ESTADÍSTICA ESPAÑOLA
Vol. 38, Núm. 141, 1996, págs. 193 a 217

La regresión logística. Una aplicación a la demanda de estudios universitarios

por
MANUEL SALAS VELASCO
Departamento de Economía Aplicada
Universidad de Granada

RESUMEN

La regresión logística se incluye dentro del conjunto de las denominadas técnicas estadísticas del análisis de datos. Su uso se hace imprescindible cuando se quiere relacionar una variable dependiente cualitativa con una o más variables independientes. Este artículo, adoptando un enfoque práctico, analiza, mediante la técnica de la regresión logística, qué factores determinan que unos alumnos elijan una carrera de ciclo largo (Licenciatura en Administración y Dirección de Empresas), frente a una carrera de ciclo corto (Diplomatura en Ciencias Empresariales), estimando la influencia que factores de tipo económico, geográfico y familiar ejercen sobre tal elección.

Palabras clave: Elección de carrera. Odds Ratio. Estadístico de Wald. Regresión Logística.

Clasificación AMS: 62J99 62P20 90A99

1. INTRODUCCIÓN

Son muchos los problemas y cuestiones de interés en Economía en los que la variable endógena no toma en la muestra todos los valores de un intervalo real,

sino sólo un número finito de ellos; a veces, esta variable ni siquiera es cuantificable. El caso más frecuente de variables endógenas discretas surge cuando el investigador pretende utilizar un modelo econométrico para explicar la decisión tomada por un agente económico utilizando para ello un vector de características de dicho individuo.

Es en este contexto en el que el presente trabajo intenta explicar las decisiones de inversión en educación superior de la cohorte de alumnos de la Universidad de Granada que inician sus estudios en la Facultad de Ciencias Económicas y Empresariales en el curso académico 1993-94, analizando los factores que determinan que unos alumnos demanden estudios de ciclo corto, Diplomatura en Ciencias Empresariales (DCE) con una duración de tres años, y otros alumnos se decanten por los estudios de ciclo largo, Licenciatura en Administración y Dirección de Empresas (LADE) con una duración de cuatro años(1).

Cuando la decisión tomada es una variable de naturaleza cualitativa, pasa a representarse mediante una variable cuantitativa que toma un valor diferente para cada una de las posibles opciones dentro del conjunto de elección. En esta situación, considerando el caso en que se pretende explicar la elección de una entre dos alternativas, la variable dependiente puede tomar dos valores: $Y = \{0,1\}$, según que el individuo escoja la primera o la segunda alternativa, y se pretende explicar la elección hecha por el decisor como función de unas variables que le caracterizan y que denotamos por X_i. Por ejemplo, en el caso de la elección de carrera, podría definirse "Y = 0" si el individuo elige DCE, e "Y = 1" si elige LADE, aunque tal asignación es arbitraria.

Este artículo, adoptando una posición más pragmática que teórica, trata uno de los métodos más apropiados de estimación de aquellos modelos econométricos en los que la variable dependiente tiene esta naturaleza cualitativa; nos referimos a la *regresión logística*(2).

---

(1) En el curso académico 1993-94 la Facultad de Ciencias Económicas y Empresariales ofertaba sólo estas dos titulaciones. Teniendo en cuenta que, por un lado, el gasto en matrícula era muy similar para ambas carreras (61.868 pesetas para DCE y 60.968 pesetas para LADE) y, por otro lado, la nota de corte en el acceso a la Universidad, en la Fase A de la preinscripción, fue muy similar para las dos titulaciones (6,6 en DCE y 6,65 en LADE), podríamos afirmar que existe, para la cohorte objeto de estudio, una "libertad" de elección de carrera.

(2) En la literatura estadística se han sugerido otros modelos de respuesta cualitativa adicionales, entre los que cabe destacar el *modelo probit* y el *modelo de probabilidad lineal*.

Entre las aplicaciones de los modelos de respuesta cualitativa a una gran variedad de decisiones económicas caben destacar, entre otras: (i) Elección de vivienda: Li (1977), (ii) Elección de transporte: Domencich y McFadden (1975), (iii) Elección de profesión: Boskin (1974); Schmidt y Straus (1975), y (iv) Elección de estudios universitarios: Mora (1989).

2. EL ANÁLISIS DE REGRESIÓN LOGÍSTICA

Los objetivos del modelo de regresión logística son, principalmente, tres: (i) determinar la existencia o ausencia de relación entre una o más variables independientes (X_i) y una variable dependiente dicotómica (Y), es decir, que sólo admite dos categorías que definen opciones o características mutuamente excluyentes u opuestas. Las variables independientes pueden ser cualitativas binarias (género: masculino o femenino) o categóricas (niveles educativos: sin estudios, estudios primarios, Bachiller o equivalente, estudios universitarios), y cuantitativas o continuas (edad en años); (ii) medir el signo de dicha relación, en caso de que exista; y (iii) estimar o predecir la probabilidad de que se produzca el suceso o acontecimiento definido como "Y = 1" en función de los valores que adoptan las variables independientes.

Un modelo de regresión logística permite, por ejemplo, predecir o estimar la probabilidad de que un individuo vaya a la Universidad una vez acabada la Enseñanza Secundaria (acudir a la Universidad, Y, sí/no) en función de determinadas características individuales (X_i): nivel económico de la familia, edad en años, género (masculino/femenino), zona de residencia (rural/urbana), etcétera. La técnica del análisis de regresión logística también se puede utilizar para predecir la probabilidad de que dicho individuo, una vez que ha decidido estudiar en la Universidad y en un área concreta, por ejemplo, Empresariales, demande una carrera de ciclo largo (LADE) frente a una carrera de ciclo corto (DCE). Al estimar la probabilidad de que un individuo elija LADE o DCE en su ingreso en la Facultad de Ciencias Económicas y Empresariales, el modelo permite definir el perfil del Licenciado frente al del Diplomado, siempre y cuando se admita el supuesto de que la elección de carrera es una característica que permanece estable en la población entrevistada(3).

El objetivo de este trabajo es analizar, mediante un modelo de regresión logística, la relación existente entre determinadas características del colectivo encuestado (variables independientes) y la elección de carrera (variable dependiente dicotómica). En el modelo propuesto se denomina "Y = 1" a la opción de respuesta

---

(3) La fuente de datos utilizada es de elaboración propia en base a un cuestionario que se les suministró a los alumnos matriculados en primero de DCE y LADE, en el curso académico 1993-94. El total de encuestas válidas, 300, representa el 43,5 por ciento de los alumnos matriculados, lo que pone de manifiesto la alta representatividad de la encuesta y, por tanto, los resultados son extrapolables a la población total.

Se ha excluido del análisis al colectivo de alumnos que proceden de FP-II ya que dichos alumnos no pueden elegir entre Diplomatura y Licenciatura, matriculándose, los admitidos en el Centro, en DCE.

"LADE" y se define "Y = 0" a su alternativa "DCE". La variable anterior es dicotómica definida por las opciones opuestas (1 = LADE; 0 = DCE)(4).

El modelo más sencillo para estudiar qué características de los individuos determinan la elección de carrera, es aquél que incluye una sóla variable independiente (X):

$$Y = \alpha + \beta X + u$$

donde "α" es el término independiente o constante; "β" es el coeficiente de regresión asociado a la variable independiente; y "*u*" es el término de perturbación aleatoria.

