilidad de obtener los valores de la variable dependiente. Estas

estimaciones requieren seguir algoritmos iterativos como el método iterativo de Newton-Raphson.

1.3. Estimación de los parámetros. Ejemplo de aplicación

Veamos a continuación un primer ejemplo de aplicación de un análisis de regresión logística binaria simple. A partir del archivo proporcionado por el software SPSS GSS1993.sav (con datos de la *General Social Survey* de EE.UU., del año 1993) consideraremos la variable VOTO para probar la hipótesis de si el abstencionismo es mayor cuanto menor es el nivel cultural en base a un modelo de regresión logística binaria simple[4]. Con la regresión logística se quiere explicar el abstencionismo (si votó o no a las elecciones, variable dependiente), con estas frecuencias:

VOTO Votó en 1992

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	0 Sí votó	1032	68,8	71,1	71,1
	1 No votó	420	28,0	28,9	100,0
	Total	1452	96,8	100,0	
Perdidos	9 Sin datos	48	3,2		
Total		1500	100,0		

en función de la lectura de los diarios (sí lee o no lee periódicos, variable independiente DIARIOS) como indicador del nivel cultural. Esta variable cualitativa dicotómica se obtiene de agrupar los valores de la variable periódica:

DIARIOS Lee diarios

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	0 Sí lee	862	57,5	85,3	85,3
	1 No lee	148	9,9	14,7	100,0
	Total	1010	67,3	100,0	
Perdidos	9 Sin datos	490	32,7		
Total		1500	100,0		

Para analizar la relación entre ambas, al tratarse de dos variables cualitativas, podríamos optar por realizar una tabla de contingencia:

VOTO Votó en 1992*DIARIOS Lee diarios tabulación cruzada

Categoría de interés

			DIARIOS Lee diarios		
			0 Sí lee	1 No lee	Total
VOTO Votó en 1992	0 Sí votó	Recuento	624	68	692
		% dentro de DIARIOS Lee diarios	74,9%	47,6%	70,9%
	1 No votó	Recuento	209	75	284
		% dentro de DIARIOS Lee diarios	25,1%	52,4%	29,1%
Total		Recuento	833	143	976
		% dentro de DIARIOS Lee diarios	100,0%	100,0%	100%

V Cramer = 0,213

[4] Seguiremos parcialmente el ejemplo utilizado por Pardo y Ruiz (2002), capítulo 28. El archivo de sintaxis ARL-Voto.sps reproduce todos los resultados que se presentan en el texto de este ejemplo.

Como se ve, el abstencionismo es mayor entre los que no leen (52,4%) que entre los que leen (25,1%). La razón de probabilidades global de no votar (valor 1) a votar (valor 0) es:

$$\frac{0,291}{0,709} = 0,410$$

Si lee diarios (x=0): $\frac{0,524}{0,476} = 1,100$

Si no lee diarios (x=0): $\frac{0,251}{0,749} = 0,335$

Cuando la variable independiente varía en una unidad (pasa de leer diarios, valor 0, a no leer, valor 1), la razón de no votar a votar aumenta en 3,293. La ecuación de regresión logística que se obtiene es, en forma aditiva (B) y en forma multiplicativa ($Exp(B)$):

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 1[a]	DIARIOS	1,192	,186	41,258	1	,000	3,293
	Constante	-1,094	,080	187,316	1	,000	,335

a. Variables especificadas en el paso 1: DIARIOS.

$$\log\left(\frac{\frac{0,524}{0,476}}{\frac{0,251}{0,749}}\right) = \log\left(\frac{1,100}{0,335}\right) = \log(3,293) = 1,192$$

Coeficiente en forma aditiva

DIARIOS (No) $\xrightarrow{1,192}$ VOTO (No)

$\xrightarrow{3,293}$

Coeficiente en forma multiplicativa

Parámetros del modelo aditivo:

Si lee diarios, cuando x=0,

$$\log\left(\frac{\Pr(Y=1|X=0)}{\Pr(Y=0|X=0)}\right) = \log\left(\frac{P}{1-P}\right) = \log\left(\frac{0,251}{1-0,251}\right) = \log\left(\frac{0,251}{0,749}\right) = \log(0,33) = -1,094 \quad \Rightarrow a = -1,094$$

Si no lee diarios, cuando x=1,

$$\log\left(\frac{\Pr(Y=1|X=1)}{\Pr(Y=0|X=1)}\right) = \log\left(\frac{P}{1-P}\right) = \log\left(\frac{0,524}{1-0,524}\right) = \log\left(\frac{0,524}{0,476}\right) = \log(1,10) = 0,098$$

El coeficiente b es cuando x pasa de 0 a 1: $b = 0,098 - (-1,094) = 1,192$

Es decir, una variación de una unidad de la variable independiente produce un cambio de 1,192 en la transformación logit. A efectos interpretativos y para no utilizar el logaritmo de la razón, puede resultar más intuitivo interpretar las variaciones de las razones, sin aplicar el logaritmo y, por tanto, considerar las expresiones exponenciales (potencias del número e) en un modelo multiplicativo.

Parámetros del modelo multiplicativo:

Si lee diarios, cuando $x=0$, $exp(a) = e^{a} = e^{-1,094} = 0,335$

Si no lee diarios, cuando $x=1$, $exp(b) = e^{b} = e^{1,192} = 3,293$

El coeficiente b es cuando x pasa de 0 a 1:

De $\frac{0,251}{0,749} = 0,335$ a $\frac{0,524}{0,476} = 1,100 \Rightarrow \frac{0,524}{0,476} \Big/ \frac{0,251}{0,749} = 1,100 / 0,335 = 3,293$

Esta versión nos permite expresar el cambio proporcional de la razón en términos de un cociente de dos razones, es lo que se conoce como razón de razones o razón de ventajas (*odds ratio*). Así interpretamos el valor 3,293 con el cambio proporcional (en este caso de aumento) que se produce en la razón del suceso que analizamos (no votó) por cada unidad de cambio que se produce en la variable independiente. Cuando la variable independiente varía en una unidad (pasa de leer periódicos, valor 0, no leer, valor 1), la razón (el cociente o la relación) de no votar a votar aumenta en 3,293.

En general,

- Si la razón de razones vale 1, esto significa que el coeficiente de regresión vale 0 ($e^{0}=1$), e implica que la variable independiente no produce ningún cambio en la razón de un éxito.
- Si la razón de razones es mayor de 1, esto significa que el coeficiente de regresión es mayor que 0 ($e^{x}> 0$ si $x>1$), e implica que la variable independiente produce un aumento en la razón del suceso.
- Si la razón de razones es menor de 1, esto significa que el coeficiente de regresión es más pequeño que 0 ($e^{x}<0$ si $x<1$), e implica que la variable independiente produce una disminución en la razón del suceso.

En el ejemplo, vemos que la probabilidad de abstenerse cuando se lee diarios es de 0,2509, la razón de este suceso (de no votar respecto a votar) es:

$$\frac{0,2509}{1-0,2509} = 0,335$$

Por otra parte, la probabilidad de abstenerse cuando no se lee es de 0,5245, la razón de este suceso es:

$$\frac{0,5245}{1-0,5245} = 1,103$$

Por tanto, cuando la probabilidad de abstenerse pasa de 0,2509 (cuando $x=0$) a 0,5245 (cuando $x=1$), su razón pasa de 0,335 a 1,103. Y la razón de las razones expresa este

aumento, la razón de no votar respecto a votar se ve aumentada 3,293 veces cuando se pasa de leer periódicos (*x=0*) a no leer periódicos (*x=1*). Así, mientras que la probabilidad de abstenerse se duplica (de 0,2509 a 0,5245) la razón de las razones se triplica (de 0,335 a 1,103).

La técnica de la regresión logística es una herramienta muy extendida en ciencias de la salud debido a que los parámetros en los que se basa tienen una interpretación en términos de riesgo. En la literatura de ciencias de la salud cuando el suceso se refiere a la aparición de una enfermedad, por ejemplo, a las variables independientes la razón de razones de las que es mayor de 1 se las identifica como factores de riesgo, y las variables independientes la razón de razones de las cuales es menor de 1 como factores de protección.

La ecuación de regresión logística genera para cada individuo valores de probabilidad pronosticada en el rango de valores entre 0 y 1. Raramente estos valores pronosticados son tan discriminantes que se distribuyen entre valores o bien 0 o bien 1, situación de máxima predictibilidad y de máxima discriminación que generaría un gráfico de la función con forma de escalón. El problema que se suscita es el de establecer un punto de corte óptimo para decidir si clasificar a los individuos en un grupo o en otro a partir de las probabilidades pronosticadas y así diferenciar al máximo a los individuos y obtener la mejor clasificación posible. En el caso de la variable dicotómica que hemos visto las probabilidades pronosticadas son 0,2509 para los que leen periódicos y 0,5245 para los que no leen. Por otra parte, sabemos que la probabilidad global de abstenerse es del 0,291 según se desprende de las frecuencias de esta variable. Con esta información nos podríamos plantear fijar un punto de corte que se encontrara entre estos dos valores, por ejemplo, podría ser el valor 0,291. Con este valor de corte, las personas con probabilidades pronosticadas mayores son clasificadas en el grupo correspondiente al valor 1 de la variable dependiente (los que no votaron) y las personas que tienen probabilidades pronosticadas menores o iguales son clasificadas en el grupo correspondiente al valor 0 de la variable dependiente (los que votaron).

Sobre esta cuestión volveremos más adelante, pero para determinar el punto de corte óptimo se puede optar por dos alternativas: generar múltiples tablas de clasificación modificando en cada caso el punto de corte hasta optimizar el porcentaje de casos correctamente clasificados, o, alternativamente se puede usar el procedimiento de Curvas COR (Característica de Operación del Receptor).

Veamos un segundo ejemplo de cálculo con la relación entre la posesión de coche y la clase social.

En forma aditiva:

Ejemplo COCHE y CSE

(y) *Var. Dependiente* — *(x)* *Var. Independiente cualitativa politómica*

Razón de probabilidades

de Sí coche a No coche global: $\frac{0,698}{0,302}=2,3$

COCHE Posesión de coche*CSE Categoría socioeconómica tabulación cruzada

			CSE Categoría socioeconómica			
			1 Categoría alta	2 Categoría media	3 Categoría baja	Total
COCHE Posesión de coche	0 No	Recuento	64	333	1036	1433
		% dentro de CSE	9,0%	21,3%	42,0%	30,2%
	1 Sí	Recuento	650	1234	1430	3314
		% dentro de CSE	91,0%	78,7%	58,0%	69,8%
Total		Recuento	714	1567	2466	4747
		% dentro de CSE	100,0%	100,0%	100,0%	100%

Categoría de referencia — *Categoría de interés*

- Si C. Baja $\frac{0,580}{0,420}=1,4$ $\log\left(\frac{\Pr(Y=1\mid X=0)}{\Pr(Y=0\mid X=0)}\right)=\log\left(\frac{P}{1-P}\right)=\log\left(\frac{0,580}{1-0,580}\right)=\log(1,38)=0,322\Rightarrow a=0,322$
- Si C. Media: $\frac{0,787}{0,213}=3,7$ $\log\left(\frac{\Pr(Y=1\mid X=1)}{\Pr(Y=0\mid X=1)}\right)=\log\left(\frac{P}{1-P}\right)=\log\left(\frac{0,787}{1-0,787}\right)=\log(3,695)=1,307$
- Si C. Alta: $\frac{0,910}{0,090}=10,1$ $\log\left(\frac{\Pr(Y=1\mid X=2)}{\Pr(Y=0\mid X=2)}\right)=\log\left(\frac{P}{1-P}\right)=\log\left(\frac{0,910}{1-0,910}\right)=\log(10,1)=2,314$

CSE(1) es cuando x pasa de 0 a 1: b = 1,307- 0,322 = 0,988

CSE(2) es cuando x pasa de 0 a 2: b = 2,314- 0,322 = 1,996

Coeficientes en **forma aditiva**

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1ª	CSE			334,974	2	,000	
	CSE(1)	,988	,074	178,039	1	,000	2,685
	CSE(2)	1,996	,137	211,554	1	,000	7,358
	Constante	,322	,041	62,408	1	,000	1,380

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

En forma exponencial:

Ejemplo COCHE y CSE

(y) *Var. Dependiente* — *(x)* *Var. Independiente cualitativa politómica*

COCHE Posesión de coche*CSE Categoría socioeconómica tabulación cruzada

			CSE Categoría socioeconómica			
			1 Categoría alta	2 Categoría media	3 Categoría baja	Total
COCHE Posesión de coche	0 No	Recuento	64	333	1036	1433
		% dentro de CSE	9,0%	21,3%	42,0%	30,2%
	1 Sí	Recuento	650	1234	1430	3314
		% dentro de CSE	91,0%	78,7%	58,0%	69,8%
Total		Recuento	714	1567	2466	4747
		% dentro de CSE	100,0%	100,0%	100,0%	100%

Categoría de referencia — *Categoría de interés*

Coeficientes en **forma exponencial**

- Si C. Baja: $exp(a)=e^{a}=e^{0,322}=1,380$
- Si C. Media: $exp(b)=e^{b}=e^{0,988}=2,685$
- Si C. Alta: $exp(b)=e^{b}=e^{1,996}=7,358$

Cuando la CSE varía de Baja a Media, la razón de "Sí coche" a "No coche" aumenta en **2,685** $\frac{0,787}{0,213}\Big/\frac{0,580}{0,420}=3,695/1,38=2,685$

Cuando la CSE varía de Baja a Alta, la razón de "Sí coche" a "No coche" aumenta en **7,358** $\frac{0,910}{0,090}\Big/\frac{0,580}{0,420}=10,1/1,38=7,358$

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1ª	CSE			334,974	2	,000	
	CSE(1)	,988	,074	178,039	1	,000	2,685
	CSE(2)	1,996	,137	211,554	1	,000	7,358
	Constante	,322	,041	62,408	1	,000	1,380

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

Con variables cualitativas un análisis de regresión logística genera resultados similares a un análisis log-lineal.

Comparación con un Log-lineal logit

Ejemplo COCHE y CSE

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1[a]	CSE			334,974	2	,000	
	CSE(1)	,988	,074	178,039	1	,000	2,685
	CSE(2)	1,996	,137	211,554	1	,000	7,358
	Constante	,322	,041	62,408	1	,000	1,380

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

COCHE Posesión de coche * CSE Categoría socioeconómica tabulación cruzada

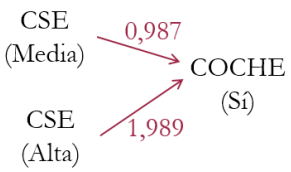
Categoría de referencia / *Categoría de interés*

			CSE Categoría socioeconómica			
			1 Categoría alta	2 Categoría media	3 Categoría baja	Total
COCHE Posesión de coche	0 No	Recuento	64	333	1036	1433
		% dentro de CSE	9,0%	21,3%	42,0%	30,2%
	1 Sí	Recuento	650	1234	1430	3314
		% dentro de CSE	91,0%	78,7%	58,0%	69,8%
Total		Recuento	714	1567	2466	4747
		% dentro de CSE	100,0%	100,0%	100,0%	100%

Estimaciones de los parámetros[c,d]

Parámetro		Estimación	Error típico	Z	Sig.	Intervalo de confianza al 95% Límite inferior	Límite superior
Constante	[CSE = 1]	4,167[a]					
	[CSE = 2]	5,810[a]					
	[CSE = 3]	6,944[a]					
[COTXE = 1]		,322	,041	7,898	,000	,242	,402
[COTXE = 2]		0[b]	.	.	.	.	.
[COTXE = 1] * [CSE = 1]		1,989	,137	14,542	,000	1,721	2,257
[COTXE = 1] * [CSE = 2]		,987	,074	13,337	,000	,842	1,132
[COTXE = 1] * [CSE = 3]		0[b]	.	.	.	.	.
[COTXE = 2] * [CSE = 1]		0[b]	.	.	.	.	.
[COTXE = 2] * [CSE = 2]		0[b]	.	.	.	.	.
[COTXE = 2] * [CSE = 3]		0[b]	.	.	.	.	.

a. Las constantes no son parámetros bajo el supuesto multinomial. Por tanto, no se calculan sus errores típicos.
b. Este parámetro se ha definido como cero ya que es redundante.
c. Modelo: Logit multinomial
d. Diseño: Constante + COTXE + COTXE * CSE



Consideremos ahora una variable independiente cuantitativa (EDUC, los años de estudios) para explicar el abstencionismo. La interpretación y la lectura de los coeficientes que se obtienen, tanto en un modelo aditivo como multiplicativo, se presenta a continuación.