Sin embargo, como Y sólo toma el valor 1 ó 0 para cada individuo en la muestra, entonces para cada observación la perturbación "*u*" debe ser una variable aleatoria que solamente puede tomar los valores [1 - (α+βX)] y [-(α+βX)], respectivamente. Además, para que E(*u*) = 0, las probabilidades con que "*u*" debe tomar estos dos valores han de ser (α+βX) y [1 - (α+βX)], respectivamente:

$$P(Y=1) = P[\, u = 1 - (\alpha + \beta X)] = \alpha + \beta X$$

$$P(Y=1) = \alpha + \beta X$$

Nosotros estaríamos interesados, pues, en estimar el siguiente modelo:

$$P = \alpha + \beta X$$

donde "P" es la probabilidad estimada de que un estudiante seleccionado al azar demande LADE(5). Sin embargo, debido a las limitaciones de la ecuación "$\alpha + \beta X$"

---

(4) En muchos casos el agente económico debe elegir una entre más de dos opciones alternativas. El modelo logístico puede también generalizarse y estimarse para estos casos de variables dependientes con múltiples categorías. Por ejemplo, y para la elección de una carrera universitaria, supongamos que cada individuo de una muestra puede escoger una de entre cuatro alternativas posibles:

Y=1 si el individuo elige estudios Científico-Técnicos (Informática, Arquitectura,...); Y=2 si el individuo elige estudios de Ciencias de la Salud (Farmacia, Medicina,...); Y=3 si el individuo elige estudios Humanísticos (Psicología, Historia,...); e Y=4 si el individuo elige estudios Socio-Jurídicos (Derecho, Empresariales,...). En este caso, un método de estimación apropiado lo ofrece el *modelo logit multinomial*. [Véanse, por ejemplo, los estudios de Schmidt y Strauss (1975), y McFadden (1974b), los cuales proporcionan aplicaciones bien conocidas de modelos de alternativas múltiples para la elección de profesión y demanda de educación superior, respectivamente].

(5) P es, por tanto, la probabilidad de que ocurra el suceso definido como "Y = 1": P = P(Y = 1).

para estimar valores de "P(Y=1)" entre su rango real de 0 a 1, se intenta ajustar el modelo:

$$P = e^{(\alpha + \beta X)}$$

expresión que a veces también se denota por:

$$P = EXP(\alpha + \beta X)$$

El modelo anterior sólo permite estimar valores de "P(Y=1) > 0", pero también mayores que 1, por lo que tampoco constituye el modelo adecuado para predecir la probabilidad de elegir LADE. El modelo que mejor estima tal probabilidad, debido a que restringe los valores predichos a su rango natural [0,1] es:

$$P = \frac{e^{(\alpha+\beta X)}}{1 + e^{(\alpha+\beta X)}} = \frac{1}{1 + e^{-(\alpha+\beta X)}}$$

expresión que se conoce como *función logística*, y que puede expresarse también de la siguiente forma:

$$\frac{P}{1-P} = e^{(\alpha + \beta X)}$$

La función logística es exponencial aunque puede transformarse, tomando logaritmos neperianos (Ln), en una función lineal:

$$Ln\left[\frac{P}{1-P}\right] = \alpha + \beta X$$

con lo que nos encontramos de nuevo con el modelo lineal clásico.

Sabemos que:

$$P = P(Y=1)$$

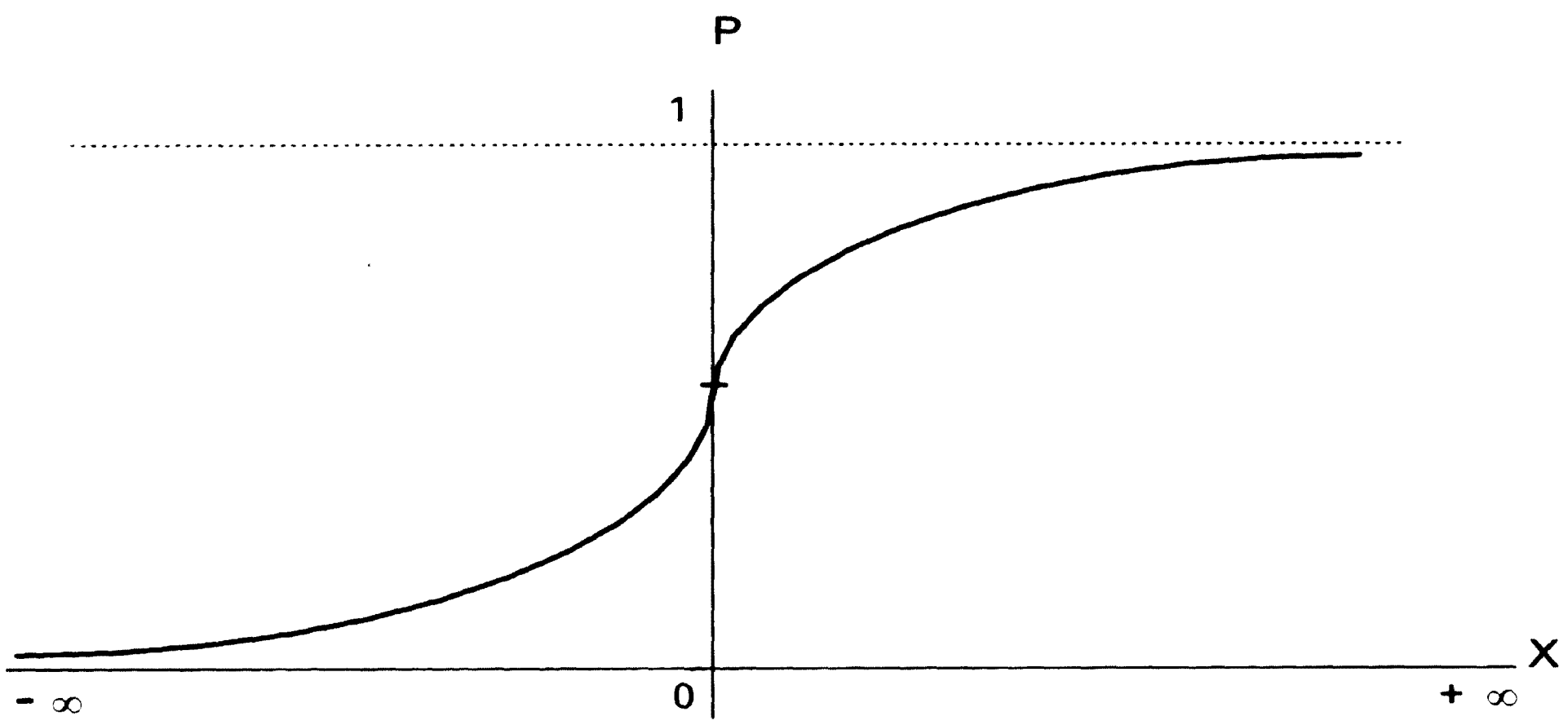
Luego:

$$Ln\left[\frac{P(Y=1)}{1-P(Y=1)}\right] = Ln\left[\frac{P(Y=1)}{P(Y=0)}\right]$$

donde el cociente [P(Y=1)/P(Y=0)] se conoce como *odds*. Por tanto, la *odds* es la razón entre la probabilidad de que se produzca un acontecimiento o suceso, y la probabilidad de que no se produzca. Esta medida se puede utilizar para valorar y estimar la magnitud de la elección de LADE. La *odds* admite valores que van desde "0", cuando "P(Y=1) = 0", hasta "∞", cuando "P(Y=1) = 1".

El logaritmo de la *odds* se conoce como "*logit*" (L): L = Ln (*odds*) = Ln [P/(1-P)]. Los posibles valores de "L" pueden oscilar entre "-∞", si "P(Y=1) = 0", y "+ ∞", cuando "P(Y=1) = 1". Los *logit*s son funciones lineales de las variables independientes. Si el *logit* "L" es una función lineal de las diferentes variables independientes, la probabilidad estimada "P(Y=1)" es una función curvilínea en forma de "*S*". La estimación sigmoidea, y por tanto no lineal, permite que las estimaciones de las probabilidades predichas se mantengan en el rango de valores comprendidos entre "0" y "1" (Figura 1).