Ejemplo VOTO y EDUC

(y) *Var. Dependiente* — *(x)* *Var. Independiente cuantitativa*

Coeficiente en forma aditiva

EDUC —(-0,212 / 0,809)→ VOTO (No)

Coeficiente en forma multiplicativa

- Cuando "EDUC" (los años de estudios) aumentan en una unidad el log de la razón del "VOTO" (no votar en relación a votar) se reduce en -0,212.
- Cuando "EDUC" (los años de estudios) aumentan en una unidad la razón del "VOTO" (no votar en relación a votar) cambia con un factor multiplicativo de 0,809. La probabilidad de "No votar" se reduce en un 19,1%

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1[a]	educ	-,212	,021	99,734	1	,000	,809
	Constante	1,772	,269	43,441	1	,000	5,880

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

- Cuando "EDUC" (los años de estudios) es 0 el log de la razón del "VOTO" (no votar en relación a votar) es 1,772.
- Cuando "EDUC" (los años de estudios) es 0 la razón del "VOTO" (no votar en relación a votar) es 5,880. Se multiplica per 5,88 la probabilidad de "No votar"

La interpretación del coeficiente de regresión como factor multiplicativo se presenta seguidamente:

Ejemplo VOTO y EDUC

EDUC -0,809 VOTO (No)

Años escolarización	Log de las razones predichas	Razones predichas	Prob.de No votar
0	**1,772**	**5,883**	0,855
1	1,560	4,759	0,826
2	1,348	3,850	0,794
3	1,136	3,114	0,757
4	0,924	2,519	0,716
5	0,712	2,038	0,671
6	0,500	1,649	0,622
7	0,288	1,334	0,572
8	0,076	1,079	0,519
9	-0,136	0,873	0,466
10	-0,348	0,706	0,414
11	-0,560	0,571	0,364
12	-0,772	0,462	0,316
13	-0,984	0,374	0,272
14	-1,196	0,302	0,232
15	-1,408	0,245	0,197
16	-1,620	0,198	0,165
17	-1,832	0,160	0,138
18	-2,044	0,130	0,115
19	-2,256	0,105	0,095
20	-2,468	0,085	0,078

$$\log\left(\frac{P(no\,votar)}{P(votar)}\right) = 1,772 - 0,212 \cdot educ$$

$$\frac{P(no\,votar)}{P(votar)} = 5,880 \times 0,809^{educ}$$

Si k>1: $e^{kb} = e^{k \cdot -0,212}$

$$P(no\,votar) = \frac{1}{1+e^{-(1,772-0,212 \cdot educ)}}$$

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1[a]	educ	-,212	,021	99,734	1	,000	,809
	Constante	1,772	,269	43,441	1	,000	5,880

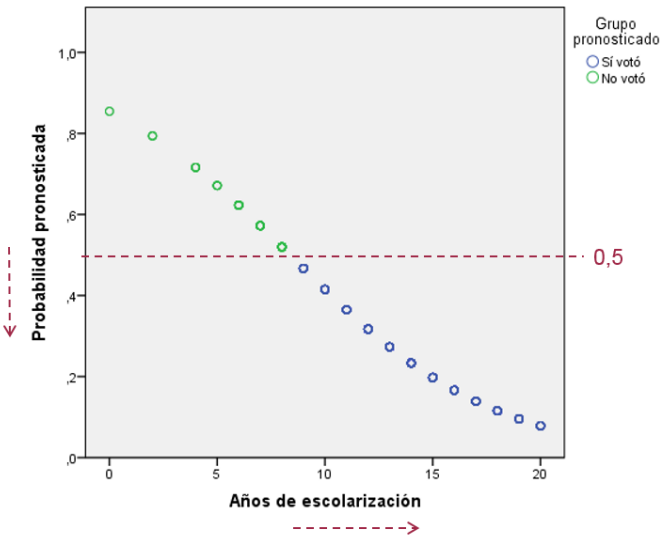
a. Variable(s) introducida(s) en el paso 1: educ.

1 año más de "EDUC" supone un cambio de las razones predichas en un factor multiplicativo constante de 0,809:

- De 0 a 1: 5,883 × 0,809 = 4,789
- De 1 a 2: 4,789 × 0,809 = 3,850
- De 2 a 3: 3,850 × 0,809 = 3,114
- ...

A medida que aumentan los años de estudios disminuye la probabilidad no votar, de abstenerse:

EDUC -0,212 VOTO (No)

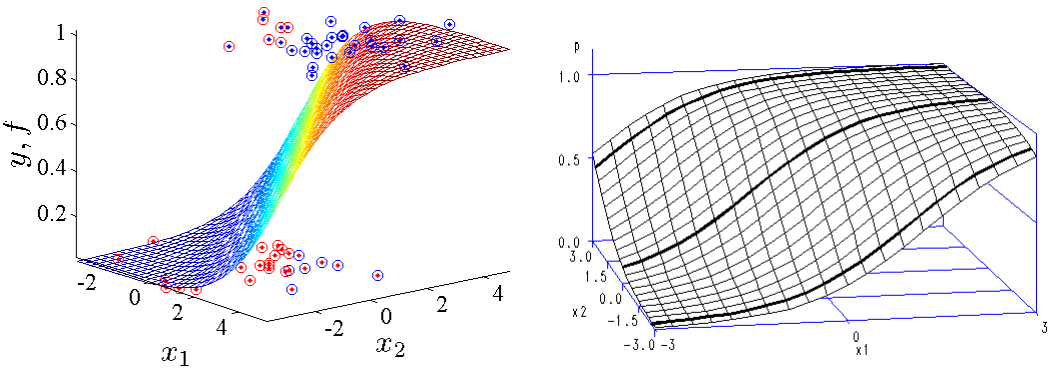


2. Análisis regresión logística binaria múltiple

Ahora consideraremos la regresión múltiple añadiendo diversas variables independientes. En este caso, con más de una variable independiente o covariable, la expresión de la ecuación de regresión logística es:

$$y = Pr(y=1|x) = \frac{1}{1+e^{-(a+b_1x_1+b_2x_2+\cdots+b_jx_j)}} = \frac{1}{1+e^{-\left(a+\sum_{j=1}^{p} b_jx_j\right)}} \qquad \textit{Ecuación 11}$$

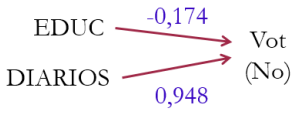
En este caso que y tome el valor 1 depende de un conjunto de p covariables $x_1,\ldots,x_p$ que inciden de forma diferente según un coeficiente b_j, siendo a la constante del modelo.



La cuestión que se suscita en un contexto multivariable es la determinación del mejor modelo explicativo, lo que significa evaluar la pertinencia de las diferentes variables independientes propuestas inicialmente en el modelo de regresión logística y escoger aquellas que mejor explican la variable dependiente. A tal efecto disponemos de varios métodos de selección de variables.

Retomamos el ejemplo anterior del voto considerando simultáneamente las variables DIARIOS y EDUC. La ecuación de regresión incluye dos coeficientes cuya lectura e interpretación se presenta a continuación.

Ejemplo VOTO y DIARIOS + EDUC



- Cuando "EDUC" aumenta en una unidad el log de la razón de "No votar" se reduce en -0,174, controlando o ajustando (si no cambia) "DIARIOS"

- Cuando "EDUC" aumenta en una unidad la razón de "No votar" cambia con un factor multiplicativo de 0,841, controlando o ajustando (si no cambia) "DIARIOS". La probabilidad de "No votar" se reduce en un 15,9%

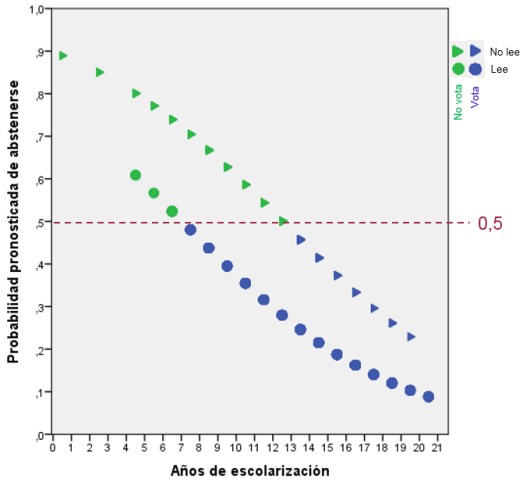
Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	EDUC	-,174	,026	45,575	1	,000	,841	,799	,884
	DIARIOS	,948	,194	23,761	1	,000	2,580	1,762	3,776
	Constante	1,136	,334	11,599	1	,001	3,115		

a. Variables especificadas en el paso 1: EDUC, DIARIOS.

- Cuando "DIARIOS" aumenta en una unidad el log de la razón de "No votar" aumenta en 0,948, controlando o ajustando (si no cambia) "EDUC"

- Cuando "DIARIOS" aumenta en una unidad la razón de "No votar" cambia con un factor multiplicativo de 2,580, controlando o ajustando (si no cambia) "EDUC". La probabilidad de "No votar" se multiplica por 2,58 (158% más)



2.1. Proceso de análisis

El proceso de análisis que se sigue tiene en cuenta los aspectos siguientes:

1) Selección de las variables del modelo.
- Se requiere la definición de un modelo de análisis que justifique un modelo de dependencia.
- Analizar las condiciones de aplicación.
- Constatar manifiesta relación de cada variable independiente con la variable dependiente.
- Análisis igualmente de las relaciones entre las variables independientes para establecer posibles situaciones de espureidad o colinealidad.
- Plantear la inclusión de interacciones.
- Seguir un proceso de selección de las mejores variables con el objetivo de ajustar el mejor modelo lo más parsimonioso.

2) Estimación de los coeficientes de les variables independientes
- Se aplica un método de estimación de máxima verosimilitud, un caso particular de Mínimos Cuadrados Ordinarios (MCO).
- Los coeficientes valoran la importancia de cada variable independiente, cuantitativa (con un único valor), cualitativa (estableciendo una categoría de referencia, *baseline*, y generando uno o más coeficientes) y las posibles interacciones.

3) Evaluación del modelo.
- Se trata de establecer la bondad de ajuste del modelo o capacidad explicativa: estadísticos de R^2 de Nagelkerke o R^2 de Cox y Snell, así como la prueba de Hosmer y Lemeshow.
- La eficacia predictiva o capacidad discriminatoria del modelo: surge del cruce de la variable dependiente observada por la pronosticada, la llamada tabla de clasificación o matriz de confusión. Se trata de mirar el porcentaje de casos bien clasificados. Para ello se establece un criterio de corte: de probabilidad de que la variable dependiente sea 1. Por defecto se considera el valor 0,5, pero puede ajustarse con un estudio específico. Para ello se dispone en particular del

denominado análisis de curvas COR (Característica de Operación del Receptor, en inglés ROC).
- Análisis de valores extremos mediante un análisis de los residuos, detectando casos influyentes, analizando medidas de influencia y la multicolinealidad.
- Interpretación de los resultados con el conjunto de la información, pero fundamentalmente dando cuenta de la tabla de coeficientes.

2.2. Condiciones de aplicación

La regresión logística no establece condiciones restrictivas de aplicación. En particular, no se imponen supuestos a la distribución de las variables.

Se establecen las siguientes condiciones:

1) El modelo debe estar correctamente especificado y ser relevante sustantivamente.
2) No se omiten variables independientes relevantes.
3) Las variables independientes se miden sin error.
4) Las observaciones son independientes entre sí.
5) Ausencia de colinealidad entre las variables independientes. Es una cuestión de grado. Correlaciones de 0,8 la implican pues incrementa los errores típicos: cuando los errores típicos sean superiores a 2 indica la existencia de multicolinealidad[5]. La colinealidad se puede detectar, pero no es fácil de resolver. Cuando es alta o no tolerable, hay que revisar el modelo y, por ejemplo, eliminar una o más de las variables colineales, cambiar la escala de medida o combinarlas en una medida única.
6) Linealidad de las variables cuantitativas.
7) Monotonicidad: cada independiente se relaciona de forma directa o indirecta.
8) En relación al tamaño de la muestra. Hosmer y Lemeshow recomiendan muestras mayores de 400 casos. De Maris (1992) sugiere 15 casos por variable. Peduzzi *et al.* (1996) sugieren que el número mínimo se calcula como:

$$n = \frac{10p}{P}$$

donde p es el número de covariables y P es la proporción de casos más pequeña en la población de presencia o ausencia. Es decir, por cada covariable contar 10 casos por cada evento de la VD con menor representación: $\frac{n}{10} \cdot P = p$. Si el valor es inferior a 100, Long (1997) sugiere incrementarlo a 100.
Esta regla se recomienda en casos de covariables cuantitativas o cualitativas con distribuciones relativamente equilibradas. Con variables cualitativas dicotómicas se consideran 10 casos por casilla de la tabla que cruza la variable dependiente binaria y la variable independiente.
9) El tanto por ciento de casos que corresponden al 0 o al 1 de la variable dependiente debe ser del 10% al menos.

[5] En SPSS no se proporcionan pruebas para determinarla, se puede recurrir a la regresión lineal: IBM SPSS (http://www-01.ibm.com/support/docview.wss?uid=swg21476696) recomienda realizar una regresión lineal y estudiar los indicadores de multicolinealidad. Si la variable es cualitativa se convierte en variable dummy (o indicador), ver http://www-01.ibm.com/support/docview.wss?uid=swg21476169.

2.3. Pruebas de significación

En el análisis de regresión logística se consideran las siguientes pruebas estadísticas de significación:

a) Evaluación del ajuste del modelo a través del cambio o del incremento del estadístico $-2log(L)$ (lejanía del modelo) donde L es la razón de verosimilitud que varía entre 0 y 1. Se determina la verosimilitud del modelo y se compara con la del modelo nulo (L_0). El estadístico L se distribuye según χ^2 al igual que el estadístico $-2log(L)$. Si el modelo incluye sólo la constante, los grados de libertad son el número de casos menos uno ($n-1$), si el modelo considera las variables independientes entonces los grados de libertad es igual al número de casos, menos el número de variables independientes menos uno ($n-k-1$). Si calculamos la diferencia entre los dos valores del estadístico de los dos modelos, el constante y el que incluye las variables independientes, obtenemos un estadístico que también se distribuye según un χ^2 con la diferencia de grados de libertad, es decir, el número de variables independientes del modelo k. Como L varía entre 0 y 1, el valor de $-2log(L)$ es cero en un modelo perfecto, en un modelo de dependencia que se ajusta a los datos, lo que significa que el valor de verosimilitud no difiere significativamente, de 1. Por lo tanto, implica plantear como hipótesis nula que $L=1$.

   Mediante $-2log(L)$ podemos evaluar el modelo de regresión logística, y en este caso que no se pueda rechazar la hipótesis nula equivale a decir que el modelo es significativo y, por tanto, que el coeficiente de regresión es significativamente diferente de cero. En conclusión, por tanto, para que ajuste el modelo se debe obtener una probabilidad superior o igual a 0,05.

b) Evaluación estadística de los coeficientes de regresión logística, si son significativamente distintos de 0, mediante el estadístico de Wald:

$$wald = \frac{b^2}{s_b^2} \qquad \textit{Ecuación 12}$$

   que sigue una distribución normal estándar. Los coeficientes significativos son los que tienen una probabilidad inferior a 0,05.

c) El cálculo del Pseudo R^2 que determina la bondad de ajuste, cuánto mejora un modelo en relación al modelo, expresado en porcentaje:

$$Pseudo R^2 = 1 - \frac{L_M^2}{L_0^2} \qquad \textit{Ecuación 13}$$

   Se trata de medidas que evalúan el incremento de la verosimilitud del modelo: el cambio del estadístico $-2log(L)$ o bien de L, la razón de verosimilitud que varía entre 0 y 1. Si $\Lambda = -2log(L)$ entonces $L = \exp(-\Lambda/2)$ y los pseudo R^2 se calculan de la forma siguiente en el caso del de Cox y Snell y del de Nagelkerke:

Pseudo R^2 de Cox y Snell:

$$R^2_{CS} = 1 - \left(\frac{L_{constante}}{L_{modelo}} \right)^{\frac{2}{n}} = 1 - \exp\left(\frac{\Lambda_{modelo} - \Lambda_{constante}}{n} \right) \qquad \textit{Ecuación 14}$$

Estadístico que varía entre 0 y 1, $0 \leq R^2 \leq 1$, y donde $L(a)$ es el modelo de la constante mientras que $L(a,b_1, b_2,\ldots,b_j)$ es el modelo completo considerado.

Pseudo R^2 de Nagelkerke:

$$R^2_N = \frac{1 - \left(\frac{L_{constante}}{L_{modelo}} \right)^{\frac{2}{n}}}{1 - (L_{constante})^{\frac{2}{n}}} = \frac{1 - \exp\left(\frac{\Lambda_{modelo} - \Lambda_{constante}}{n} \right)}{1 - \exp\left(\frac{-\Lambda_{constante}}{n} \right)} \qquad \textit{Ecuación 15}$$

Con $1-(L_{constante})^{\frac{2}{n}}$ que corresponde al $R^2_{\max}$. Es un estadístico que varía también entre 0 y 1, con valores algo superiores en relación anterior. Tanto éste como el anterior son indicadores de la variabilidad explicada que siempre proporcionan valores muy bajos en relación a la medida homóloga en regresión lineal clásica. Es habitual encontrar resultados de 0,2 y 0,3, mientras que un pseudo R^2 de 0,6 es poco habitual.

d) Evaluación del estadístico de bondad de ajuste z^2 :

$$z^2 = \sum_{i=1}^{n} \frac{(p_i - \hat{p}_i)^2}{\hat{p}_i(1-\hat{p}_i)} = \sum_{i=1}^{n} \frac{R_i^2}{\hat{p}_i(1-\hat{p}_i)} \qquad \textit{Ecuación 16}$$

donde R_i es el residuo entre la probabilidad observada y la probabilidad estimada del i-ésimo caso. Sigue una distribución de chi-cuadrado. Cuando el modelo es significativo la probabilidad asociada es mayor o igual a 0,05.

e) Prueba de Hosmer y Lemeshow de bondad de ajuste del modelo. Para corroborar si el modelo se ajusta se utiliza este contraste de distribución. Para ello se calcula la probabilidad pronosticada del suceso para todos los individuos de la muestra y se calcula la diferencia con los valores observados. La prueba consiste en dividir el recorrido de la probabilidad en deciles y se comparan las distribuciones de frecuencias esperada y observada mediante un contraste de chi-cuadrado con 8 grados de libertad.

$$\chi^2_{HL} = \sum_{g=1}^{G} \frac{(O_g - E_g)^2}{E_g(1 - E_g/n_g)} \qquad \textit{Ecuación 17}$$

La hipótesis nula establece que no hay diferencias entre los valores observados y pronosticados: el modelo ajusta. Por tanto, si el ajuste es bueno, se esperar un valor alto de probabilidad, superior o igual a 0,05. Se trata de una prueba adecuada para muestras pequeñas y covariables continuas.

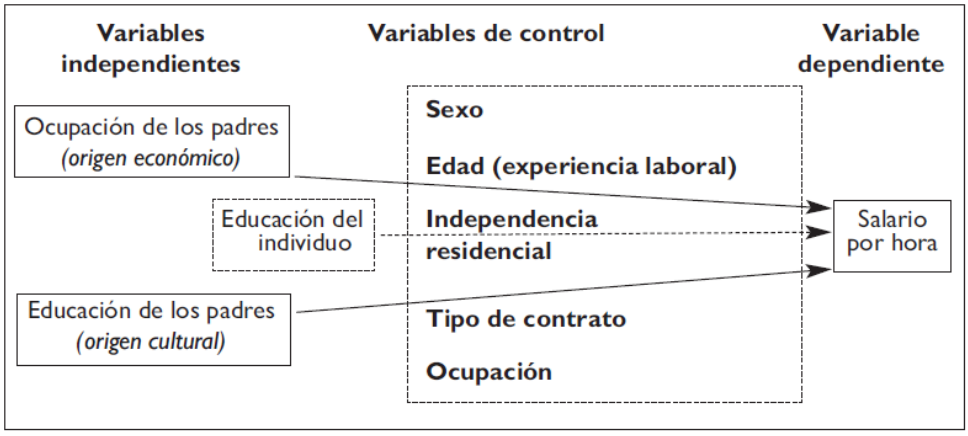
2.4. Inserción laboral de graduados universitarios: ejemplo de aplicación

En el marco de una investigación financiada por el Ministerio de Ciencia e Innovación, concretamente por el Plan Nacional de Investigación Científica, Desarrollo e Investigación Tecnológica (CSO 2010-19271), con el título de "Itinerarios universitarios, equidad y movilidad ocupacional (ITUNEQMO)", se desarrolló un análisis del papel del origen social en la inserción laboral de las personas graduadas entre los años 1999 y 2002 en España Los resultados fueron publicados en la Revista de Educación (Fachelli, Torrents y Navarro-Cendejas, 2014) y los utilizaremos como ejemplo de aplicación de la regresión logística binomial múltiple.

Los autores plantean que en varios estudios realizados en el marco del proyecto han observado que el origen social tiene una baja influencia en la inserción laboral de las personas graduadas en Cataluña y se proponen probar los resultados a nivel estatal. Para ello utilizan los datos disponibles de la Encuesta de Condiciones de Vida de 2005, concretamente los de un módulo realizado a personas de entre 26 y 65 años en el que se indaga sobre la ocupación y educación de sus padres.[6]

Se parte de un modelo de análisis tradicional donde se analiza el impacto del origen social en términos económicos y culturales (variables independientes) sobre la inserción laboral de los universitarios teniendo en cuenta las características individuales como el sexo, la experiencia laboral, el nivel educativo, el tipo de contrato y la ocupación (variables de control).

Gráfico III.10.8. Modelo de análisis de de la Inserción laboral de graduados universitarios



Fuente: Fachelli, Torrents y Navarro-Cendejas (2014: 128)

La variable dependiente, dicotómica, es el salario bruto por hora que sintetiza varios aspectos de la inserción laboral. Los salarios altos son los iguales o superiores a la media de los salarios de cada uno de los grupos analizados: graduados y trabajadores de toda la población. El criterio de corte utilizado para fijar el límite entre unos y otros es la media recortada, es decir, la media de los salarios eliminando el 5% más alto y el 5% más bajo. De esta manera, se logra un valor medio más estable no sujeto a la presencia de casos extremos. Así, los graduados universitarios que ganan 8,89 euros por hora o

[6] El artículo puede ser consultado aquí: https://ddd.uab.cat/pub/artpub/2014/118532/revedu_a2014m4-6n364p119iSPA.pdf . Este análisis se puede reproducir tal y como se comenta en el último apartado del capítulo.

más están dentro de la categoría Salarios altos y el resto en la categoría Salarios bajos. En el caso de todos los trabajadores, el salario medio por hora es de 8,69 euros y se siguió el mismo procedimiento para la clasificación.

El modelo de análisis presentado se aplica de dos maneras concretas:

a) A los graduados universitarios entre 1999 y 2002.
b) A toda la población que está trabajando en el momento de la encuesta.

Esto hace que el modelo inicial se vea modificado de la siguiente manera. Para el primer modelo se agrega la variable sobre si la persona se ha independizado y diferenciar si el graduado vive o no con sus padres. Para el segundo modelo se agrega la educación del trabajador/a ya que se cuenta con toda la muestra.

De esta manera se testean dos hipótesis que a su vez contienen dos subhipótesis:

Hipótesis 1. El origen familiar de los graduados entre 1999 y 2002 influye en su inserción laboral.

- Subhipótesis 1a: El origen económico de los graduados entre 1999 y 2002 (máximo nivel ocupacional de los padres) influye en su inserción laboral.
- Subhipótesis 1b: El origen cultural de los graduados entre 1999 y 2002 (máximo nivel educativo de los padres) influye en su inserción laboral.

Hipótesis 2. El origen familiar de la población trabajadora influye en su inserción laboral.

- Subhipótesis 2a: El origen económico de la población trabajadora (máximo nivel ocupacional de los padres) influye en su inserción laboral.
- Subhipótesis 2b: El origen cultural de la población trabajadora (máximo nivel educativo de los padres) influye en su inserción laboral.

A su vez para analizar el efecto del origen económico y cultural en forma separada se utilizó para contrastar cada hipótesis un modelo básico, que contenía dichas variables y un modelo ampliado que incluía las de control.

El ejemplo de aplicación que utilizaremos a continuación se corresponde con el testeo del modelo ampliado de la Hipótesis 1.

De esta manera se presentan los resultados de la regresión logística correspondiente a la Hipótesis 1: el origen familiar de los graduados entre 1999 y 2002 influye en su inserción laboral.

Junto con comentar la tabla con los resultados (Tabla III.10.1), hacemos sugerencias que orientan la presentación y realización de la lectura de los resultados de la regresión.

1. La primera recomendación es poner al pie de la tabla que se va a presentar en la publicación de los resultados el número de casos utilizados, el punto de corte utilizado, los datos del resumen del modelo y, si corresponde, la prueba Hosmer y Lemenshow.

Por tanto, en el ejemplo que estamos analizando la redacción sería: la bondad de ajuste del modelo utilizado para analizar los graduados universitarios (417 casos), presenta

un pseudo R^2 de Nagelkerke del 19,5% y la prueba de Hosmer y Lemeshow también muestra un buen ajuste pues su significación es mayor o igual a 0,05 (Rodríguez y Gutiérrez, 2007).[7]

2. En segundo lugar, analizar los resultados de las variables independientes que han intervenido en la ecuación. Así, se interpreta la significación del coeficiente de regresión Beta de las variables más importantes de nuestro modelo.

En este caso, el modelo ampliado muestra que el origen social no es significativo y que del resto de las variables podemos rescatar dos que tienen un rol activo en el modelo: el sexo de los graduados y su propia ocupación.

Tabla III.10.1. Resultados de la regresión logística de la Inserción laboral de graduados universitarios. Modelo ampliado

Variable dependiente: **Salario por hora**, Bajo vs Alto*	B	Error típico	Wald	Grados de libertad	Sig.	Exp(B)
Ocupacion de los padres			3,011	2	0,222	
Cualificados	−0,239	0,275	0,753	1	0,385	0,787
No cualificados	0,597	0,456	1,716	1	0,190	1,817
No manual*	–					
Estudios de los padres			1,206	2	0,547	
Primario	0,227	0,287	0,628	1	0,428	1,255
Secundario	0,334	0,309	1,167	1	0,280	1,397
Universitario*	–					
Sexo (Mujer)	−0,532	0,225	5,605	1	0,018	0,587
Varón*	–					
Emancipado (No)	−0,477	0,249	3,669	1	0,055	0,621
Sí*	–					
Contrato(Indefinido)	0,202	0,227	0,793	1	0,373	1,224
Temporal*	–					
AñosT	0,018	0,022	0,657	1	0,418	1,018
Ocupación entrevistado/a			43,805	2	0,000	
Cualificados	−1,506	0,236	40,587	1	0,000	0,222
No cualificados	−2,158	0,841	6,590	1	0,010	0,116
No manual*	–					
Constante	0,498	0,356	1,952	1	0,162	1,645
Número de casos	417					
-2 log de verosimilitud	507,04					
R^2 de Nagelkerke	0,195					
Sig. Hosmer-Lemeshow	0,569					
Punto de corte	0,5					
% de casos bien clasificados	66,9					

* Categoría de referencia
Fuente: Fachelli, Torrents y Navarro-Cendejas (2014)

[7] La expresión del pseudo R^2 puede hacerse en porcentaje o como una proporción.

3. Analizar los coeficientes en el modelo aditivo (columna B) que podemos denominar como indicadores de la jerarquía e intensidad de las variables que son significativas. En ese sentido se puede observar el valor del coeficiente y ordenar la lectura de variables en función de su importancia. En el caso de una variable politómica tomar en consideración el valor más alto de alguna de sus categorías.

En el caso que nos ocupa diremos que, de entre los elementos que resultan significativos la ocupación tiene más relevancia que el sexo.

4. Leer el signo de Beta junto al exponencial de Beta (columna Exp(B)) de los coeficientes con el fin de interpretar lo que podemos identificar como el impacto de la categoría en cuestión de la variable independiente, en relación a la categoría de referencia, sobre la variable dependiente (salarios altos en relación a salarios bajos).

En este sentido, ser mujer, baja en un 59% las probabilidades de tener un salario alto con respecto a ser varón. En el caso de la ocupación, ser trabajador no cualificado y agrícola reduce el 89% las probabilidades de tener salario alto y ser trabajador cualificado las reduce en casi un 80%, al compararlos con los trabajadores no manuales. Si bien es conveniente usar las categorías de referencia (de la independiente y de la dependiente) no conviene agobiar la presentación del resultado mareando al lector. Con una aclaración al comienzo de qué categorías de referencia se están utilizando en cada caso es suficiente.

5. Dependiendo del objetivo del estudio puede resultar de mucho interés el detenerse en la lectura de las variables que no son significativas que pueden resultar ser tan interesantes como las que son significativas.

En este caso, el hecho de que las variables sobre el origen social arrojen un resultado no significativo es muy importante. No sólo porque refuta la hipótesis planteada, sino que, como se analiza en la introducción del artículo, se aportan nuevas evidencias de un comportamiento determinado a nivel español, de hallazgos previamente observados a nivel autonómico en el caso de los graduados catalanes en el año 2008 (Fachelli y Planas, 2014; Fachelli, Planas y Navarro, 2012, Fachelli y Planas, 2011, Planas y Fachelli, 2010). Resultados que se siguieron constatando en la cohorte de graduados catalanes en 2011 (Fachelli y Navarro-Cendejas, 2015) y en graduados españoles durante la democracia, donde se analizaron 6 cohortes distribuidas entre los años 1975 y 1998 (Torrents y Fachelli, 2015).

Finalmente, el resto de las hipótesis estudiadas pueden consultarse en el artículo, aquí presentamos directamente los resultados del estudio sintetizándolos en el siguiente cuadro (Tabla III.10.2).

Tabla III.10.2. Conclusiones de las regresiones logísticas

ORIGEN FAMILIAR	GRADUADOS ENTRE 1999 Y 2002 HIPÓTESIS 1	TODOS LOS ENTREVISTADOS HIPÓTESIS 2
Económico	No influye	Influye
Cultural/Formativo	No influye	Influye
R^2 de Nagelkerke	19,5%	38,5%

Variable dependiente: Salario por hora
Fuente: Elaboración propia sobre la base de la ECV-INE, 2005.

Fuente: Fachelli, Torrents y Navarro-Cendejas (2014: 137)

El cuadro presenta los resultados de los modelos que han sido testeados. Así la hipótesis 1, que hace referencia al grupo de personas graduadas entre 1999 y 2002, queda refutada, pues constatamos que el origen social de los universitarios no tiene influencia en su inserción laboral. En este caso, la universidad diluye las diferencias entre orígenes sociales, tanto económicos como culturales. Solo la variable sexo y la propia ocupación son significativas. La capacidad explicativa del modelo completo es del 19,5%.

Por su parte, la Hipótesis 2 queda corroborada con una capacidad explicativa del 38,5%, pues se observa que en la población asalariada el origen social sí tiene influencia en su inserción laboral.

2.5. Ejemplos de aplicación en la literatura

El análisis de regresión logística se ha convertido en una de las técnicas más utilizadas en la actividad científica en general, y de las ciencias sociales y la sociología en particular. En la bibliografía se pueden encontrar diversos trabajos de investigación donde se ha aplicado la técnica. Referenciamos los textos de Escribà (2006), García, Alvarado y Jiménez (2000), Kelley (1990), Luque, Ferrer y Capdevila (2005), Miguélez et al. (2011), Miguélez y López-Roldán (2014), Reyneri (2006), Rodríguez-Ayán (2005), Stanek (2011), Verge y Tormos (2012).

3. Análisis regresión logística con SPSS

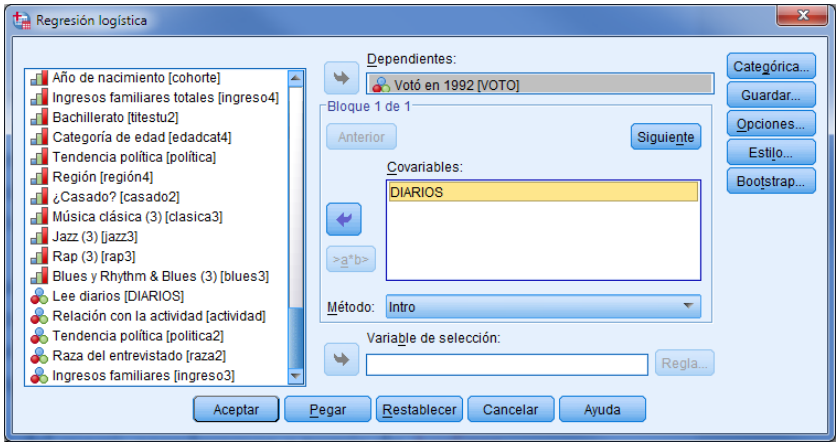
Reproduciremos el ejemplo presentado sobre el comportamiento abstencionista con los datos de la matriz GSS1993.sav (*General Social Survey* de EE.UU., del año 1993). Todos los resultados de SPSS que se presentan a partir de ahora son el resultado de ejecutar las instrucciones que se pueden encontrar en el archivo de sintaxis ARL-Voto.sps que se adjunta en la página web considerando diversos modelos.

3.1. Regresión binaria simple

3.1.1. *Una variable independiente dicotómica*

En primer lugar, consideraremos la variable VOTO para probar la hipótesis de si el abstencionismo en función de la lectura de los diarios (sí lee o no lee periódicos, variable independiente DIARIOS) como indicador del nivel cultural, en base a un modelo de regresión logística binaria simple.

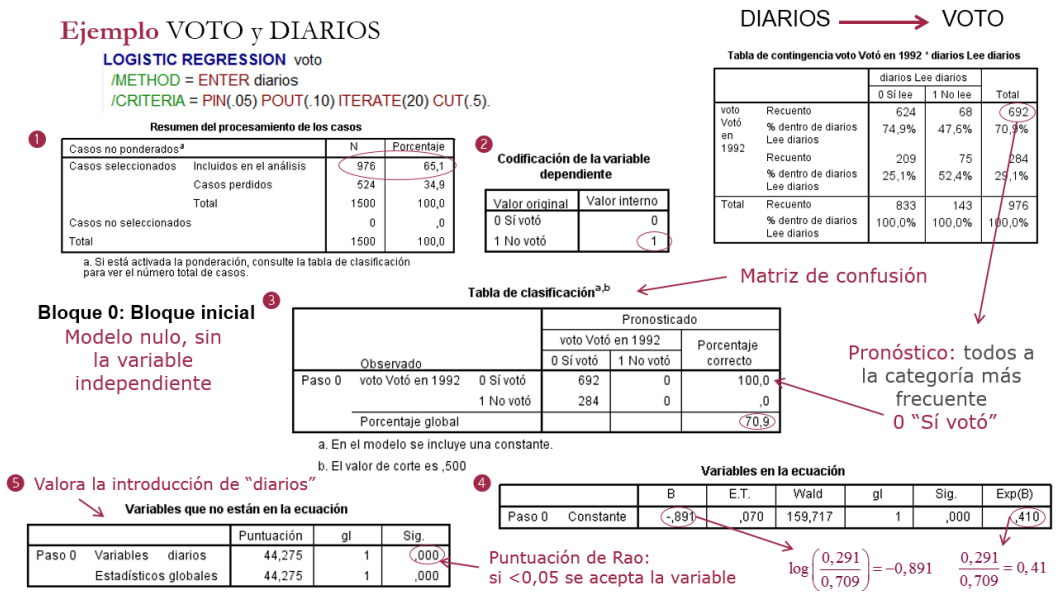
Para realizar el análisis con SPSS a través del menú, accederemos a través de Analizar / Regresión / Logística binaria, que corresponde al comando **LOGISTIC REGRESSION**. En el cuadro de diálogo principal introduciremos la variable dependiente, que debe ser dicotómica, en nuestro caso VOTO, y la variable independiente DIARIOS en el recuadro de Covariables:



Ejecutaremos directamente el procedimiento para ver los resultados que se obtienen con las especificaciones por defecto. Un análisis de regresión logística binaria simple nos proporciona los resultados que se presentan a continuación.