Figura 1
FUNCIÓN DE DISTRIBUCIÓN LOGÍSTICA



Fuente: Elaboración propia

3. ESTIMACIÓN DEL MODELO DE REGRESIÓN LOGÍSTICA CON UNA SÓLA VARIABLE INDEPENDIENTE

El modelo que vamos a estimar es el siguiente:

$$Ln\left[\frac{P}{1-P}\right] = \alpha + \beta X$$

O bien:

$$Ln\left[\frac{P(Y=1)}{P(Y=0)}\right] = \alpha + \beta X$$

que en nuestro caso, para estudiar qué características de los individuos encuestados determinan la elección de la Licenciatura frente a la Diplomatura, adopta la siguiente formulación:

$$Ln\left[\frac{P(CARRERA=1)}{P(CARRERA=0)}\right] = \alpha + \beta\ DOMIC$$

Vemos, pues, que la adopción de la decisión individual de demandar unos u otros estudios, donde la variable dependiente, CARRERA, es una variable dicotómica que toma dos valores (1 si el alumno demanda LADE y 0 si demanda DCE), dependería, a priori, de una única variable independiente: DOMIC (variable explicativa que toma el valor 1 si el alumno tiene su domicilio familiar en Granada capital, y toma el valor 0 en caso contrario) (6).

El análisis descriptivo de la elección de carrera, atendiendo a la localización geográfica del domicilio familiar, se recoge en el Cuadro 1.

Cuadro 1
ELECCIÓN DE CARRERA SEGÚN DOMICILIO FAMILIAR

Titulación		*Domicilio familiar*		*Total*
		Granada DOMIC =1	*Pueblos (*) DOMIC =0*	
LADE	CARRERA=1	63	25	88
DCE	CARRERA=0	52	94	146
Total		115	119	234

(*) Incluye a los alumnos cuyo domicilio familiar está en un pueblo de la provincia de Granada, o bien en otra capital o pueblo de ésta.

Fuente: Elaboración propia

(6) Los datos disponibles de la encuesta ponen de manifiesto que la variable territorio cobra especial relevancia como determinante de la elección entre estudios de ciclo largo frente a estudios de ciclo corto (Ver Cuadro 1). En este caso asumimos que la variable explicativa es dicotómica (o binaria). Sin embargo, en el modelo de regresión logística la variable independiente (o explicativa) también puede ser politómica (más de dos categorías), o bien ser una variable cuantitativa (continua).

El modelo estimado es el siguiente (7):

$$Ln\left[\frac{P(CARRERA = 1)}{P(CARRERA = 0)}\right] = -1,3244 + 1,5163\ DOMIC$$

El hecho de que el coeficiente estimado β sea positivo ya nos indica que la probabilidad de que un estudiante demande LADE es mayor si vive en Granada capital que si vive en un pueblo. Así, la probabilidad estimada de que un alumno de Granada capital (DOMIC=1) demande LADE es 0,55:

$$Ln[P(CARRERA=1)/P(CARRERA=0)] = -1,3244 + 1,5163 \cdot 1$$

$$Ln[P(CARRERA=1)/P(CARRERA=0)] = 0,1919$$

$$P(CARRERA=1)/P(CARRERA=0) = e^{0,1919} = 1,2115$$

$$P(CARRERA=1) = 1,2115/(1+1,2115) = 0,55\ (55\%)$$

Por su parte, la probabilidad estimada de que un alumno de un pueblo demande LADE ["P(CARRERA=1) si DOMIC=0"] es de 0,21:

$$Ln[P(CARRERA=1)/P(CARRERA=0)] = -1,3244 + 1,5163 \cdot 0$$

$$Ln[P(CARRERA=1)/P(CARRERA=0)] = -1,3244$$

$$P(CARRERA=1)/P(CARRERA=0) = e^{-1,3244} = 0,2659$$

$$P(CARRERA=1) = 0,2659/(1+0,2659) = 0,21\ (21\%)$$

Las probabilidades complementarias son:

Probabilidad de que un individuo de Granada capital demande DCE:1-0,55=0,45 (45%)

Probabilidad de que un individuo de un pueblo demande DCE:1-0,21=0,79 (79%)

El cálculo directo de estas probabilidades queda recogido en el Cuadro 2.

---

(7) Los parámetros de la ecuación de regresión logística se estiman por el *Método de Máxima Verosimilitud*. El método se fundamenta en la estimación de los parámetros que maximizan la función logística para el conjunto de valores muestrales. La ecuación logística es una expresión de la probabilidad de obtener los valores observados en la muestra en función de los parámetros incluidos en el modelo.

Cuadro 2
VALORES DEL MODELO DE REGRESIÓN LOGÍSTICA CUANDO LA VARIABLE INDEPENDIENTE ES DICOTÓMICA

Variable dependiente (CARRERA)	*Variable independiente (DOMIC)*	
	DOMIC = 1	*DOMIC = 0*
CARRERA = 1	$Prob = \frac{e^{\alpha+\beta}}{1+e^{\alpha+\beta}} = \frac{e^{-1,3244+1,5163}}{1+e^{-1,3244+1,5163}} = 0,55$	$Prob = \frac{e^{\alpha}}{1+e^{\alpha}} = \frac{e^{-1,3244}}{1+e^{-1,3244}} = 0,21$
CARRERA = 0	$Prob = \frac{1}{1+e^{\alpha+\beta}} = \frac{1}{1+e^{-1,3244+1,5163}} = 0,45$	$Prob = \frac{1}{1+e^{\alpha}} = \frac{1}{1+e^{-1,3244}} = 0,79$
Total	1,00	1,00

Fuente: Elaboración propia

Podemos resumir diciendo que la variable de tipo territorial, DOMIC, pone de relieve la existencia de barreras de tipo geográfico en el acceso a estos estudios. El hecho de que el coeficiente β sea positivo ya indica que la probabilidad de que un alumno estudie la Licenciatura es mayor si vive en Granada capital que si vive en cualquier otro municipio. La probabilidad estimada de que un alumno de Granada capital haga LADE es del 55 por ciento, mientras que la probabilidad de que se matricule en dicha titulación, siendo de un pueblo de la provincia de Granada o de cualquier otra provincia (o municipio de ésta), es del 21 por ciento. La probabilidad de que un alumno, que no tiene su domicilio familiar en Granada capital, haga la Diplomatura es de un 79 por ciento.

Los alumnos de Granada capital, y también aquellos de municipios "próximos" a la capital (y que se desplazan diariamente a la Facultad), demandan principalmente los estudios de ciclo largo(8). Por su parte, los alumnos que viven "lejos" de la capital se alojan durante el curso académico en un piso de estudiantes o en un

(8) Es evidente que el sentido de "proximidad" hay que entenderlo bajo las perspectivas de coste y duración del desplazamiento, por lo que las infraestructuras de comunicaciones, la tecnología del transporte y su propia organización (horarios de autobuses, etcétera) serían variables que influyen directamente y en cada caso en la magnitud y delimitación del ámbito de "proximidad".

Los datos empíricos, y para aquellos alumnos de municipios granadinos, ponen de manifiesto que si un alumno vive en un municipio que dista menos de 30 kilómetros de Granada capital, se desplaza diariamente a la Facultad, y si vive a más de 30 kilómetros decide trasladarse a la capital durante el curso académico; aunque existe una excepción, el caso de los alumnos de Loja que, aunque viven a 54 kilómetros, se desplazan diariamente hasta el Centro educativo debido, principalmente, a la buena infraestructura viaria y al disponer de un servicio de transporte universitario.