El procedimiento asigna el valor 0 a los casos que presentan el menor valor de la variable dependiente (ya sea una variable de tipo numérico o de tipo cadena) y el valor 1 a los casos con mayor valor. Habrá que tener en cuenta el criterio de codificación de cara a la interpretación de los resultados del modelo, la categoría de referencia será la codificada con valor 1.

Los resultados se inician en un Bloque 0 o Bloque inicial. El procedimiento de estimación del modelo de regresión logística se realiza a través del método de máxima verosimilitud e implica realizar un proceso iterativo en el que se va valorando la mejora en el ajuste en relación al modelo nulo, es decir, el modelo que considera solamente la constante, sin la variable independiente.



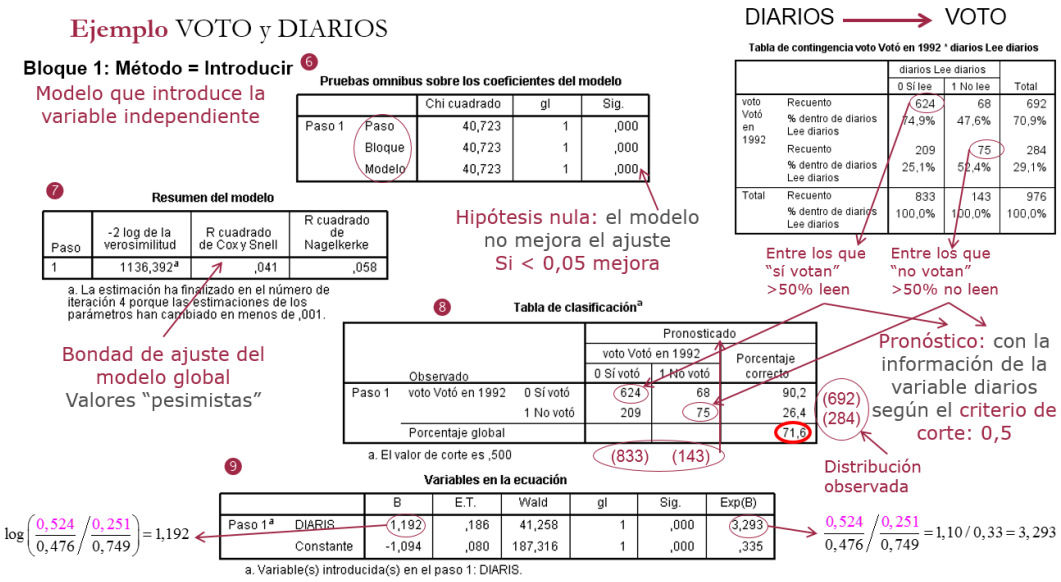
En esta etapa inicial se presenta una primera tabla de clasificación (o matriz de confusión o matriz de clasificación correcta) con el cruce de la variable dependiente observada con los pronósticos bajo el modelo nulo. Sin tener información de las variables independientes y, simplemente con el conocimiento de la distribución observada entre votantes y abstencionistas, el resultado pronosticado es la clasificación de todos los casos en la categoría más frecuente de la variable dependiente observada (en este caso la que corresponde al valor 0, sí votó). Los datos de la matriz de confusión son las frecuencias de la variable dependiente (sin los valores perdidos). En consecuencia, el porcentaje de casos correctamente clasificados coincide con el porcentaje de casos que pertenecen a la categoría más numerosa (70,9%).

La etapa o paso 0 con este modelo nulo proporciona la estimación del término constante (B=−0,891), los estadísticos asociados y la significación (0,000).

A continuación, se ofrece un contraste de hipótesis para valorar cada una de las variables independientes del modelo. En este caso sólo se valora la variable DIARIOS, que aún no se ha introducido en el modelo, pero si se introdujera, se formula la hipótesis nula según la cual el efecto de la variable es nulo. Siempre que el nivel de significación sea inferior a 0,05 se puede rechazar la hipótesis nula y aceptar que la variable independiente contribuye significativamente a mejorar el ajuste del modelo y, por tanto, a explicar el comportamiento de la variable dependiente. La tabla será de especial interés cuando se consideren diversas variables independientes y se utilice el método de introducción por pasos.

En este caso vemos como la introducción de la variable DIARIOS en el modelo mejoraría significativamente el ajuste (Sig. = 0,000).

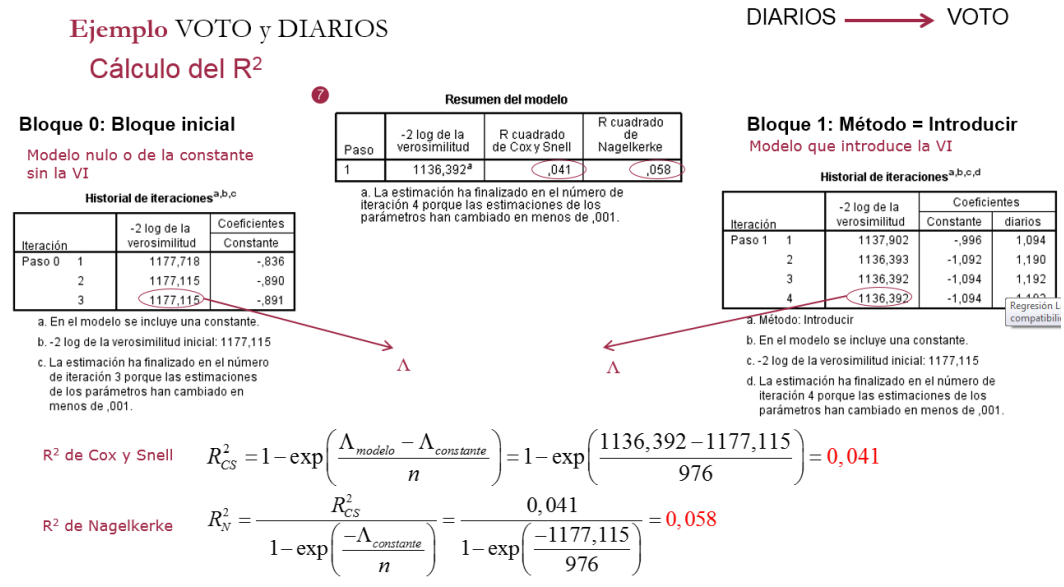
El paso siguiente consiste precisamente en introducir en el modelo la variable independiente, o las diversas variables en el caso de que fuera una regresión logística múltiple (Bloque 1: Método = Introducir).



En primer lugar, se contrasta la hipótesis nula según la cual el modelo no mejora el ajuste con la inclusión de la variable independiente. En este caso se rechaza esta hipótesis y se concluye que la introducción de la variable DIARIOS mejora significativamente el ajuste y la capacidad predictiva del modelo (todas las pruebas dan el mismo resultado dado que hay una única variable independiente).

Con el modelo global considerado se determina su bondad de ajuste. Los estadísticos de pseudo R^2 suelen presentar valores muy bajos por el hecho de tratarse de variables cualitativas, a pesar de que el modelo estimado sea adecuado. Por lo que hay que tomarse esta información como orientativa y teniendo presente que valores de 0,6 son elevados y de 0,2 y 0,3 son suficientes para un buen nivel de ajuste.

A continuación, se detallan los cálculos del ejemplo.



$$R^2_{CS} = 1 - \exp\left(\frac{\Lambda_{modelo} - \Lambda_{constante}}{n}\right) = 1 - \exp\left(\frac{1136{,}392 - 1177{,}115}{976}\right) = 0{,}041$$

$$R^2_{N} = \frac{R^2_{CS}}{1 - \exp\left(\frac{-\Lambda_{constante}}{n}\right)} = \frac{0{,}041}{1 - \exp\left(\frac{-1177{,}115}{976}\right)} = 0{,}058$$

Con el nuevo modelo de regresión logística se extrae la correspondiente tabla de clasificación que resulta del conocimiento de la variable independiente. En filas tenemos los valores de la variable VOTO observados en la muestra (No votó, un total de 692 casos, y Sí votó, un total de 284 casos). En columnas aparecen los valores pronosticados (No votó, un total de 833 casos, y Sí votó, un total de 143 casos). En este caso se ha procedido a realizar la asignación de casos pronosticados del voto utilizando el conocimiento de la distribución según la variable independiente de lectura de periódicos. Para asignar un caso a una categoría u otra del valor pronosticado se utiliza como criterio de corte por defecto el valor de probabilidad o la proporción 0,5 (cuando el porcentaje por fila de la tabla de contingencia sea superior al 50%). Como la proporción de casos, entre los que votan, que leen es superior a 0,5 (en este caso 624/692 = 0,902, según la tabla de contingencia), entonces hay que prever que es más probable que estos sean los que voten y se les asigna el valor 0 (Sí votó) en la variable pronosticada. Lo mismo se hace con los que no votan; es decir, como entre los que no votan predomina leer (209/284 = 0,736), es más probable que estos sean los no votantes y se les asigna el valor 1 (No votó) en la variable pronosticada. Este criterio de 0,5 es adecuado cuando el número de variables es alto y los grupos pronosticados son aproximadamente del mismo tamaño. Pero este valor puede ser estudiado con mayor detenimiento y variarlo.

Una vez realizada esta asignación obtenemos el cruce entre los valores observados y pronosticados. La diagonal principal contiene los casos correctamente clasificados por el modelo (699, es decir, 624 más 75), que son los que coinciden en ambas variables, y representan el 71,6% del total de casos, un porcentaje superior que en el caso del modelo nulo anterior. Este porcentaje es un indicador de validez del modelo ya que nos muestra su capacidad predictiva, su capacidad para clasificar correctamente los casos. Para obtener una máxima capacidad predictiva, del 100%, deberíamos tener una distribución extrema de los datos observados, en el que el 100% de los que votan son los que leen, y el 100% de los que no votan son los que no leen. Es la situación que corresponde al gráfico (d) de curvas logísticas que vimos anteriormente.

Finalmente, aparece la tabla con las estimaciones de los coeficientes del modelo. Como vimos el valor 1,192 nos indica que cuando pasamos de leer a no leer sube el abstencionismo, un valor, pues, positivo y de una magnitud que es el resultado de calcular el logaritmo neperiano del cociente de razones como analizamos.

3.1.2. *Una variable independiente cuantitativa*

Podemos reproducir otro análisis de regresión simple, pero con la variable cuantitativa EDUC y también con los resultados por defecto del procedimiento.

El modelo ajusta con un R^2 de Nagelkerke de 0,108.

Resumen del modelo

Escalón	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	1625,527[a]	,075	,108

a. La estimación ha terminado en el número de iteración 4 porque las estimaciones de parámetro han cambiado en menos de ,001.

Se obtienen los coeficientes de la ecuación:

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 1[a]	EDUC	-,212	,021	99,734	1	,000	,809
	Constante	1,772	,269	43,441	1	,000	5,880

a. Variables especificadas en el paso 1: EDUC.

Donde se evidencia la relación negativa entre educación y abstencionismo como vimos anteriormente. Cuando EDUC (los años de estudios) es 0 el logaritmo neperiano de la razón del VOTO (no votar en relación a votar) es 1,772. En términos multiplicativos cuando EDUC (los años de estudios) es 0 la razón del VOTO (no votar en relación a votar) es 5,880. Se multiplica per 5,88 la probabilidad de abstenerse.

3.1.3. *Una variable independiente politómica*

Una situación particular se produce cuando tratamos con una variable independiente cualitativa politómica que es necesario tener en consideración en el análisis.

Cuando se consideran variables independientes cuantitativas o independientes cualitativas dicotómicas, codificadas con 0 y 1, se pueden utilizar directamente y el análisis implica considerar un solo parámetro o coeficiente en la ecuación. Si son politómicas necesario un tratamiento particular en el análisis donde se aplicar un sistema de codificaciones determinado. Cuando una variable cualitativa está codificada con valores 0 y 1, o simplemente tiene dos valores válidos, se dice que es una variable indicador y la variable se puede introducir directamente en el análisis sin más. Pero si la variable presenta más de dos categorías es necesario definir la variable como categórica e indicar el tratamiento que deberá recibir.

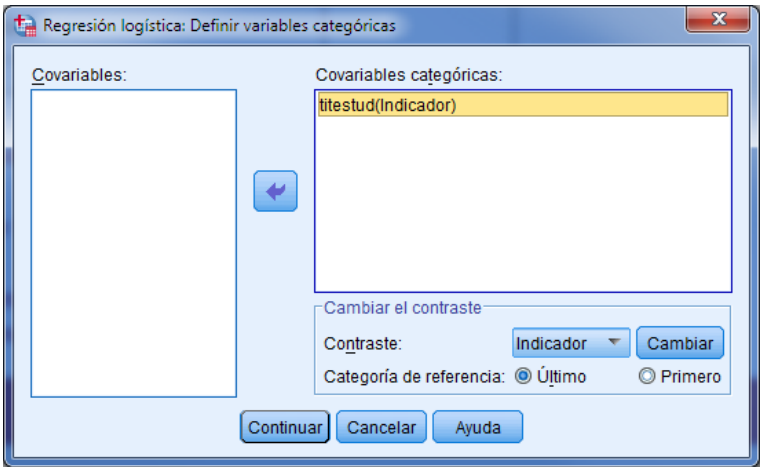
La definición de las variables categóricas implica especificar un método de contraste o esquema de codificación que recibirán las categorías de la variable durante el análisis. El procedimiento genera tantos contrastes como categorías tiene la variable menos uno. Cada contraste estima un coeficiente de regresión para cada una de las categorías de la variable excepto para una de ellas, la que se considera como categoría de referencia. Los contrastes generan una matriz de códigos en la que cada categoría adquiere un significado particular según las características del contraste.

En SPSS encontramos diferentes propuestas de contrastes. Pasamos a comentarlos seguidamente.

- Indicador. Cada categoría se compara con una categoría de referencia (que no es evaluada), y puede ser la primera o la última. La categoría de referencia se representa en la matriz de contraste como una fila de ceros. Cada categoría (con código 1, presencia) se compara con el resto (con código 0, ausencia) en relación a la categoría de referencia.
- Simple. Cada categoría de la variable predictora (excepto la misma categoría de referencia) se compara con la categoría de referencia. Esta puede ser la primera o la última y se identifica con unos.

- Diferencia. Cada categoría de la variable predictora, excepto la primera categoría, se compara con el efecto promedio de las categorías anteriores. También se conoce como contrastes de Helmert inversos.
- Helmert. Cada categoría de la variable predictora, excepto la última categoría, se compara con el efecto promedio de las categorías subsiguientes.
- Repetido. La primera categoría de la variable predictora se compara con el promedio de las restantes, la primera y la segunda con el promedio de las restantes, y así sucesivamente.
- Polinómica. Contrastes polinómicos ortogonales. Los códigos se asignan de forma que el primer contraste define una tendencia lineal, el segundo cuadrática, el tercero cúbica, etc. Los contrastes polinómicos sólo están disponibles para variables numéricas con valores igualmente espaciados.
- Desviación. Cada categoría de la variable predictora, excepto la categoría de referencia (la primera o la última), se compara con el efecto total, el promedio del resto de categorías, media total no ponderada.

El tipo contraste se escoge a través del botón Categórica que nos muestra el cuadro de diálogo principal que aparece seguidamente, destinado a definir las variables categóricas con el método de contraste. Para ilustrar este tipo de análisis consideraremos el caso de la relación simple entre el absentismo (variable VOTO) y la variable independiente cualitativa politómica referida al nivel educativo (variable titestud).



Se considerará como categoría de referencia el último valor de la variable (en este caso corresponde a la situación laboral Licenciado, valor 4 de la variable). Se presentan seguidamente los principales resultados.

En primer lugar, consideramos los datos de la tabla de contingencia que relacionan ambas variables. La relación es destacable y pone de manifiesto la relación inversa entre educación y abstencionismo: cuanto mayor es el nivel de estudios menor es el porcentaje de personas que no votan.

```
LOGISTIC REGRESSION voto
 /METHOD = ENTER titestud
 /CONTRAST (titestud)=Indicator
 /PRINT=GOODFIT
 /CRITERIA = PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .
```

Categórica: indicador

titestud → voto

Tipo de contraste

titestud Título escolar

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	0 Elemental	279	18,6	18,6	18,6
	1 Medio	780	52,0	52,1	70,8
	2 Bachillerato	90	6,0	6,0	76,8
	3 Diploma universitario	234	15,6	15,6	92,4
	4 Licenciado	113	7,5	7,6	100,0
	Total	1496	99,7	100,0	
Perdidos	8 No sabe	2	,1		
	9 No contesta	2	,1		
	Total	4	,3		
Total		1500	100,0		

Tabla de contingencia voto Votó en 1992 * titestud Título escolar

			titestud Título escolar					
			0 Elemental	1 Medio	2 Bachillerato	3 Diploma universitario	4 Licenciado	Total
voto Votó en 1992	0 Sí votó	Recuento	136	523	71	201	100	1031
		% dentro de titestud Título escolar	51,1%	68,6%	81,6%	88,9%	91,7%	71,1%
	1 No votó	Recuento	130	239	16	25	9	419
		% dentro de titestud Título escolar	48,9%	31,4%	18,4%	11,1%	8,3%	28,9%
Total		Recuento	266	762	87	226	109	1450
		% dentro de titestud Título escolar	100,0%	100%	100,0%	100,0%	100,0%	100,0%

V de Cramer: 0,283 (sig.=0,000)

Aplicando un contraste de tipo indicador, el que se aplica por defecto y es habitual, se generan estos resultados:

Categórica: indicador

titestud → voto

1 Resumen del procesamiento de los casos

		N	Porcentaje
Casos no ponderados[a]			
Casos seleccionados	Incluidos en el análisis	1450	96,7
	Casos perdidos	50	3,3
	Total	1500	100,0
Casos no seleccionados		0	,0
Total		1500	100,0

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

2 Codificación de la variable dependiente

Valor original	Valor interno
0 Sí votó	0
1 No votó	1

Codificación variable independiente
4 variables indicador (1),(2),(3),(4)
con valor 1 y el resto 0

Codificaciones de variables categóricas

		Frecuencia	Codificación de parámetros (1)	(2)	(3)	(4)
titestud Título escolar	0 Elemental	266	1,000	,000	,000	,000
	1 Medio	762	,000	1,000	,000	,000
	2 Bachillerato	87	,000	,000	1,000	,000
	3 Diploma universitario	226	,000	,000	,000	1,000
	4 Licenciado	109	,000	,000	,000	,000

Bloque 0: Bloque inicial
Modelo nulo, sin la variable independiente

3 Tabla de clasificación[a,b]

			Pronosticado		
			voto Votó en 1992		Porcentaje correcto
Observado			0 Sí votó	1 No votó	
Paso 0	voto Votó en 1992	0 Sí votó	1031	0	100,0
		1 No votó	419	0	,0
	Porcentaje global				71,1

a. En el modelo se incluye una constante.

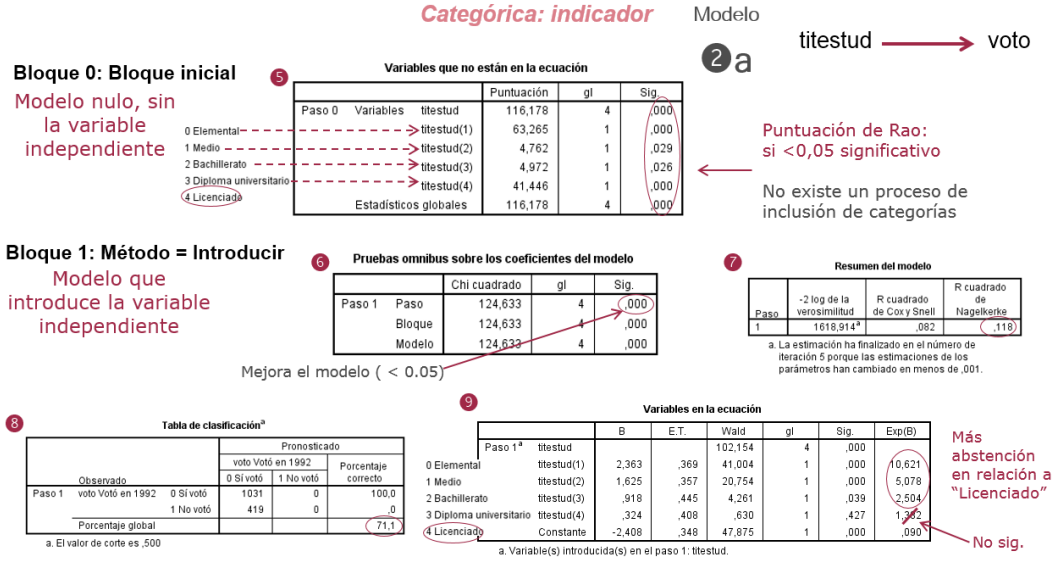
b. El valor de corte es ,500

4 Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 0	Constante	-,900	,058	241,540	1	,000	,406

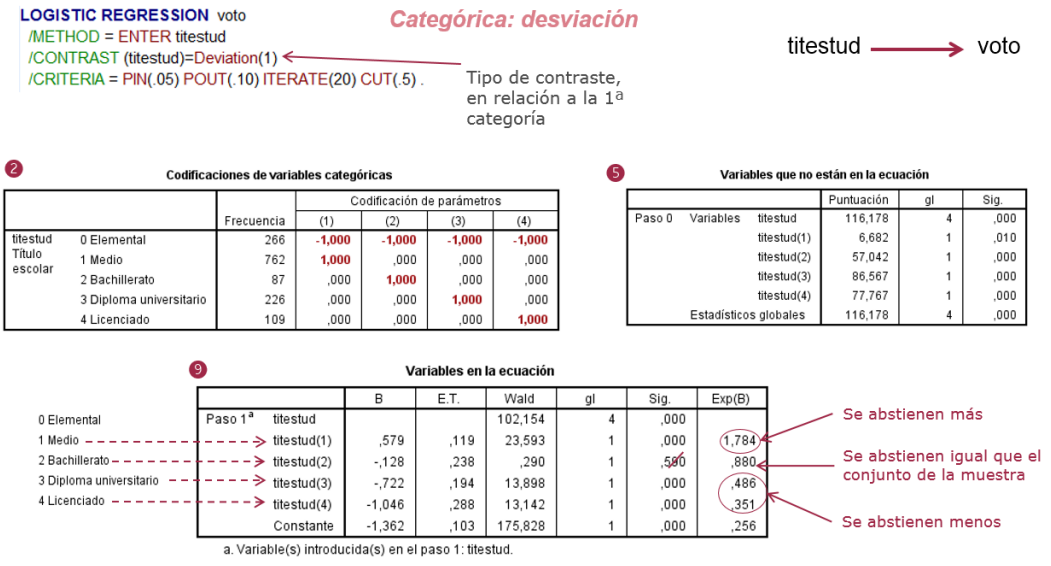
En la tabla de codificación se muestra el tipo de contraste que implica: valor 1 en la categoría en cuestión y 0 en el resto, tomando como referencia el nivel de estudios de Licenciado.

Constatamos la relevancia de la variable para explicar el abstencionismo, pues su introducción mejora el modelo, y vemos qué categorías son significativas en relación a la de referencia. En este caso todas ellas son significativas.



El modelo alcanza un destacable R^2 de Nagelkerke de 0,118. La ecuación que se obtiene pone de manifiesto que la abstención aumenta a medida que baja el nivel de estudios, observándose en particular que los Diplomados no difieren en su comportamiento abstencionista de los Licenciados.

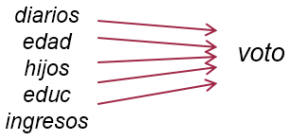
Si cambiamos el tipo de contrate por el de desviación, y cambiamos además la categoría de referencia, ahora la primera categoría, el nivel de estudios Elemental, se puede observar cómo las personas con Bachillerato se comportan como el promedio del conjunto de la muestra y por ello no difieren significativamente, y que las personas con estudios Medios se abstienen más que la media mientras que Diplomados y Licenciados se abstienen menos.



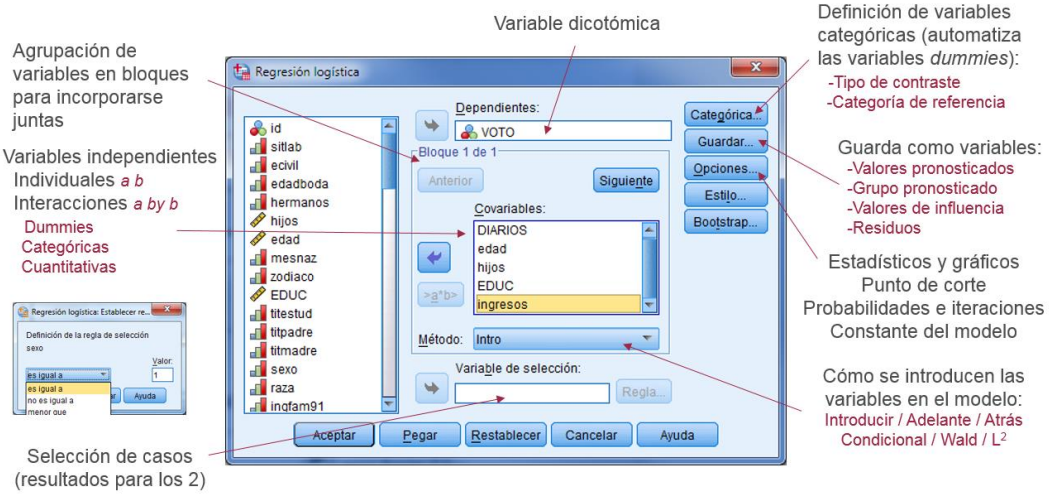
3.2. Regresión binaria múltiple

3.2.1. *Abstencionismo*

Ahora consideraremos la regresión múltiple para explicar el abstencionismo, la variable VOTO, considerando como variables independientes, la lectura de periódicos (DIARIOS), y las variables cuantitativas: edad (edad), el número de hijos (hijos), los años de escolarización (EDUC) y los ingresos individuales (ingresos).



Se busca determinar el mejor modelo explicativo a partir de estas variables consideradas inicialmente. A tal efecto disponemos de varios métodos de selección de variables que pueden elegirse en el cuadro de diálogo principal, en Método.

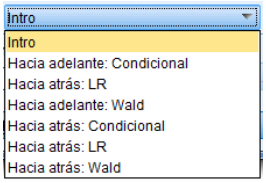


Una primera alternativa que consideraremos para la construcción de modelos de regresión logística es el método de un único paso en el que se consideran todas las variables independientes seleccionadas vez (introducción forzosa). Pero hay otros métodos posibles que pueden ser más adecuados. Los métodos de selección de variables que proporciona el SPSS son:

- Método de introducción forzosa: incluye todas las variables independientes seleccionadas. Tiene la ventaja de permitir establecer el efecto conjunto de todas las variables cuando existe colinealidad, pero también puede incluir variables que no contribuyen al ajuste del modelo.
- Método de selección por pasos: utiliza criterios estadísticos para incluir de forma automática las variables que son significativas y excluir las que no lo son. Una consecuencia de este procedimiento es la posibilidad de dejar fuera del modelo variables que teórica o conceptualmente se consideran relevantes.

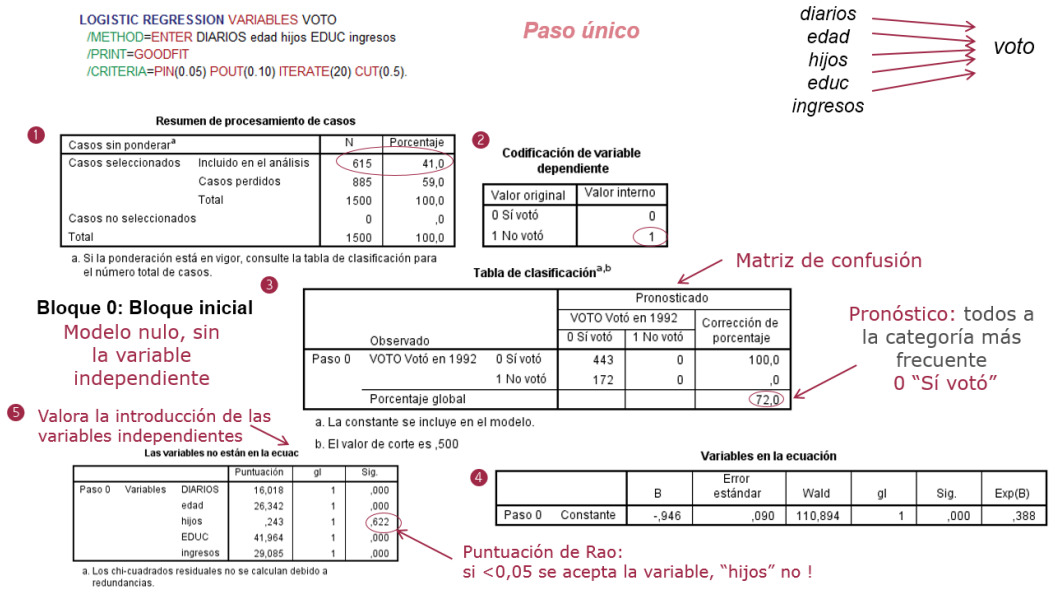
- Método de selección de bloque: permite la inclusión o exclusión de variable mediante la combinación secuenciada de diferentes procedimientos, con la posibilidad de generar modelos jerárquicos.

El estadístico empleado para incluir (o excluir) las variables que contribuyen al ajuste global del modelo en el procedimiento por pasos es el estadístico de puntuación de Rao. Para la exclusión de variables se puede escoger entre el estadístico de Wald, el cambio en la razón de verosimilitudes y el estadístico condicional.



Los métodos hacia adelante parten del modelo nulo y van introduciendo variables paso a paso hasta que no queden variables significativas para incluir. En los métodos hacia atrás se parte del modelo saturado que incluye todas las variables y se excluyen las variables paso a paso hasta que no quedan variables no significativas para excluir.

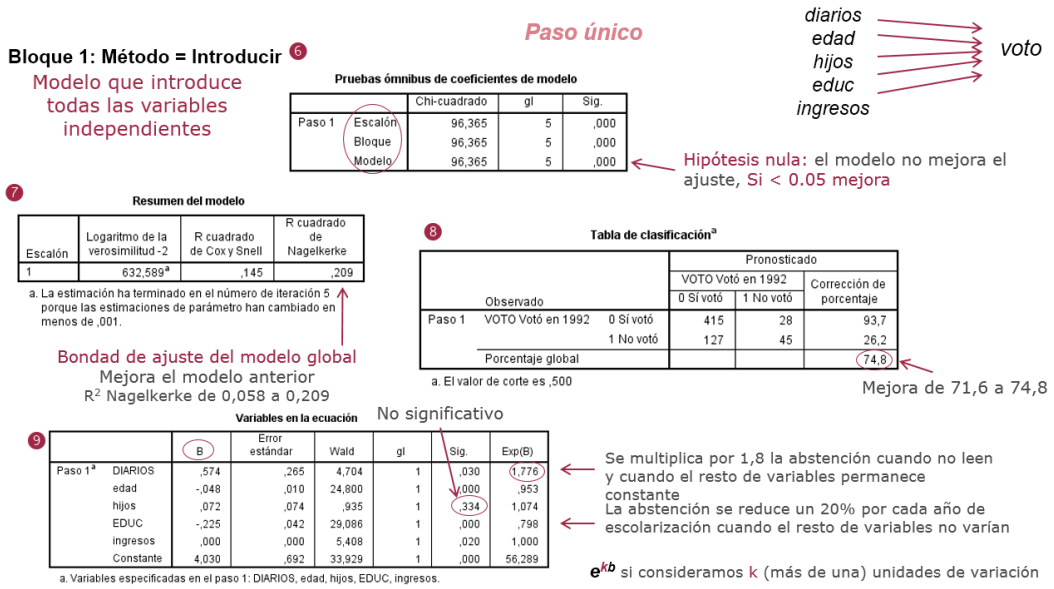
Consideraremos en primer lugar el método de paso único (método Introducir), que implica considerar simultáneamente todas las variables a la vez.



La tabla de Variables que no están en la ecuación nos proporciona el estadístico de puntuación de Rao que mide la contribución individual de cada variable en la mejora del ajuste global del modelo. De todas las variables introducidas solamente la variable hijos resulta no significativa, por tanto, el resto son candidatas a formar parte del modelo de regresión logística. También disponemos de la prueba estadística global que considera todas las variables independientes conjuntamente que resulta significativo.

En el primer paso (Bloque 1: Método = Introducir) se valora si la introducción de las cinco variables independientes logra un incremento significativo del ajuste global en

relación al modelo nulo. El estadístico de chi-cuadrado que contrasta la hipótesis nula de que el incremento es nulo resulta significativo y, por tanto, el valor de chi-cuadrado indica una mejora respecto al modelo nulo del paso 0.



El resumen del modelo nos proporciona tres estadísticos que valoran el ajuste global del modelo en el paso 1, en este caso considerando todas las variables. Si bien muestran valores bajos estos han mejorado en relación al modelo de regresión simple.

La matriz de clasificación nos muestra igualmente como el porcentaje global de clasificación correcta mejora en relación a la regresión simple, pasando del 71,6% al 74,8%.

En la tabla de coeficientes podemos ver como el estadístico de Wald, similar a la t^2, que valora su significación en relación a la hipótesis nula de que el estadístico vale 0 en la población, nos sugiere que la variable hijos no está significativamente relacionada con la variable dependiente. Un defecto de este estadístico es su sensibilidad en relación al tamaño del coeficiente, cuanto mayor es en valor absoluto más propensión a que sea significativo y es poco fiable. En este caso es recomendable evaluar la significación a través del método por pasos.

Mirando la tabla de coeficientes de la ecuación se observa que las variables edad, EDUC e ingresos dan coeficientes negativos y DIARIOS positivo. Por tanto, con esta última variable la razón de razones será un valor mayor de 1: en la columna *Exp(B)* vemos que el valor de la *odd ratio* es de 1,776, que nos cuantifica en qué grado aumenta la abstención cuantos los individuos no leen los periódicos y las demás variables permanecen constantes. En este caso, la razón del abstencionismo es casi el doble entre los que no leen periódicos en relación a los que sí leen.

Las otras variables con coeficientes negativos implican que el abstencionismo disminuye: cuando aumenta la edad, aumenta el nivel de estudios y aumentan los ingresos la abstención es menos probable. En el caso de la variable EDUC, un valor de razón de razones de 0,798, inferior a 1, significa que por cada año más de

escolarización la razón de no votar se reduce proporcionalmente en 1-0,798 = 0,202, es decir, que la razón del abstencionismo se reduce en un 20%.
Cuando disponemos de una variable cuantitativa y queremos considerar más de una unidad de variación, por ejemplo, un lustro o una década cuando consideramos años, la razón de razones asociada a k unidades de cambio se obtiene mediante e^{kb}.