Colegio Mayor. Estos alumnos tienen un gasto en educación mucho mayor y es por este motivo por el que deciden, principalmente, hacer una carrera más corta(9). Por tanto, es evidente que la variable territorio condiciona de forma importante la elección de unos u otros estudios, existiendo barreras geográficas importantes en el acceso a la educación superior; barreras que nos indican que a mayor distancia desde un municipio hasta la sede universitaria, más costoso es permanecer en la Universidad por lo que se opta, preferentemente, por titulaciones de ciclo corto. Así, del Cuadro 1 se desprende que el 71,6 por ciento de los alumnos de Licenciatura tienen su domicilio familiar en Granada capital, y el 64,4 por ciento de los alumnos de Diplomatura tienen su domicilio familiar en un pueblo(10).

Por su parte, la *odds* estimada de que un alumno de Granada capital demande LADE es:

$$L = \text{Ln}(odds) = -1{,}3244 + 1{,}5163 \cdot 1 = 0{,}1919$$

$$odds = e^{0\,1919}$$

La *odds* estimada de que un estudiante de un pueblo demande LADE es:

$$L = \text{Ln}(odds) = -1{,}3244 + 1{,}5163 \cdot 0 = -1{,}3244$$

$$odds = e^{-1{,}3244}$$

La razón de *odds* (*odds ratio*) estimada de demandar LADE para un estudiante de Granada capital, en comparación con un estudiante de un pueblo, se estima de la siguiente forma:

$$\text{odds ratio} = \frac{e^{0{,}1919}}{e^{-1{,}3244}} = e^{(0{,}1919 + 1{,}3244)} = e^{1{,}5163} = 4{,}5553$$

---

(9) El gasto en educación estimado fue de 393.941 pesetas/curso para aquellos individuos que se alojan en un piso de estudiantes, y de 611.087 pesetas/curso si se alojan en un Colegio Mayor o Residencia universitaria. Por su parte, el gasto en educación estimado fue de 98.648 pesetas/curso si el alumno vive en Granada capital, y de 126.102 pesetas/curso si el alumno se desplaza diariamente desde su pueblo hasta la Facultad.

El gasto privado medio en educación superior se ha calculado considerando los siguientes conceptos de gasto: matrícula, material escolar (libros de texto y otros libros), transporte urbano y, en su caso, gastos en desplazamientos al domicilio familiar, gasto en comida y gasto en alojamiento (piso de estudiantes o Colegio Mayor).

(10) En relación a este último colectivo el 67 por ciento tiene su residencia durante el curso fuera del domicilio paterno/materno, alojándose en pisos de estudiantes o en Colegios Mayores. El resto se desplaza diariamente desde el pueblo hasta la Facultad.

lo que nos indica que la *odds* estimada de que un alumno de Granada capital demande LADE es 4,55 la *odds* de que sea de un pueblo. Otra interpretación sería que los alumnos de Granada capital demandan 4,55 veces más la LADE que los alumnos procedentes de pueblos(11). Pero la razón de *odds* (*odds ratio*) es también una medida de la magnitud de la asociación entre dos variables. Una *odds ratio* mayor que 1 indica la existencia de una relación positiva o directa entre la variable dependiente e independiente, mientras que una *odds ratio* menor que 1 señala la presencia de una relación negativa o inversa. Una *odds ratio* igual a 1 es indicativo de la ausencia de relación entre las dos variables.

La *odds ratio* también puede calcularse a partir de la estimación de los parámetros del modelo. Hemos concluido que:

odds ratio = $e^{1,5163}$, donde se puede observar que el exponente (1,5163) coincide con el valor del coeficiente beta estimado del modelo de regresión. Luego:

$$\textit{odds ratio} = e^{\beta} = EXP(\beta)$$

$$\textit{odds ratio} = EXP(\beta) = e^{1,5163} = 4,5553$$

El modelo de regresión logística también permite calcular errores estándares y, consecuentemente, contrastar hipótesis y estimar intervalos de confianza en torno al valor del coeficiente beta estimado y al valor *odds ratio* estimado. El error estándar (EE) del coeficiente beta puede calcularse directamente por medio de la siguiente expresión:

$$EE(\beta) = \sqrt{\frac{1}{n_1\ p_1\ q_1} + \frac{1}{n_2\ p_2\ q_2}}$$

donde:

n_1 = número de alumnos de Granada capital (115)

n_2 = número de alumnos de pueblos (119)

p_1 = proporción de alumnos de Granada capital que demandan LADE (63/115)

p_2 = proporción de alumnos de pueblos que demandan LADE (25/119)

q_1 = proporción de alumnos de Granada capital que demandan DCE (52/115)

q_2 = proporción de alumnos de pueblos que demandan DCE (94/119)

---

(11) La *odds* estimada de que un estudiante de Granada capital demande LADE es 63/52; mientras que la *odds* estimada de que un estudiante, cuyo domicilio familiar no está en Granada capital, demande LADE es 25/94. La razón de *odds* (*odds ratio*) es: [(63/52)/(25/94)] = 4,5553 (Cuadro 1).

$$EE(\beta) = \sqrt{\frac{1}{115\ \frac{63}{115}\ \frac{52}{115}} + \frac{1}{119\ \frac{25}{119}\ \frac{94}{119}}} = 0,2928$$

Sabemos que:

$$\text{Ln}(odds) = -1,3244 + 1,5163\ DOMIC$$

El coeficiente de regresión (β = 1,5163) implica que el logaritmo de la *odds* de que un individuo demande LADE aumenta 1,5163 si es de Granada capital, que si no lo es(12). El aumento del logaritmo de la *odds* se traduce en un incremento de la probabilidad de que un individuo demande LADE si es de Granada capital, que si es de un pueblo. La pregunta siguiente sería: ¿tiene alguna significación estadística este hecho? Para comprobarlo se puede contrastar la hipótesis nula "H_0:β=0". En otras palabras, se trata de contrastar la hipótesis nula de que elegir LADE es independiente de DOMIC (de si se vive en Granada capital o en un pueblo). Para realizar el contraste se construye el denominado «Estadístico de Wald». El «Estadístico de Wald» (*W*) se calcula de la siguiente forma: β^2/EE^2, siendo β el valor del coeficiente y *EE* el error estándar del mismo. Así, para la variable *DOMIC* el «Estadístico de Wald» sería: $W = (1,5163)^2/(0,2928)^2 = 26,8181$. Este valor se compara con la distribución χ^2 con la finalidad de contrastar la hipótesis nula: "la variable independiente no explica la dependiente", frente a la hipótesis alternativa: "la variable independiente explica la dependiente". Al comparar el valor estimado de *W* con el de la tabla de valores, a un nivel de significación de 0,05 y un grado de libertad, se puede comprobar que $W > 3,84$, por lo que se rechaza la hipótesis nula y se puede afirmar que la variable independiente DOMIC incide en la probabilidad de elegir la Licenciatura(13).

---

(12) El valor de β también se puede calcular como: $\beta = [\ \text{Ln}(p_1/q_1) - \text{Ln}(p_2/q_2)]$. En nuestro caso:

$\beta = \{\text{Ln}[(63/115)/(52/115)] - \text{Ln}[(25/119)/(94/119)]\} = \text{Ln}(63/52) - \text{Ln}(25/94) = 0,1918 - (-1,3244) = 1,5163$

(13) Cuando se estudia la relación entre dos variables dicotómicas resulta menos complejo usar el análisis de las «Tablas de Contingencia 2x2», y aplicar la *prueba de la Chi-cuadrado*, para contrastar la "H_0 de independencia", o de no relación, entre ambas variables.

4. ANÁLISIS MULTIVARIABLE

4.1 Introducción

El objetivo básico de la regresión logística es intentar explicar un fenómeno que se puede clasificar de forma dicotómica (Y= {1,0}) a partir de un conjunto de variables independientes (X_i). La inclusión en la ecuación de regresión logística de dos o más variables independientes (X_1, X_2,X_3,...,X_n) constituye el denominado *modelo multivariable*:

$$Ln[P/(1-P)]= \alpha+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+....+\beta_n X_n$$

El proceso de estimación de los valores de los parámetros poblacionales por el método de la máxima verosimilitud requiere adoptar las mismas asunciones o supuestos ya anticipados cuando se hacía referencia al modelo de regresión logística con una sóla variable independiente. Así, en el caso de una variable dependiente dicotómica "Y", se asume que los valores estimados por la función logística, o valores de los *logits* "L" de "Y", se relacionan de forma lineal con las variables explicativas (X_i). En cambio, la *odds* y las probabilidades estimadas no son funciones lineales. Otras asunciones a tener en cuenta en la construcción de modelos de regresión logística multivariable son: la independencia entre observaciones y la ausencia de error en la medida de las variables.

4.2 Selección de variables independientes

La selección de las variables a incluir en el modelo como independientes se debe realizar siguiendo dos criterios:

A) *Modelización sustantiva*: el investigador decide qué variables independientes debe incluir en el modelo en función de sus hipótesis de investigación. Las variables independientes que pueden incidir, a priori, en la elección de una u otra carrera, las agrupamos en(14):

(14) Aunque el rango de valores posibles de una variable económica es usualmente un intervalo de la recta real, sin embargo la teoría econométrica considera también modelos de regresión en los que alguna de las variables explicativas toma valores en un conjunto discreto y finito. Un ejemplo de este tipo de variables lo constituye las llamadas *variables ficticias*. En este trabajo se introduce en el modelo una variable discreta para cada una de las características que se pretenden tomar en consideración. A cada una de las posibles modalidades que puede presentar la característica se le asocia un valor numérico, y la variable ficticia así definida se utiliza en la estimación del modelo como una variable explicativa más.

a) Características generales

SEXO: variable explicativa que toma el valor 1 si el alumno es hombre; toma el valor 0 si es mujer.

EDAD: variable explicativa que toma el valor 1 si el alumno tiene entre 17 y 20 años; y toma el valor 0 si tiene más de 20 años. Intentamos ver si el hecho de tener más edad es un factor determinante en la elección de una carrera más corta.

b) Características socioeconómicas

TIPOCENT: variable explicativa que toma el valor 1 si el alumno estudió en Secundaria en un Centro público; toma el valor 0 si lo hizo en un Centro privado. Esta variable nos suministra información sobre el origen socioeconómico del individuo e intentamos ver si se demanda más educación cuando se dispone de más renta.

BECA: variable explicativa que toma el valor 1 si el alumno estudia con algún tipo de beca; toma el valor 0 en caso contrario. Aquí también se intenta captar información sobre el origen socioeconómico del individuo.

PADREMPR: variable explicativa que toma el valor 1 si el padre es empresario; toma el valor 0 en caso contrario.

PADRPROF: variable explicativa que toma el valor 1 si el padre es profesional y toma el valor 0 en caso contrario.

CLASE: variable explicativa que toma el valor 1 si el alumno dice pertenecer a una clase social media-alta; toma el valor 0 en caso contrario.

c) Características territoriales

DOMIC: variable explicativa que toma el valor 1 si el alumno tiene su domicilio familiar en Granada capital, y toma el valor 0 en caso contrario.

RESIF: variable explicativa que toma el valor 1 si el universitario durante el curso académico reside con sus padres; toma el valor 0 en caso contrario. Esta variable complementa a la anterior como indicadora de costes que serán notoriamente más elevados en el caso de tener que residir fuera del domicilio paterno/materno.

d) Características familiares o de entorno

ESTPADR: variable explicativa que toma el valor 1 si el padre cuenta con un nivel de estudios de Bachillerato Superior o Universitarios; toma el valor 0 para niveles educativos inferiores.

ESTMADR: variable explicativa que toma el valor 1 si la madre tiene estudios de Bachillerato Superior o Universitarios; toma el valor 0 para niveles educativos inferiores.

e) Características académicas

REPCOU: variable explicativa que toma el valor 1 si el alumno no repite COU; toma el valor 0 en caso contrario.

NOTMED: variable explicativa que toma el valor 1 si el alumno obtuvo en Secundaria una nota media de Notable o Sobresaliente; toma el valor 0 si la calificación media fue de Suficiente o Bien.

B) *Modelización estadística*: el criterio estadístico sólo admite en el modelo aquellas variables independientes que, una vez incluidas en el mismo, tienen una capacidad de predicción estadísticamente significativa. Para conocer si estas variables independientes o explicativas son o no significativas, recurrimos al análisis de las «Tablas de Contingencia 2x2» y, utilizando el *Contraste* χ^2 *de Pearson* a un nivel de significación de 0,05, llegamos a las siguientes conclusiones:

(i) Ni el sexo ni la edad influyen en la elección de la Licenciatura frente a la Diplomatura. Estas variables no son significativas.

(ii) De las variables indicadoras de la situación socioeconómica del individuo sólo resultan significativas las variables: TIPOCENT, PADRPROF y CLASE. Ni la variable PADREMPR ni la variable BECA resultan significativas en nuestro análisis.

(iii) Los factores geográficos, medidos por las variables DOMIC y RESIF, son determinantes no sólo para que un alumno decida continuar o no sus estudios una vez finalizada la Enseñanza Secundaria, sino también para elegir, una vez tomada la decisión de ir a la Universidad, una u otra carrera.

(iv) Las variables ESTPADR y ESTMADR resultan bastante significativas en el análisis. Por tanto el nivel educativo de los padres ejerce una influencia importante en la elección de alternativa.

(v) Las características académicas, medidas por las variables REPCOU y NOTMED, no influyen en la elección de una u otra carrera.

4.3 Estimación

Nuestro modelo de demanda de educación superior obedecería, pues, a la siguiente formulación:

CARRERA = f (TIPOCENT, PADRPROF, CLASE, DOMIC, RESIF, ESTPADR, ESTMADR)

El modelo a estimar sería el siguiente:

$$\text{Ln}\,[P(\text{CARRERA}=1)/P(\text{CARRERA}=0)] = \alpha + \beta_1\,\text{TIPOCENT} + \beta_2\,\text{PADRPROF} +$$

$$+ \beta_3\,\text{CLASE} + \beta_4\,\text{DOMIC} + \beta_5\,\text{RESIF} + \beta_6\,\text{ESTPADR} + \beta_7\,\text{ESTMADR}$$

El modelo de regresión logística ajustado a los datos ponía de relieve que solamente tres variables (ESTMADR, DOMIC y PADRPROF) incidían, simultáneamente, en la probabilidad de hacer la Licenciatura en Administración y Dirección de Empresas. El resto de variables independientes no explicaban la variable depen-

diente y esto se debía, principalmente, al fenómeno de la multicolinealidad (15). El modelo de regresión logística estimado fue, consecuentemente, un modelo con sólo tres variables explicativas(16):

$$Ln\left[\frac{P(CARRERA = 1)}{P(CARRERA = 0)}\right] = \alpha + \beta_1\ ESTMADR + \beta_2\ DOMIC + \beta_3\ PADRPROF$$

Por tanto, el *ambiente cultural del hogar*, los *factores territoriales* y los *factores socioeconómicos*, son las variables que inciden, simultáneamente, en la demanda de cuatro años (LADE), frente a tres años de educación universitaria (DCE). Pero, ¿cómo inciden? (Cuadro 3).