Bloque 1: Método = Introducir
Modelo que introduce todas las variables independientes

Paso único

diarios
edad
hijos
educ
ingresos
→ voto

Prueba de Hosmer y Lemeshow

Escalón	Chi-cuadrado	gl	Sig.
1	6,365	8	,606

Hipótesis nula: el modelo ajusta si sig ≥ 0,05
0,606 > 0,05 ⇒ el modelo ajusta

Tabla de contingencia para la prueba de Hosmer y Lemeshow

		VOTO Votó en 1992 = 0 Sí votó		VOTO Votó en 1992 = 1 No votó		
		Observado	Esperado	Observado	Esperado	Total
Paso 1	1	58	59,158	4	2,842	62
	2	59	55,838	3	6,162	62
	3	53	53,325	9	8,675	62
	4	46	50,788	16	11,212	62
	5	48	48,059	14	13,941	62
	6	47	44,881	15	17,119	62
	7	44	41,281	18	20,719	62
	8	34	36,892	28	25,108	62
	9	33	31,755	29	30,245	62
	10	21	21,023	36	35,977	57

La prueba de Hosmer y Lemeshow muestra adicionalmente la bondad del modelo.

Ahora consideraremos el método de selección por pasos, y la opción o método hacia adelante condicional (Adelante: condicional).

```
LOGISTIC REGRESSION VOTO
  /METHOD = FSTEP(COND) DIARIOS edad hijos EDUC ingresos
  /PRINT=GOODFIT
  /CRITERIA = PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .
```

Hacia adelante condicional

diarios
edad
hijos
educ
ingresos
→ voto

Bloque 1: Método = Avanzar por paso (Condicional)
Modelo que introduce una a una las variables seleccionándolas

Pruebas ómnibus de coeficientes de modelo

		Chi-cuadrado	gl	Sig.
Paso 1	Escalón	43,333	1	,000
	Bloque	43,333	1	,000
	Modelo	43,333	1	,000
Paso 2	Escalón	41,430	1	,000
	Bloque	84,763	2	,000
	Modelo	84,763	2	,000
Paso 3	Escalón	5,798	1	,016
	Bloque	90,561	3	,000
	Modelo	90,561	3	,000
Paso 4	Escalón	4,873	1	,027
	Bloque	95,434	4	,000
	Modelo	95,434	4	,000

El modelo mejora el ajuste si sig < 0.05

"Paso": informa sobre la mejora en el ajuste en relación al paso previo debido a la variable que acaba de incorporarse

"Bloque" nos dice qué mejora se da por la incorporación de un bloque de variables (no es el caso)

"Modelo" nos indica la mejora producida debido al total de las variables incluidas en relación al paso 0.

Evaluación del modelo que incluye EDUC y edad

Evaluación del modelo que incluye EDUC, edad y Ingfam91

Evaluación del modelo que incluye EDUC, edad, ingresos y DIARIOS

Pas 1: se incorpora EDUC
Pas 2: se incorpora edad
Pas 3: se incorpora ingresos
Pas 4: se incorpora DIARIOS

La prueba de ajuste global muestra a cada paso las variaciones producidas en el ajuste como resultado de la incorporación o eliminación de cada nueva variable. En cada paso se muestran tres informaciones:

- La fila Escalón informa sobre la mejora en el ajuste en relación al paso previo debido a la variable que acaba de incorporarse.

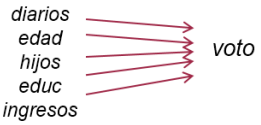
- La fila Bloque nos dice qué mejora se produce como resultado de la incorporación del bloque de variables que se considere en un procedimiento de selección de bloque.
- Y la fila Modelo nos indica la mejora producida debida al total de las variables incluidas en relación al paso 0.

En el ejemplo, el primer paso incluye la variable EDUC y supone una mejora significativa del ajuste. En el segundo paso incluye la variable edad y su inclusión (fila Escalón, del Paso 2) supone también una mejora significativa en relación al paso anterior, y el modelo resultante (fila Modelo) que incluye la constante y las variables EDUC y edad también es significativo. El Paso 3 y el Paso 4 suponen igualmente mejoras de ajuste con las variables de ingresos y de lectura de periódicos, ofreciendo el mejor ajuste el último paso que incluye todas las variables. En los resultados que aparecen seguidamente se pueden observar las mejoras progresivas del ajuste que se traducen en una mejora igualmente los estadísticos de bondad de ajuste global, hasta alcanzar el valor de 0,205.

Bloque 1: Método = Avanzar por paso (Condicional)

Modelo que introduce una a una las variables seleccionándolas

Hacia adelante condicional



7

Resumen del modelo

Escalón	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	685,621[a]	,068	,098
2	644,191[b]	,129	,185
3	638,393[b]	,137	,197
4	633,520[b]	,144	,207

a. La estimación ha terminado en el número de iteración 4 porque las estimaciones de parámetro han cambiado en menos de ,001.

b. La estimación ha terminado en el número de iteración 5 porque las estimaciones de parámetro han cambiado en menos de ,001.

Bondad de ajuste del modelo global

Mejora progresiva de los R^2

8

Tabla de clasificación[a]

Observado			Pronosticado		
			VOTO Votó en 1992		Corrección de porcentaje
			0 Sí votó	1 No votó	
Paso 1	VOTO Votó en 1992	0 Sí votó	429	14	96,8
		1 No votó	157	15	8,7
	Porcentaje global				72,2
Paso 2	VOTO Votó en 1992	0 Sí votó	411	32	92,8
		1 No votó	127	45	26,2
	Porcentaje global				74,1
Paso 3	VOTO Votó en 1992	0 Sí votó	414	29	93,5
		1 No votó	128	44	25,6
	Porcentaje global				74,5
Paso 4	VOTO Votó en 1992	0 Sí votó	412	31	93,0
		1 No votó	126	46	26,7
	Porcentaje global				74,5

a. El valor de corte es ,500

Validez del modelo

Mejora de los casos correctamente clasificados

La tabla de clasificación refuerza la idea de una mejora de la validez del modelo, si bien el último paso no implica un aumento de los casos correctamente clasificados.

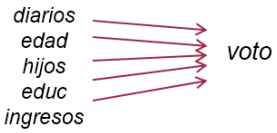
La tabla que aparece seguidamente informa de las variables introducidas en cada paso en el modelo, así como de las estimaciones de los parámetros y su significación. El último paso es el que suele tener más interés ya que es el que hace referencia al modelo final elegido. Este modelo ha supuesto la eliminación de la variable hijos. Se observa igualmente el proceso de introducción de las variables y el momento en que aparece cada una de ellas.

Los resultados son similares a los obtenidos en el caso anterior de introducción forzosa. El indicador cultural (no leer) es el que tiene un efecto positivo en el absentismo, y es el más importante. El resto tienen un efecto negativo, es decir, van en

contra del absentismo: primero, por orden de importancia, si se tienen más años de educación, segundo, si tienen más ingresos y, tercero, si se tienen más años.

Bloque 1: Método = Avanzar por paso (Condicional)
Modelo que introduce una a una las variables seleccionándolas

Hacia adelante condicional



9 **Variables en la ecuación**

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 1[a]	EDUC	-,220	,036	38,218	1	,000	,802
	Constante	1,982	,472	17,617	1	,000	7,260
Paso 2[b]	edad	-,050	,008	36,313	1	,000	,951
	EDUC	-,272	,039	48,533	1	,000	,762
	Constante	4,622	,670	47,573	1	,000	101,730
Paso 3[c]	edad	-,045	,008	28,503	1	,000	,956
	EDUC	-,243	,041	35,819	1	,000	,784
	ingresos	,000	,000	5,503	1	,019	1,000
	Constante	4,375	,677	41,710	1	,000	79,477
Paso 4[d]	DIARIOS	,588	,264	4,948	1	,026	1,801
	edad	-,044	,009	26,178	1	,000	,957
	EDUC	-,233	,041	32,385	1	,000	,792
	ingresos	,000	,000	5,103	1	,024	1,000
	Constante	4,083	,690	35,067	1	,000	59,344

a. Variables especificadas en el paso 1: EDUC.
b. Variables especificadas en el paso 2: edad.
c. Variables especificadas en el paso 3: ingresos.
d. Variables especificadas en el paso 4: DIARIOS.

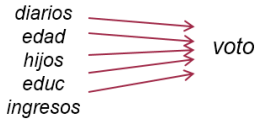
Pasos seguidos y estimación de los parámetros

Modelo final
Resultados similares a todas simultáneamente

La tabla siguiente titulada Modelo si se elimina el término proporciona, para cada paso, una evaluación de la pérdida de ajuste que se produciría en el modelo si se eliminaran, una a una, las variables ya incluidas. Siendo un procedimiento de selección hacia adelante permite la exclusión de una variable previamente incluida si se aprecia una pérdida de significación como resultado de la incorporación de nuevas variables. No es el caso.

Bloque 1: Método = Avanzar por paso (Condicional)
Modelo que introduce una a una las variables seleccionándolas

Hacia adelante condicional



10 **Modelo si el término se ha eliminado[a]**

Variable		Logaritmo de la verosimilitud de modelo	Cambio en el logaritmo de la verosimilitud -2	gl	Sig. del cambio
Paso 1	EDUC	-364,750	43,879	1	,000
Paso 2	edad	-343,155	42,119	1	,000
	EDUC	-351,172	58,153	1	,000
Paso 3	edad	-335,390	32,387	1	,000
	EDUC	-339,430	40,468	1	,000
	ingresos	-322,114	5,834	1	,016
Paso 4	DIARIOS	-319,198	4,877	1	,027
	edad	-331,572	29,624	1	,000
	EDUC	-335,005	36,489	1	,000
	ingresos	-319,455	5,390	1	,020

a. Se basa en estimaciones de parámetro condicionales.

Evaluación de la pérdida de ajuste que se produciría en el modelo si se eliminaran, una a una, las variables ya incluidas. La excluye si hay una pérdida de significación

11 **Las variables no están en la ecuación[a]**

			Puntuación	gl	Sig.
Paso 1	Variables	DIARIOS	9,539	1	,002
		edad	38,617	1	,000
		hijos	4,582	1	,032
		ingresos	14,129	1	,000
Paso 2	Variables	DIARIOS	5,483	1	,019
		hijos	,798	1	,372
		ingresos	5,596	1	,018
Paso 3	Variables	DIARIOS	5,013	1	,025
		hijos	1,178	1	,278
	Estadísticos globales		5,943	2	,051
Paso 4	Variables	hijos	,937	1	,333
	Estadísticos globales		,937	1	,333

a. Los chi-cuadrados residuales no se calculan debido a redundancias.

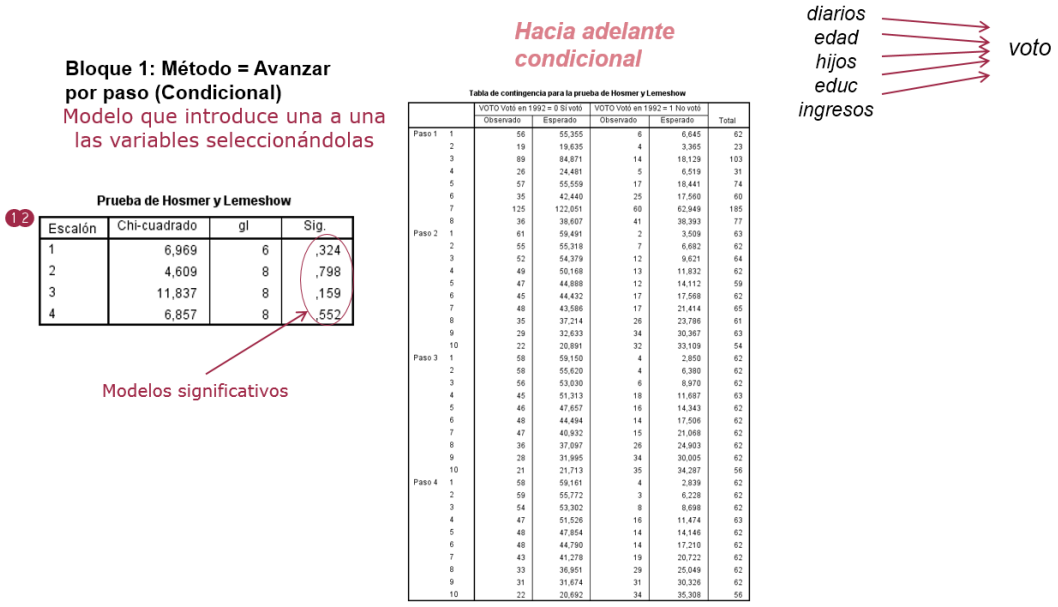
Proceso seguido donde se evalúa qué variable será incluida en el modelo en el paso siguiente: la que tenga un valor más alto

No significativa

La tabla de Variables que no están en la ecuación nos da la información de las variables aún no incluidas en el modelo a cada paso. El estadístico de puntuación de Rao se utiliza para escoger qué variable será incluida en el modelo al siguiente paso: aquella que tenga un valor más alto. Esta tabla también es interesante para observar el efecto

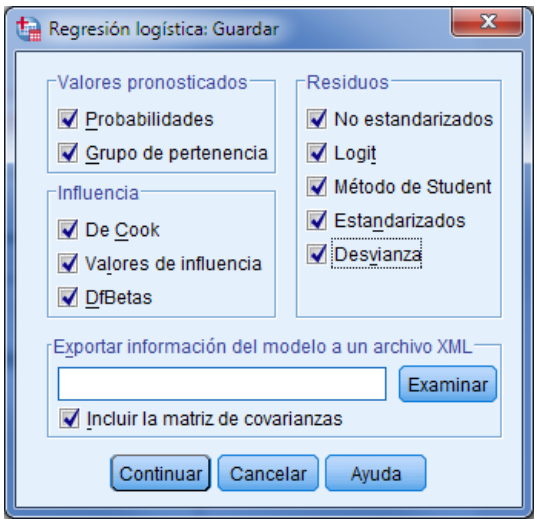
de la existencia de colinealidad, ya que algunas variables pierden significación antes de ser incluidas en el modelo.

Finalmente, se adjunta la prueba de Hosmer y Lemeshow que muestra también la bondad del ajuste al obtenerse probabilidades por encima de 0,05.



		VOTO Votó en 1992 = 0 Sí votó		VOTO Votó en 1992 = 1 No votó		
		Observado	Esperado	Observado	Esperado	Total
Paso 1	1	56	55,355	6	6,645	62
	2	19	19,635	4	3,365	23
	3	89	84,871	14	18,129	103
	4	26	24,481	5	6,519	31
	5	57	55,559	17	18,441	74
	6	35	42,440	25	17,560	60
	7	125	122,051	60	62,949	185
	8	36	38,607	41	38,393	77
Paso 2	1	61	59,491	2	3,509	63
	2	55	55,318	7	6,682	62
	3	52	54,379	12	9,621	64
	4	49	50,168	13	11,832	62
	5	47	44,888	12	14,112	59
	6	45	44,432	17	17,568	62
	7	48	43,586	17	21,414	65
	8	35	37,214	26	23,786	61
	9	29	32,633	34	30,367	63
	10	22	20,891	32	33,109	54
Paso 3	1	58	59,150	4	2,850	62
	2	58	55,620	4	6,380	62
	3	56	53,030	6	8,970	62
	4	45	51,313	18	11,687	63
	5	46	47,657	16	14,343	62
	6	48	44,494	14	17,506	62
	7	47	40,932	15	21,068	62
	8	36	37,097	26	24,903	62
	9	28	31,995	34	30,005	62
	10	21	21,713	35	34,287	56
Paso 4	1	58	59,161	4	2,839	62
	2	59	55,772	3	6,228	62
	3	54	53,302	8	8,698	62
	4	47	51,526	16	11,474	63
	5	48	47,854	14	14,146	62
	6	48	44,790	14	17,210	62
	7	43	41,278	19	20,722	62
	8	33	36,951	29	25,049	62
	9	31	31,674	31	30,326	62
	10	22	20,692	34	35,308	56

En cualquiera de los análisis se puede realizar un estudio detallado de los casos de la muestra y usar los resultados que se obtienen en la regresión logística en otros procedimientos para realizar comprobaciones de supuestos y obtener representaciones gráficas. Así, es posible utilizar una serie de especificaciones del comando a través del botón Guardar, con el cuadro de diálogo siguiente:



En relación a los valores pronosticados podemos crear nuevas variables basadas en los pronósticos del modelo:

- Probabilidades: crea una variable en la que cada caso contiene la probabilidad pronosticada por el modelo, es decir, qué probabilidad tiene un individuo de pertenecer a la categoría 1 de la variable dependiente.
- Grupo de pertenencia: crea una variable en la que cada caso es asignado al grupo de la variable dependiente donde queda clasificado segundo el punto de corte considerado, por defecto, el 0,5.

En relación a la cuestión de la existencia de casos influyentes en el modelo, se pueden crear nuevas variables con información sobre la importancia de esta influencia:

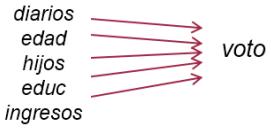
- Distancia de Cook: mide el cambio en los coeficientes de regresión cuando se elimina cada caso en la estimación de la ecuación. Un valor alto implica un peso considerable en el caso en cuestión en la determinación de los coeficientes. En general un valor superior a 1 en esta distancia exige su revisión.
- Valores de influencia: representan la influencia potencial de cada caso en el modelo. Estos valores miden la distancia de un punto en relación al centro de la distribución. Estos valores se sitúan entre 0 y 1, y su promedio es p/n, el número de parámetros estimados dividido el número de casos.
- DfBetas (Diferencia de Betas): mide el cambio del coeficiente de regresión estandarizados (betas) como resultado de eliminar cada caso de la ecuación de regresión. El software crea tantas variables como coeficientes tiene la ecuación, incluida la constante.