Cuadro 3
INFLUENCIA DE LAS VARIABLES INDEPENDIENTES EN LA PROBABILIDAD DE ELEGIR LADE

Variables independientes	*Coeficiente*	*Significación estadística (1)*	*Error estándar*	*Estadístico de Wald*	*odds ratio*
Constante (2)	-1,6390	**	0,2586	40,1571	--------
ESTMADR	1,2699	**	0,3697	11,8002	3,5606
DOMIC	1,1210	**	0,3193	12,3284	3,0678
PADRPROF	0,6331	*	0,3565	3,1543	1,8835

Estadístico Chi-cuadrado = 44,761 (p = 0,0000)

Número de observaciones: 219

(1) * * Coeficientes significativos a un nivel de significación de 0,05; * Coeficientes significativos a un nivel de significación de 0,10.

(2) Individuo de referencia: estudiante de un pueblo cuyo padre es no-profesional y su madre cuenta con un nivel educativo inferior al de Bachillerato Superior o equivalente.

Fuente: Elaboración propia

(15) El fenómeno de la multicolinealidad se produce cuando existe una relación lineal entre variables independientes incluidas en el modelo, hecho que dificulta la estimación del efecto separado que cada una de las variables independientes pudiera ejercer en la predicción de la variable dependiente. Así, por ejemplo, existe correlación importante entre las variables PADRPROF y ESTPADR: PADRPROF = *f* (ESTPADR).

(16) Existen varios métodos para seleccionar el "mejor" modelo de entre varios. Un procedimiento aceptable consiste en partir de un modelo saturado (que contiene todos los posibles términos) e ir eliminando progresivamente variables explicativas. Aquella variable con el valor de **"W"** («Estadístico de Wald») más bajo y cercano al valor 0 que no sea estadísticamente significativa será excluida del modelo (*método backward o de eliminación progresiva*).

Existe el criterio de optar, de entre varios modelos igualmente aceptables, por el más sencillo, esto es, el que contenga menos términos: *criterio de parsimonia*.

La *odds ratio* para cada variable aparece recogida en la última columna. Para todas las variables la *odds ratio* es mayor que 1, lo que indica que existe una relación directa entre la variable dependiente y la independiente. En el modelo multivariable cualquier *odds ratio* que estima la relación entre una variable independiente y la variable dependiente dicotómica, está ajustada o condicionada por los valores que adoptan las otras variables independientes (covariables) incluidas en el modelo. Por ejemplo, el valor de la *odds ratio* para la variable *DOMIC* es de 3,0678, mientras que en el modelo estimado en la sección anterior, cuando sólo incluíamos una única variable independiente, era de 4,5553.

La probabilidad de que se produzca el acontecimiento o suceso definido como Y=1, vendría dada por la expresión:

$$P(CARRERA = 1) = \frac{1}{1 + e^{-(\alpha+\beta_1 ESTMADR+\beta_2 DOMIC+\beta_3 PADRPROF)}}$$

En este caso la probabilidad de elegir la Licenciatura está condicionada por el conjunto de valores que adopten las diversas variables independientes incluidas en el modelo de regresión logística (Cuadro 4).

Cuadro 4
PROBABILIDAD DE ELEGIR LADE [P(CARRERA=1)], SEGÚN DIVERSOS VALORES ADOPTADOS POR LAS VARIABLES INDEPENDIENTES

Variables independientes	*Probabilidad (%)*	*Incremento de la probabilidad*
ESTMADR=0; DOMIC=0; PADRPROF=0	16,26	------
ESTMADR=0; DOMIC=0; PADRPROF=1	26,78	10,52
ESTMADR=0; DOMIC=1; PADRPROF=0	37,33	10,55
ESTMADR=1; DOMIC=0; PADRPROF=0	40,87	3,54
ESTMADR=0; DOMIC=1; PADRPROF=1	52,87	12,00
ESTMADR=1; DOMIC=0; PADRPROF=1	56,56	3,69
ESTMADR=1; DOMIC=1; PADRPROF=0	67,96	11,40
ESTMADR=1; DOMIC=1; PADRPROF=1	80,00	12,04

Fuente: Elaboración propia

Vemos como cobran importancia en la demanda de estudios de ciclo largo, principalmente, los factores de tipo socioeconómico, la localización geográfica del lugar de residencia y el nivel cultural del hogar. Así, un estudiante cuya madre tiene un nivel educativo alto, su padre es un profesional (por ejemplo, directivo de una empresa) y vive en Granada capital, tiene un 80 por ciento de probabilidad de

demandar una carrera de ciclo largo (LADE), frente al 16,26 por ciento de probabilidad de demandar estos estudios por parte del individuo de referencia.

4.4 Contraste de hipótesis

Para el contraste de hipótesis para el valor de un coeficiente, puede usarse el «Estadístico de Wald». Los resultados del análisis, recogidos en el Cuadro 3, permiten aceptar la hipótesis alternativa, para todas las variables independientes, de que el coeficiente de regresión es diferente de cero. Comparando el «Estadístico de Wald» con una distribución chi-cuadrado (con un grado de libertad) volvemos a poner de manifiesto cómo las variables independientes explican la dependiente y, por tanto, influyen en la probabilidad de que un alumno elija la Licenciatura en Administración y Dirección de Empresas (LADE). El *Test de Wald* evalúa, por consiguiente, la significación estadística individual de cada uno de los coeficientes estimados ($\beta_1;\beta_2;\beta_3$).

Por su parte, para evaluar la significación global del modelo se utiliza el «Estadístico de la Razón de Verosimilitud» (*ERV*)(17). Si denotamos por L(MV) la función de verosimilitud para el modelo formulado, y denotamos por L(R) la función de verosimilitud para el modelo restringido en el que únicamente se considera al término independiente o constante, se define el «Estadístico de la Razón de Verosimilitud» como:

$$ERV = -2[Ln\frac{L(R)}{L(MV)}] = -2\{Ln[L(R)] - Ln[L(MV)]\} = \{-2\ Ln[L(R)]\} - \{-2\ Ln[L(MV)]\}$$

Este estadístico sigue una distribución chi-cuadrado con (k-1) grados de libertad (χ^2_{k-1}), donde k es el número de parámetros incluidos en el modelo formulado y que han sido estimados por máxima verosimilitud. En nuestro caso tenemos que:

$$-2\ Ln[L(R)] = 289{,}636$$

$$-2\ Ln[L(MV)] = 244{,}875(18)$$

$$ERV = (289{,}636) - (244,\ 875) = 44{,}761$$

El valor de este estadístico (conocido también como «Estadístico Chi-cuadrado») se utiliza para el contraste de la significación global del modelo, cuya

---

(17) Traducción del término anglosajón *Likelihood Ratio Statistic.*

(18) Si esta cantidad la dividimos por (-2) se obtiene un valor de -122,4375, que se conoce como *loglikelihood.*

hipótesis nula es que todos los coeficientes de la ecuación, excepto la constante, son nulos (H_0: $\beta_1=\beta_2=\beta_3=0$). La hipótesis nula se rechaza si el valor del «Estadístico de la Razón de Verosimilitud» excede del valor crítico. El valor de la chi-cuadrado, según tabla de valores, es de 7,81 para 3 grados de libertad y un nivel de significación de 0,05. En nuestro caso: *ERV* > 7,81. Los resultados de esta prueba permiten, pues, rechazar la hipótesis nula, H_0, de que el valor de los tres coeficientes estimados, los β's, es igual a cero.

Por tanto, ambas pruebas estadísticas (el *Contraste Chi-cuadrado* para el conjunto de coeficientes estimados, y el *Test de Wald* para cada uno de los coeficientes estimados de forma individual) permiten afirmar que la elección de carrera se relaciona, simultáneamente, con las variables *ESTMADR, DOMIC* y *PADRPROF*.

4.5 La bondad de ajuste del modelo

Una medida de la *bondad de ajuste* es un "estadístico-resumen" que indique la precisión con la cual un modelo se aproxima a los datos observados.

Para evaluar la idoneidad del modelo de regresión logística se pueden utilizar, entre otras: (i) aquellas medidas que son análogas al coeficiente de determinación múltiple (R^2) utilizado en la regresión lineal; y (ii) aquellas que estiman la bondad de ajuste mediante la comparación del número de casos, o individuos, observados con los esperados o predichos por el modelo estimado.

Dentro del primer grupo podemos calcular el «R^2-directo», o el «Pseudo-R^2 indirecto» derivado del *Test de la Razón de Verosimilitud* (prueba de la Chi-cuadrado). El método del «R^2-directo» se puede usar sólo en aquellos casos en los que la variable dependiente es binaria, y vendría dado por:

$$R^2 = 1 - \left[\frac{L(R)}{L(MV)}\right]^{\frac{2}{n}}$$

donde "n" es el tamaño muestral.

Nosotros obtenemos un R^2 igual a 0,1848. Este resultado le lleva a uno a pensar que, al estar próximo a cero, la capacidad explicativa del modelo es muy reducida. Sin embargo, Morrison (1972) argumenta que precisamente los valores de R^2 que normalmente se obtienen cuando se calculan correlaciones entre una variable dependiente binaria y las probabilidades predichas, son valores bajos, pero no implica, necesariamente, que el modelo no sea bueno.

De hecho, cuando se trabaja con modelos de respuesta cualitativa no es fácil interpretar los valores de R^2 entre 0 y 1. En estos casos el valor máximo de R^2 es mucho menor que 1, a diferencia de lo que ocurre con el modelo lineal. La razón

está en que la función de verosimilitud alcanza un máximo absoluto de 1. Así, pues, el intervalo de variación para el R^2 definido iría desde 0 hasta $1-[L(R)]^{2/n}$ [Maddala (1983)]. Nosotros obtenemos un límite superior para R^2 de 0,7335.

Una medida mejor, dentro también de este primer grupo, la constituye el denominado «Pseudo-R^2», que se define como:

$$\text{Pseudo} - R^2 = \frac{1 - [L(R) / L(MV)]^{2/n}}{1 - [L(R)]^{2/n}}$$

que en nuestro caso toma un valor de 0,2520(19).

Ninguna de las diferentes medidas de la bondad de ajuste análogas al coeficiente de determinación múltiple tienen la capacidad de explicación tan precisa como tiene tal medida en el caso de la regresión lineal, por lo que constituyen meras aproximaciones. Incluso hay autores que plantean que el uso del coeficiente de determinación como estadístico-resumen debe evitarse en aquellos modelos que contengan variables dependientes cualitativas [Aldrich y Nelson (1984)].

Introducimos, pues, un segundo grupo de medidas de la bondad de ajuste que intentan juzgar la precisión con la que el modelo se aproxima a los datos observados, comparando el número de casos observados con los predichos por el modelo estimado. El porcentaje de individuos que eligieron la alternativa predicha por el modelo puede utilizarse como indicador de la bondad del ajuste.

El porcentaje de predicciones correctas es, en nuestro caso:

$$\frac{115 + 44}{219} 100 = 72{,}60\%$$

tomando como punto de corte 0,5 (Cuadro 5)(20).

---

(19) Como medida del «Pseudo-R^2» también está la propuesta por McFadden (1974a): $R^2 = 1 - [LnL(MV)/LnL(R)]$. En nuestro caso, el «$R^2$ de McFadden» vale 0,1545.

Otras medidas de la bondad de ajuste en los modelos de respuesta cualitativa pueden encontrarse en el trabajo realizado por Amemiya (1981).

(20) El punto de corte 0,5 implica que si la probabilidad de elegir LADE estimada por el modelo es superior a 0,5 [$P(Y=1) \geq 0{,}5$], se pronostica que el estudiante estudia una carrera de ciclo largo (LADE), o que no lo hace en caso contrario, esto es, que para valores de $P < 0{,}5$ se pronostica que ese individuo estudia una carrera de ciclo corto (DCE).

El porcentaje de aciertos que se tendría con el modelo se obtiene comparando los casos en que el modelo genera una probabilidad mayor de 0,5 de elegir LADE, con los casos en que el estudiante realmente elige tal alternativa. Este porcentaje de aciertos se puede tomar como medida de la capacidad predictiva del modelo.

Cuadro 5
CLASIFICACIÓN DE LOS CASOS OBSERVADOS Y ESPERADOS PARA LA VARIABLE *CARRERA*

Casos observados	*Casos esperados o predichos*		*Total*
	DCE	LADE	
DCE	115	22	137
LADE	38	44	82
Total	153	66	219

Fuente: Elaboración propia

En base a los datos del Cuadro 5 se puede calcular la *sensibilidad del modelo* (proporción de estudiantes de LADE clasificados correctamente) y la *especificidad* del mismo (proporción de estudiantes de DCE clasificados correctamente). La sensibilidad del modelo es del 53,66% [(44/82) ·100], mientras que la especificidad es del 83,94% [(115/137) · 100].

Por otro lado, y si asignamos el mismo peso a cada caso (i = 1, ..., n), podemos definir el estadístico S:

$$S = \Sigma \frac{(y_i - p_i)^2}{p_i (1 - p_i)}$$

donde "p_i" es la probabilidad predicha por el modelo para el caso i, e "y_i" es el verdadero valor (1 ó 0).

La medida de la bondad de ajuste, como ya adelantamos antes, se basa en comparar la relación existente entre el número de casos o individuos observados y los esperados o predichos, bajo la hipótesis nula de que "el modelo seleccionado ajusta bien los datos". El estadístico S se distribuye, aproximadamente, en muestras grandes como una chi-cuadrado con (n-k) grados de libertad. Fijado un nivel de significación α, si $S < \chi^2_{(n-k),\alpha}$ entonces se acepta el modelo como bueno [Novales (1993)].

En nuestro caso tenemos un valor para S de 218,821. Este valor debe compararse con el valor según tablas de $\chi^2_{(219-4),0,05}$, y que es igual a 257,017(21). Por tanto, como $S <$ 257,017 se acepta el modelo como bueno.

---

(21) Cuando los grados de libertad son superiores a 30, Harnett y Murphy (1987) obtienen los valores críticos, para la distribución chi-cuadrado, utilizando la siguiente expresión: $\chi^2_{\alpha} = 1/2\left[Z_{\alpha} + (2 \cdot gl - 1)^{1/2}\right]^2$, donde Z_α son los valores de la Normal Estandarizada, y gl son los grados de libertad.

En nuestro caso, tendríamos que: $\chi^2_{215,\ 0,05} = 1/2\left[1,96 + (2 \cdot 215 - 1)^{1/2}\right]^2 = 257,017$

5. CONCLUSIONES

Son muchos los fenómenos de interés en el campo económico en los cuales la variable dependiente depende de la elección de los individuos en un conjunto formado por un número finito de alternativas. Si la elección abarca solamente dos alternativas, tendremos un *modelo de elección dicotómica* (por ejemplo: utilizar transporte público o privado; elegir una carrera de ciclo largo o una de ciclo corto; etcétera). En estas situaciones, en las que la variable respuesta es dicotómica, la regresión "clásica" (que trata de explicar el nivel de una variable respuesta continua en función de un conjunto de variables explicativas) no es el método más adecuado, ya que sus propiedades óptimas están basadas en unos supuestos que dejan de cumplirse cuando la variable respuesta es cualitativa [Teijeiro (1991)]. El *modelo de regresión logística* permite, consecuentemente, relacionar una variable dependiente dicotómica con una o más variables independientes, las cuales pueden ser dicotómicas, politómicas o continuas.

Los resultados del modelo de regresión logística utilizado en este trabajo ponen de manifiesto la existencia de barreras económicas, geográficas y culturales, para la Universidad de Granada, en la elección de estudios de ciclo largo (LADE), frente a estudios de ciclo corto (DCE).