En relación a los residuos, la diferencia entre los valores observados y los pronosticados, se pueden calcular las siguientes variables:

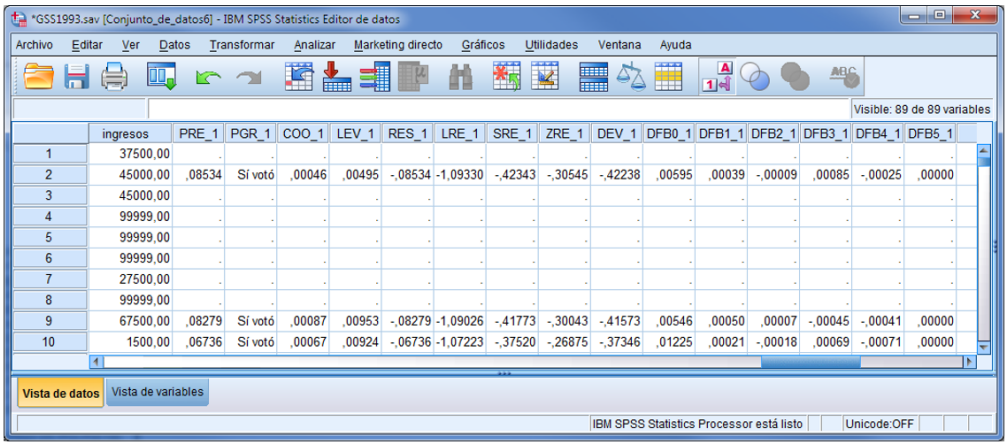
- No tipificados: residuo no tipificado o sucio resultado de restar la probabilidad pronosticada por el modelo de la probabilidad observada, siempre en relación a la categoría de la variable dependiente codificada con un 1.
- Logit: es el residuo no tipificado en la escala logit, dividido por la varianza de su pronóstico.
- Método de student: son los residuos studentitzados que nos indican el cambio observado en la desviación del modelo cuando se excluye cada caso del modelo.
- Tipificados (residuos de Pearson o residuos estandarizados): residuos resultado de dividir el residuo bruto por la estimación del error típico que se distribuyen según una $N(0,1)$. Se divide por la raíz cuadrada del producto de las probabilidades $P_i \cdot (1-P_i)$.
- Desvianza: se define como la raíz cuadrada de $-2\log \hat{P}_i$, donde $\hat{P}_i$ es la probabilidad pronosticada de pertenecer al grupo de la variable dependiente al que realmente pertenece. La desvianza se distribuye aproximadamente como una normal.

Si guardamos toda esta información ejecutando el procedimiento obtenemos una matriz de datos ampliada como la de la imagen siguiente donde aparecen en las columnas los nombres asignados a cada variable guardada. En particular: PRE_1 para los valores pronosticados, PRG_1 para el grupo de pertenencia, RES_1 para los residuos o SRE_1 para los residuos estandarizados.

Pronósticos y residuos

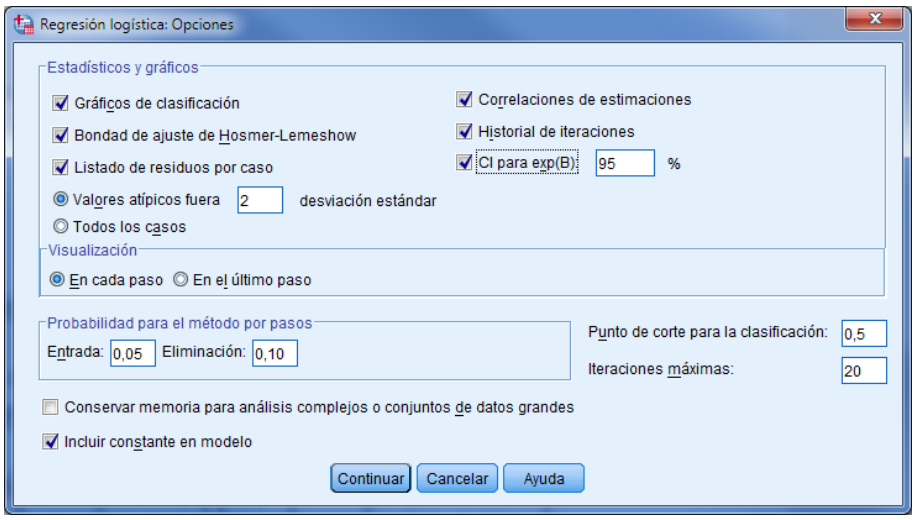


Variables guardadas de pronóstico, clasificación, influencia y residuos



	ingresos	PRE_1	PGR_1	COO_1	LEV_1	RES_1	LRE_1	SRE_1	ZRE_1	DEV_1	DFB0_1	DFB1_1	DFB2_1	DFB3_1	DFB4_1	DFB5_1
1	37500,00	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
2	45000,00	,08534	Sí votó	,00046	,00495	-,08534	-1,09330	-,42343	-,30545	-,42238	,00595	,00039	-,00009	,00085	-,00025	,00000
3	45000,00	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	99999,00	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
5	99999,00	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
6	99999,00	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
7	27500,00	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
8	99999,00	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
9	67500,00	,08279	Sí votó	,00087	,00953	-,08279	-1,09026	-,41773	-,30043	-,41573	,00546	,00050	,00007	-,00045	-,00041	,00000
10	1500,00	,06736	Sí votó	,00067	,00924	-,06736	-1,07223	-,37520	-,26875	-,37346	,01225	,00021	-,00018	,00069	-,00071	,00000

Finalmente se pueden tener en consideración algunas opciones adicionales mediante el botón Opciones con el cuadro de diálogo siguiente:



En primer lugar, podemos seleccionar algunos estadísticos y gráficos:

- Gráficos de clasificación: histograma apilado con las probabilidades pronosticadas por el modelo en el que se distingue el grupo al que pertenece cada caso, el punto de corte y los territorios de clasificación.
- Bondad de ajuste de Hosmer-Lemeshow: índice para evaluar el ajuste global del modelo, sobre todo cuando se disponen de muchas variables independientes, o cuando algunas variables independientes son continuas.
- Listado de residuos por caso: listado de los residuos no tipificados, de las probabilidades pronosticadas, del grupo observado y del grupo pronosticado, bien

de todos los casos o limitado a los casos con un residuo tipificado alejado de la media k unidades de desviación típica.

- Correlaciones de las estimaciones: ofrece la matriz de correlaciones entre las estimaciones de los parámetros del modelo.
  - Historial de iteraciones: listado con los coeficientes estimados y del logaritmo de la función de verosimilitud en cada iteración del proceso de iteración.
  - IC Para Exp(B): incluye en la tabla de estimación de los coeficientes del intervalo de confianza del valor exponencial de cada coeficiente. Por defecto con un nivel del 95%.

Además, se pueden mostrar los estadísticos, tablas y gráficos de cada paso de la estimación, o bien tan sólo los que corresponden al modelo final de cada bloque, con un resumen de los pasos intermedios.

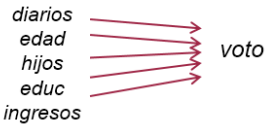
Las probabilidades para los casos determinan los niveles de significación de los métodos de selección por pasos, para valorar una variable que candidata a entrar en el modelo y a ser excluida del mismo.

El punto de corte para asignar un caso cada grupo pronosticado de la variable dependiente predeterminado es 0,5. Si el caso tiene una probabilidad superior al punto de corte se asigna al grupo codificado con el 1 de la variable dependiente.

También se puede controlar el número de iteraciones que por defecto son 20.

Y finalmente podemos considerar o no el parámetro constante en el modelo. Si no se considera es porque hemos comprobado previamente que no es significativo.

Siguiendo el ejemplo anterior presentamos a continuación algunos de estos resultados. En la primera tabla de los parámetros estimados en la ecuación se puede observar cómo se han añadido dos columnas con el intervalo de confianza de cada coeficiente con un nivel del 95%.



Opciones adicionales

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	DIARIOS	,574	,265	4,704	1	,030	1,776	1,057	2,984
	edad	-,048	,010	24,800	1	,000	,953	,936	,971
	hijos	,072	,074	,935	1	,334	1,074	,929	1,242
	EDUC	-,225	,042	29,086	1	,000	,798	,736	,866
	ingresos	,000	,000	5,408	1	,020	1,000	1,000	1,000
	Constante	4,030	,692	33,929	1	,000	56,289		

a. Variables especificadas en el paso 1: DIARIOS, edad, hijos, EDUC, ingresos.

Deben tender a ser bajos
Indicios de colinealidad

Matriz de correlaciones

		Constante	DIARIOS	edad	hijos	EDUC	ingresos
Paso 1	Constante	1,000	-,168	-,578	-,068	-,870	,111
	DIARIOS	-,168	1,000	,080	-,049	,083	,032
	edad	-,578	,080	1,000	-,444	,203	-,157
	hijos	-,068	-,049	-,444	1,000	,188	-,083
	EDUC	-,870	,083	,203	,188	1,000	-,270
	ingresos	,111	,032	-,157	-,083	-,270	1,000

La matriz de correlaciones entre las estimaciones del coeficiente nos permite hacer un análisis de la existencia de colinealidad. Estos valores deben ser bajos entre las variables

independientes, si la correlación es alta entonces la medida o la importancia del efecto de la variable independiente puede está afectada (sesgada) por la relación que mantiene con las otras variables. En este caso se puede considerar que todas las correlaciones son bajas.

La lista de casos extremos en el modelo de regresión, aquellos residuos tipificados por encima de ±2 unidades de desviación, se presenta en la tabla siguiente.

Opciones adicionales

diarios
edad
hijos
educ
ingresos
→ voto

Lista por casos[b]

Caso	Estado seleccionado [a]	Observado: VOTO Votó en 1992	Pronosticado	Grupo pronosticado	Variable temporal: Resid	Variable temporal: ZResid
262	S	N**	,137	S	,863	2,515
285	S	N**	,050	S	,950	4,352
450	S	N**	,046	S	,954	4,529
724	S	N**	,129	S	,871	2,603
759	S	N**	,050	S	,950	4,373
787	S	N**	,119	S	,881	2,719
788	S	N**	,118	S	,882	2,739
875	S	N**	,044	S	,956	4,666
1186	S	N**	,131	S	,869	2,581
1408	S	N**	,089	S	,911	3,196

a. S = Seleccionado, U = casos sin seleccionar, y ** = casos clasificados incorrectamente.

b. Se listan los casos con residuos estudentizados mayores que 2,000.

Listado de los residuos estandarizados mayores de 2

Si eliminamos los residuos tipificados con valores superiores a 3:
-El R^2 de Nagelkerke pasa de 0,209 a 0,249
-El % de clasificados correctamente pasa de 74,8% a 75,4%

Si volviéramos a ejecutar el análisis eliminando los casos que tienen un residuo tipificado superior a 3 los resultados del modelo mejoran en cierta medida: aumenta el R^2 a casi el 25% y también el porcentaje de casos correctamente clasificados.

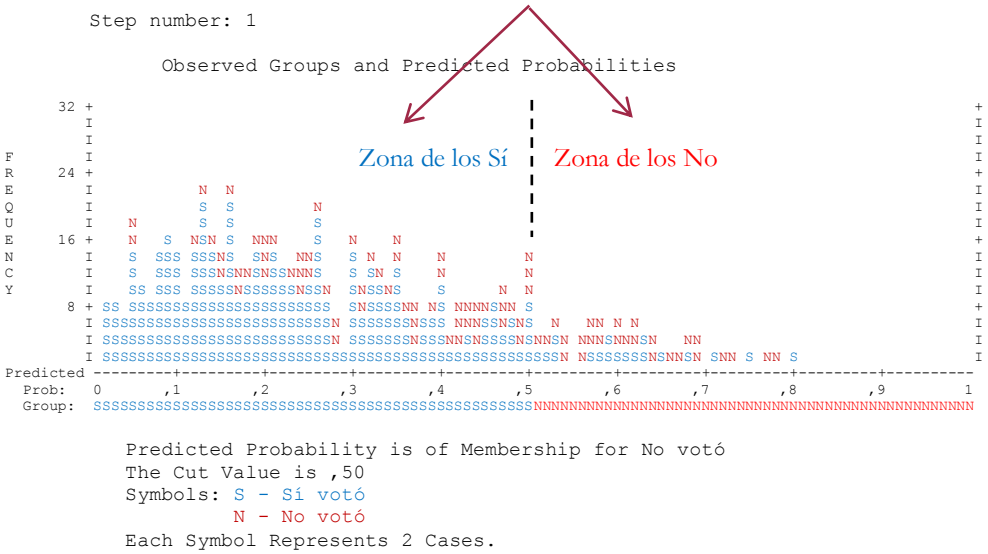
Finalmente se presenta el histograma de las probabilidades predichas de no votar y la clasificación en los dos grupos a partir del punto de corte de 0,5.

Opciones adicionales

diarios
edad
hijos
educ
ingresos
→ voto

Histograma de probabilidades pronosticadas

Los que no coinciden están mal clasificados



Predicted Probability is of Membership for No votó
The Cut Value is ,50
Symbols: S - Sí votó
N - No votó
Each Symbol Represents 2 Cases.

Cada caso se identifica con una letra (S es Sí votó y N es No votó). Cada símbolo, en este caso, representa 2 casos. El valor de corte considerado fue de 0,5 y por debajo del eje de abscisas podemos ver la región o el territorio que corresponde a cada pronóstico segundo este valor de corte. La situación ideal, de ajuste o clasificación perfecto, sería aquella en la que los símbolos del histograma fueran los mismos en la vertical de su territorio.

► **Ejercicio 1. Propuesto**
Con la misma matriz de datos GSS1993.sav se puede considerar la variable dependiente deportes y analizar los posibles factores explicativos que determinan si "Asistió a un acontecimiento deportivo en el último año". También se puede analizar la variable museos, sobre si "Visitó museos o galerías en el último año", o la variable penacap sobre la "Postura sobre la pena de muerte por asesinato". Como variables independientes se pueden considerar: DIARIOS, sexo, edad, edadcat4, ingreso4 ingreso3, EDUC, titestu2, raza, raza2, actividad, indsocec, horastv, región4, casado, hijos, titestu2, politica2, VOTO, casaprop o vida.

3.2.2. *Inserción laboral de graduados universitarios*

Complementaremos el ejemplo anterior con el análisis del ejemplo aplicado relativo a la inserción laboral de graduados universitarios que vimos. Trabajaremos con la matriz de datos ARL-AQU.sav que se encuentra en la página web junto con los archivos de sintaxis ARL-AQU.sps, que reproducen los diferentes análisis, y de resultados ARL-AQU.spv, donde se pueden comprobar los resultados que se generan.

Consideramos como variable dependiente dicotómica a los ingresos del entrevistado/a, variable ingresos, en función del máximo nivel de ocupación de los padres (OcupaciónP), del máximo nivel educativo de los padres (EstudiosP), el sexo (Sexo), si se han independizado (Emancipado), el tipo de contrato (Contrato), el número de años pasados en trabajo remunerado (AñosT) y la ocupación de los entrevistados/as (OcupaciónH). Para realizar el análisis en SPSS se debe tener en cuenta que OcupaciónP, EstudiosP, y OcupaciónH son variables cualitativas que requieren la especificación del tipo de contraste y de la categoría de referencia. Se comentan los resultados obtenidos a continuación.

El primer grupo de tablas de resultados muestra el resumen de casos, previamente haber filtrado graduados y que sean trabajadores asalariados. De esta manera se muestra que la base contiene 541 graduados de los cuales 417 se pueden utilizar para realizar el análisis.

Se muestra la codificación de la variable dependiente: 0 para ingresos bajos y 1 para ingresos altos. A su vez se presenta la codificación de las variables categóricas. El hecho de que tanto en los parámetros 1 y 2 figure cero implica que es la categoría que se toma como referencia.

En la tabla de clasificación (o matriz de confusión o matriz de clasificación correcta) se observan los pronósticos bajo el modelo nulo. Sin tener información de las variables

independientes y, simplemente con el conocimiento de la distribución observada entre graduados con ingresos altos o bajos. Como se clasifican los casos en la categoría más frecuente de la variable dependiente observada (que en este caso corresponde al valor 0, ingresos bajos). El 100% de los casos correctamente clasificados coincide con el porcentaje de casos que pertenecen a la categoría más numerosa (ingresos bajos). Si no hemos modificado el punto de corte, por defecto será del 0,5.

Modelo 1
Ampliado

```
LOGISTIC REGRESSION VARIABLES IngresosG
 /METHOD=ENTER OcupacionP EstudiosP Sexo  Emancipado Contrato AñosT OcupacionH
 /CONTRAST (OcupacionP)=Indicator(1)
 /CONTRAST (EstudiosP)=Indicator
 /CONTRAST (Sexo)=Indicator(1)
 /CONTRAST (Emancipado)=Indicator(1)
 /CONTRAST (Contrato)=Indicator(1)
 /CONTRAST (OcupacionH)=Indicator(1)
  /PRINT=GOODFIT
 /SAVE=ZRESID
 /CASEWISE OUTLIER(3)
 /CRITERIA=PIN(.05) POUT(.10) ITERATE(20) CUT(.5).
```

Codificación de la variable dependiente

Valor original	Valor interno
,00 Bajos	0
1,00 Altos	1

Resumen del procesamiento de los casos

Casos no ponderados[a]		N	Porcentaje
Casos seleccionados	Incluidos en el análisis	417	77,1
	Casos perdidos	124	22,9
	Total	541	100,0
Casos no seleccionados		0	,0
Total		541	100,0

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

Codificaciones de variables categóricas

		Frecuencia	Codificación de parámetros (1)	(2)
OcupacionH Ocupación de los hijos	1,00 Trabajadores no manuales	250	,000	,000
	2,00 Trabajadoras cualificados	157	1,000	,000
	3,00 Trabajadoras no cualificados y agrícolas	10	,000	1,000
EstudiosP Nivel estudio padres en 3 categorías	1,00 Hasta Primario	196	1,000	,000
	2,00 Secundario	113	,000	1,000
	3,00 Universitario	108	,000	,000
OcupacionP Ocupación de los padres	1,00 Trabajadores no manuales	295	,000	,000
	2,00 Trabajadoras cualificados	94	1,000	,000
	3,00 Trabajadoras no cualificados y agrícolas	28	,000	1,000
Emancipado Emancipado	,00 No vive con los padres	181	,000	
	1,00 Vive con los padres	236	1,000	
Contrato Tipo de contrato	,00 Contrato temporal	158	,000	
	1,00 Contrato indefinido	259	1,000	
Sexo Sexo	1 Varón	183	,000	
	2 Mujer	234	1,000	

Tabla de clasificación[a,b]

Observado			Pronosticado: IngresosG Ingresos horarios de los graduados ,00 Bajos	1,00 Altos	Porcentaje correcto
Paso 0	IngresosG Ingresos horarios de los graduados	,00 Bajos	232	0	100,0
		1,00 Altos	185	0	,0
	Porcentaje global				55,6

a. En el modelo se incluye una constante.
b. El valor de corte es ,500

La etapa o paso 0 con este modelo nulo proporciona la estimación del término constante (B=−0,226), los estadísticos asociados y la significación (0,022).

A continuación, se ofrece un contraste de hipótesis para valorar cada una de las variables independientes del modelo. Viendo la significación de cada variable se puede observar cuáles de ellas mejorarían el modelo inicial, en nuestro caso sexo y ocupación.

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 0	Constante	-,226	,099	5,275	1	,022	,797

Variables que no están en la ecuación

			Puntuación	gl	Sig.
Paso 0	Variables	OcupacionP	2,497	2	,287
		OcupacionP(1)	1,810	1	,179
		OcupacionP(2)	1,031	1	,310
		EstudiosP	,164	2	,921
		EstudiosP(1)	,163	1	,686
		EstudiosP(2)	,063	1	,802
		Sexo(1)	5,506	1	,019
		Emancipado(1)	5,414	1	,020
		Contrato(1)	1,534	1	,215
		AñosT	6,905	1	,009
		OcupacionH	44,394	2	,000
		OcupacionH(1)	38,887	1	,000
		OcupacionH(2)	2,464	1	,116
	Estadísticos globales		61,688	10	,000

Pruebas omnibus sobre los coeficientes del modelo

		Chi cuadrado	gl	Sig.
Paso 1	Paso	65,735	10	,000
	Bloque	65,735	10	,000
	Modelo	65,735	10	,000

Resumen del modelo

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	507,041[a]	,146	,195

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

Prueba de Hosmer y Lemeshow

Paso	Chi cuadrado	gl	Sig.
1	6,697	8	,570

Tabla de clasificación[a]

Observado			Pronosticado: IngresosG Ingresos horarios de los graduados ,00 Bajos	1,00 Altos	Porcentaje correcto
Paso 1	IngresosG Ingresos horarios de los graduados	,00 Bajos	157	75	67,7
		1,00 Altos	63	122	65,9
	Porcentaje global				66,9

a. El valor de corte es ,500

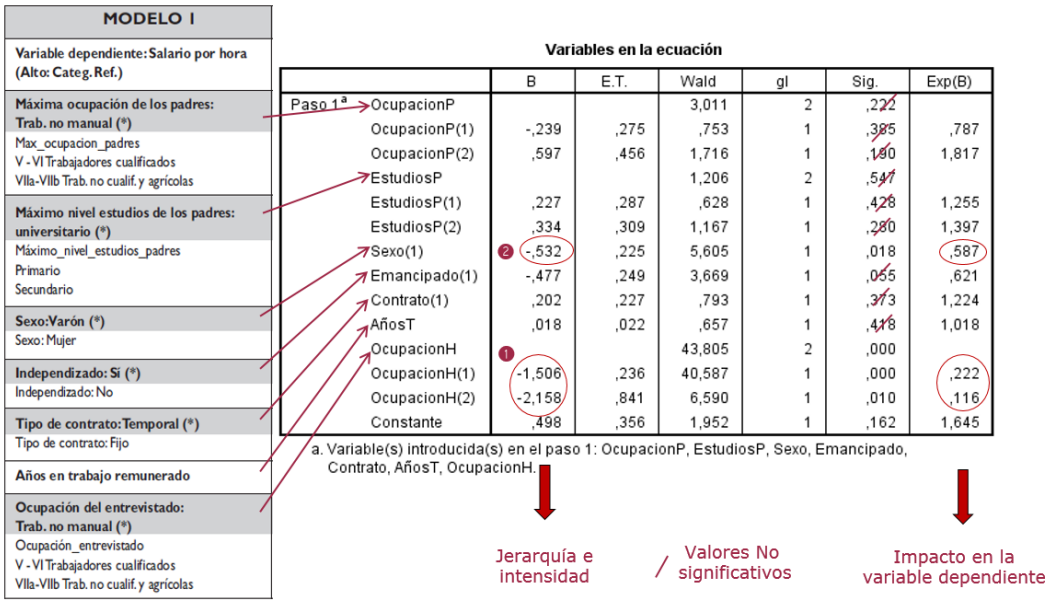
La prueba ómnibus muestra que la introducción de las variables independientes mejora el ajuste del modelo nulo.

El resumen del modelo presenta la bondad de ajuste del modelo considerando todas las variables introducidas. Es decir, los Pseudo R^2 de Cox y Snell y de Nagelkerke.

Otro indicador de la bondad de ajuste lo presenta la prueba de Hosmer y Lemeshow, en este caso el indicador de un buen ajuste implica obtener un valor mayor o igual a 0,05.

La tabla de clasificación ahora muestra los casos ya clasificados teniendo en cuenta el modelo con todas las variables y en ese sentido muestra que el 66,9% de los casos está bien clasificado.

Finalmente se presenta la tabla de estimación de los coeficientes de la ecuación que comentamos en el texto precedentemente.



		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1[a]	OcupacionP			3,011	2	,222	
	OcupacionP(1)	-,239	,275	,753	1	,385	,787
	OcupacionP(2)	,597	,456	1,716	1	,190	1,817
	EstudiosP			1,206	2	,547	
	EstudiosP(1)	,227	,287	,628	1	,428	1,255
	EstudiosP(2)	,334	,309	1,167	1	,280	1,397
	Sexo(1)	-,532	,225	5,605	1	,018	,587
	Emancipado(1)	-,477	,249	3,669	1	,055	,621
	Contrato(1)	,202	,227	,793	1	,373	1,224
	AñosT	,018	,022	,657	1	,418	1,018
	OcupacionH			43,805	2	,000	
	OcupacionH(1)	-1,506	,236	40,587	1	,000	,222
	OcupacionH(2)	-2,158	,841	6,590	1	,010	,116
	Constante	,498	,356	1,952	1	,162	1,645

► Ejercicio 2. Propuesto

Repetir el análisis para la población trabajadora asalariada (Hipótesis 2) de la ECV utilizando el archivo ARL-ECV2005.sav. Reproducir el análisis de los datos y realizar la lectura de los mismos siguiendo las recomendaciones realizadas en el texto. Para guiar la realización de la regresión se puede recurrir al archivo ARL-ECV2005.sps. En el archivo ARL-ECV2005.spv se podrán comprobar los resultados de la aplicación de los cuatro modelos.

4. Bibliografía

Achen, Christopher (1982). *Interpreting and Using Regression*. Beverly Hills: Sage

Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley and Sons.

Agresti, A.; Finlay, B. (2009). *Statistical Methods for the Social Sciences*. Upper Saddle River (New Jersey): Pearson Prentice Hall. 4ª edición.

Aldrich, J. H., and F. D. Nelson (1984). *Linear probability, logit, and probit models*. Beverly Hills, California: Sage Publications.

Andersen, E. (1997). *Introduction to the statistical analysis of categorical data*. Springer.

Boix, C.; Riba, C. (2000). Las bases sociales y políticas de la abstención en las elecciones generales españolas: Recursos individuales, movilización estratégica e institucione electorales. *Revista Española de Investigaciones Sociológicas*, 90, 95-128.

Cornfield, J.; Gordon, T.; Smith, W. N. (1961). Quantal response curves for experimentally uncontroled variables. *Bulletin of the International Statistical Institute*, 38, 97-115.

Cortés, F. (1997). Regresión Logística en la investigación social: potencialidades y limitaciones. *Revista de Ciencias Sociales*, 13, 67-78.

Christensen, R. (1997). *Log-linear models and logistic regression*. New York: Springer-Verlag.

Demaris, A. (1992). *Logit Modeling. Practical Aplications*. Beverly Hills: Sage

Escribà, A. (2006). Estructura familiar, estatus ocupacional y movilidad social intrageneracional en España. *Revista Internacional de Sociología*, LXIV, 45, septiembre-diciembre, 145-170.

Fachelli, S.; Planas, J. (2011). Equidad y movilidad intergeneracional de los titulados universitarios catalanes". *Papers Revista de Sociología*, 94, 6, 1307-1331. http://ddd.uab.cat/pub/papers/02102862v96n4/02102862v96n4p1307.pdf

Fachelli, S., Planas, J. y Navarro-Cendejas, J. (2012). En qué medida la trayectoria académica y el origen social de los titulados universitarios catalanes influyen en su inserción laboral. En M. Venegas (coord.), *La sociología y los retos de la educación en la España Actual*. XV Conferencia de Sociología de la Educación. Editorial Germania: Valencia, 1-24.

Fachelli, S. y Planas, J. (2014). Inserción profesional y movilidad intergeneracional de los universitarios: de la expansión a la crisis. *Revista Española de Sociología*. 21, 69-98. http://ddd.uab.cat/pub/artpub/2014/125654/revespsoc_a2014n21p69iSPA.pdf

Fachelli, S.; Torrents, D.; Navarro-Cendejas, J. (2014). ¿La universidad española suaviza las diferencias de classe en la inserción laboral? *Revista de Educación*, 364. abril-junio, 119-144. https://ddd.uab.cat/record/118532

Fachelli, S. y Navarro-Cendejas, J. (2015) Relationship between social origin and university graduates' entry into the labour market. *Electronic Journal of Educational Research, Assessment and Evaluation RELIEVE*. 21, 2, pp. 1-22. http://dx.doi.org/10.7203/relieve.21.2.7812

García, M. V.; Alvarado, J. M.; Jiménez, A. (2000). La predicción del rendimiento académico: regresión lineal versus regresión logística. *Psicothema*, 12, 2, 248-252.

Hair, J. F. et al. (2011). *Multivariate Data Analysis*. Upper Saddle River: Prentice Hall.

Hellevik, O. (2007). Linear versus logistic regression when the dependent variable is a dichotomy. *Quality and Quantity*, 43, 1, 59-74.

Hosmer, D. W.; Lemeshow, S. (1989). *Applied Logistic Regresión*. New York: John Wiley.

Jaccard, J. (2001). *Interaction Effects in Logistic Regression*. Thousand Oaks, CA: Sage.

Jovell, A. J. (1995). *Análisis de regresión logística*. Madrid: Centro de Investigaciones Sociológicas.

Kelley, M. R. (1990). New process technology, job design, and work organisation: a contingency model. *American Sociological Review*, 55, 2, abril, 191-208.

Kleinbaum, D. G. (1992). *Logistic Regresion*. New York: Springer.

Kleinbaum, D. G.; Klein, M. (2010). *Logistic Regression. A Self-Learning Text*. New York: Springer.

Kutner, M. H.; Nachtsheim, C. J.; Neter, J.; Li, W. (2005). *Applied Linear Statistical Models*. New York: McGraw-Hill Irvin.

Landau, S.; Everitt, B. S. (2004). *A Handbook of Statistical Analyses using SPSS*. Boca Raton, Florida: Chapman & Hall/CRC

Luque, M. E.; Ferrer, M.; Capdevila, M. (2005). *La reincidència penitenciària a Catalunya*. Barcelona: Generalitat de Catalunya, Centre d'Estudis Jurídics i Formació Especialitzada.

Marrero, J. R.; Abdul-Jalbar, B. (2015). Las exigencias emocionales en el trabajo. El caso español. *Papers. Revista de Sociologia*, 100, 2, 173-193.

Martín-Artiles, A.; López-Roldán, P.; Molina, O. (2011). Movilidad ascendente de la inmigración en España: ¿asimilación o segmentación ocupacional?. *Papers. Revista de Sociologia*, 96, 4, 1311-1338. http://papers.uab.cat/article/view/225/papers_96_4-martin_artiles

Meil Landwerlin, G. (2005). El reparto desigual del trabajo doméstico y sus efectos sobre la estabilidad de los proyectos conyugales. *Revista Española de Investigaciones Sociológicas*, 111, 163-179.

Menard, S. (1995). *Applied Logistic Regression Analysis*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-106. Thousand Oaks, CA: Sage.

McCullag, P.; Nelder, J. A. (1989). *Generalized linear models*. London: Chapman and Hall.

Miguélez, F; López-Roldán, P. (Coord.) (2014). *Crisis, empleo e inmigración en España. Un análisis de las trayectorias laborales*. Bellaterra (Barcelona): Universitat Autònoma de Barcelona.

Miguélez, F.; Martín, A.; de Alós-Moner, R.; Esteban, F.; López-Roldán, P.; Molina, Ó.; Moreno, S. (2011). *Trayectorias laborales de los inmigrantes en España*. Barcelona: Obra Social "la Caixa". http://multimedia.lacaixa.es/lacaixa/ondemand/obrasocial/pdf/Trayectorias_laborales_de_los_inmigrantes_en_Espana.pdf.

O'Connell, A. A. (2006). *Logistic Regression Models for Ordinal Responses Variables*. Thousand Oaks: Sage.

Pampel, F. C. (2000). *Logistic Regression: A Primer*. Thousand Oaks, CA: Sage.

Pardo, A.; Ruiz, M. A. (2002). *Guía para el análisis de datos*. Madrid: McGraw-Hill.

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

Planas, J.; Fachelli, S. (2010). *Catalan universities as a factor of equity and professional mobility*. Barcelona: AQU. https://ddd.uab.cat/record/114675

Reyneri, E. (2006). De la economía sumergida a la devaluación profesional: nivel educativo e inserción en el mercado de trabajo de los inmigrantes en Italia. *Revista Española de Investigaciones Sociológicas*, 116, 213-237.

Rindskopf, D. (2002). Infinite Parameter Estimates in Logistic Regression: Opportunities, Not Problems. *Journal of Educational and Behavioral Statistics*, 27, 2, 147-161.

Rodríguez, C. y Gutiérrez, J. (2007) Empleo de modelos de regresión logística binomial para el estudio de variables determinantes en la inserción laboral de egresados universitarios. *Investigación y Postgrado*, 22, 1, 109-144.

Rodríguez-Ayán, M. N. (2005). La perspectiva estudiantil sobre el desempeño del profesor: un modelo de regresión logística ordinal. *Revista Electrónica de Metodología Aplicada*, 10, 1, 1-13.

Santos Peñas, J.; Muñoz, A.; Juez, P.; Guzmán, L. (1999). *Diseño y tratamiento estadístico de encuestas para estudios de mercado*. Madrid: Centro de Estudios Ramón Areces.

Scrucca, L. (2003). Graphics for studying logistic regression models. *Statistical Methods & Applications*, 11, 371-394.

Sánchez Vizcaíno, G. (2000). Regresión Logística. En T. Luque Martínez (ed.), *Técnicas de Análisis de Datos en Investigación de Mercados*. Madrid: Pirámide.

Silva, L. C.; Barroso, I. M. (2004). *Regresión logística*. Madrid: La Muralla.

Silva, L. C. (1994). *Excursión a la regresión logística en ciencias de la salud*. Madrid: Díaz Santos.

Stanek, M. (2011). Nichos étnicos y movilidad socio-ocupacional. El caso del colectivo polaco en Madrid. *Revista Española de Investigaciones Sociológicas*, 135, julio-septiembre, 69-88.

Tabachnick, B.; Fidell, L. S. (2007). *Using Multivariate Statistics*. Boston: Pearson.

Tejeiro, E. (1991). Algunas técnicas multivariantes útiles para la presentación de los resultados de una encuesta. *Estadística Española*, 33, 127, mayo-agosto, 305-324.

Tonidandel, S.; LeBreton, J.M. (2010). Determining the Relative Importance of Predictors in Logistic Regression.. An Extension of Relative Weight Analysis. *Organizational Research Methods*, 13, 4, 767-781.

Torrents, D.; Fachelli, S. (2015) El efecto del origen social con el paso del tiempo: la inserción laboral de los graduados universitarios españoles durante la democracia. *Revista Complutense de Educación*, 26, 2, 331 -349. https://ddd.uab.cat/record/131441

Verge, T.; Tormos, R. (2012). La persistencia de las diferencias de género en el interés por la política. *Revista Española de Investigaciones Sociológicas*, 138, abril-junio, 89-108.

Walter, S., Duncan, D. (1967). Estimation of the probability of an event as a function of several variables. *Biometrika*, 54, 167-79.