En primer lugar, existe una asociación significativa entre el nivel de renta familiar y la elección de alternativa. La variable explicativa PADRPROF se ha introducido en el modelo como *proxy* del nivel de renta familiar, porque la encuesta no nos facilita información directa de los ingresos del cabeza de familia. El coeficiente estimado asociado a esta variable es positivo y estadísticamente significativo, por lo que podríamos afirmar que a medida que aumenta el nivel de renta del estudiante, también aumenta la probabilidad de demandar estudios de ciclo largo frente a estudios de ciclo corto(22).

En segundo lugar, la evidencia empírica nos revela cómo la distancia a la sede universitaria está inversamente relacionada con la escolarización universitaria, siendo los alumnos de Granada capital los que demandan, principalmente, cuatro años de educación superior, mientras que los alumnos de los pueblos demandan, preferentemente, tres años de estudios universitarios. Por tanto, la localización geográfica del domicilio familiar (variable explicativa DOMIC) influye significativamente en la elección de alternativa. El hecho de que el coeficiente estimado asociado a esta variable sea positivo, ya indica que la probabilidad de que un alumno estudie la Licenciatura es mayor si vive en Granada capital que si vive en cualquier otro municipio. La *odds ratio* igual a 3,07 nos indica que los alumnos de Granada

---

(22) Estudios recientes realizados en nuestro país también ponen de manifiesto cómo existe una asociación significativa entre el tipo de estudios universitarios que se cursan y los deciles de renta. Así, las rentas bajas están más asociadas a estudios de grado medio en Escuelas Universitarias [Dávila y González (1994)].

capital multiplican por 3,07 la probabilidad de hacer la Licenciatura con respecto a la Diplomatura. Existen, pues, barreras geográficas en el acceso a la educación superior, ya que a mayor distancia desde un municipio hasta la sede universitaria, más costoso (en tiempo y en dinero) es permanecer en la Universidad por lo que se opta, preferentemente, por titulaciones de ciclo corto.

En tercer y último lugar, el análisis de la variable ESTMADR -indicadora del nivel cultural del hogar- nos permite afirmar que el nivel de formación de los padres también incide en la elección de una u otra carrera. La probabilidad estimada de que un individuo estudie la Licenciatura se ve multiplicada por 3,56 (*odds ratio*) si su madre tiene estudios de Bachillerato Superior o Universitarios, frente a niveles educativos inferiores. Son, pues, las personas cuyos padres tienen un mayor nivel educativo las que tienen también más probabilidad de completar un mayor nivel de educación, ejerciendo de este modo las transferencias de capital humano de padres a hijos una influencia decisiva en la elección de estudios(23).

Un análisis más exhaustivo de la información estadística disponible permite establecer algunas consideraciones acerca del mecanismo en virtud del cual el nivel educativo de los padres influye en el de sus hijos. Por un lado, hay una influencia directa de los intereses de los padres a través del denominado «efecto imitación»(24) y, por otro lado, la influencia también es indirecta causada por el ambiente cultural del hogar(25).

En resumen, podemos concluir diciendo que en la demanda de estudios universitarios cobran importancia, principalmente, los factores de tipo socioeconómico, la localización geográfica del domicilio familiar y el nivel educativo de los padres; todos ellos son factores que nos ayudan a explicar los diferentes valores que en cada familia se otorgan a la educación.

---

(23) En el estudio realizado en nuestro país por Mora (1996), se analiza la influencia positiva que tienen los factores culturales familiares para la consecución de niveles más altos de estudios. Los niveles educativos del sustentador principal y del cónyuge son las principales variables que explican que un joven adquiera estudios postobligatorios. Así, la probabilidad de acceder a la educación postobligatoria (secundaria y universitaria), teniendo padres universitarios o con estudios secundarios frente a tener padres sin estudios, se ve multiplicada por valores entre 2,09 y 3,06. En general, el nivel educativo del cónyuge parece más relevante para determinar que los hijos sigan estudios postobligatorios.

(24) El «efecto imitación» es la tendencia de los hijos a preferir el tipo general de ocupación de sus padres. El 70 por ciento de los alumnos cuyo padre es empresario, manifiesta una preferencia por trabajar en la empresa privada o instalarse por su cuenta.

(25) El porcentaje de alumnos de LADE cuyos padres (ambos) tienen estudios de Bachillerato Superior o Universitarios es del 31,8 por ciento. Este porcentaje sólo es del 9,5 por ciento en el colectivo de estudiantes de DCE.

REFERENCIAS

ALDRICH, J. Y NELSON, F. (1984). «Linear Probability, Logit and Probit Models», *Sage Publicaciones*. Beverly Hills, California

AMEMIYA, T. (1981). «Qualitative Response Models: A Survey». *Journal of Economic Literature*, 19, 1483-1536.

AMEMIYA, T. (1994). «Qualitative Response Models. Part IV». En *Studies in Econometric Theory. The Collected Essays of Takeshi Amemiya*, Edward Elgar. USA.

BOSKIN, M.J. (1974). «A Conditional Logit Model of Occupational Choice». *Journal of Political Economy*, 82, 389-398.

DÁVILA, D. Y GONZÁLEZ, B. (1994). «Renta y acceso a la educación superior en España» Ponencia presentada en las *III Jornadas de la Asociación de Economía de la Educación*, Barcelona, noviembre.

DOMENCICH, T. Y MCFADDEN D.L. (1975). «Urban Travel Demand: A Behavioral Analysis», *North-Holland*. Amsterdam

HARNETT, D.L. Y MURPHY, J.L. (1987). «Introducción al análisis estadístico», *Addison-Wesley Iberoamericana*. México

HOSMER, D.W. Y LEMESHOW, S. (1989). «Applied Logistic Regression», *Wiley Publications*. Nueva York

JOVELL, A.J. (1995). «Análisis de regresión logística», *Cuadernos Metodológicos del CIS*. Madrid

LASSIBILLE, G. Y NAVARRO, L. (1981). «Tratamiento econométrico de las variables cualitativas». *Cuadernos de Ciencias Económicas y Empresariales*, 8, 49-92.

LI, M.M. (1977). «A Logit Model of Home Ownership». *Econometrica*, 45, 1081-1098.

MADDALA, G.S. (1983). «Limited-Dependent and Qualitative Variables in Econometrics», *Cambridge University Press*. Cambridge.

MCFADDEN, D.L. (1974a). «The Measurement of Urban Travel Demand». *Journal of Public Economics*, 3, 303-328.

MCFADDEN, D.L. (1974b). «Conditional Logit Analysis of Qualitative Choice Behaviour». En Zarembka, P. (ed.). *Frontiers in Econometrics*, *Academic Press*. Nueva York.

MORA, J.G. (1989).«La demanda de educación superior. Un modelo analítico», *Consejo de Universidades*. Madrid

MORA, J.G. (1996). «Influencia del origen familiar en el acceso a la educación, en la obtención de empleo y en los salarios». En Grao, J. e Ipiña, A. (eds.). *Economía de la Educación. Temas de estudio e investigación*, Servicio Central de Publicaciones del Gobierno Vasco. Vitoria.

MORRISON, D.G. (1972). «Upper Bounds for Correlations Between Binary Outcomes and Probabilistic Predictions». *Journal of the American Statistical Association*, 7, 68-70.

NOVALES, A. (1993). «Econometría», *McGraw-Hill*. Madrid

RUIZ-MAYA, L. (1990). «Metodología estadística para el análisis de datos cualitativos», *Centro de Investigaciones Sociológicas*. Madrid

SCHMIDT, P. Y STRAUS, R.P. (1975). «The Predictions of Occupation Using Multiple Logit Models». *International Economic Review*, 16 (2), 471-486.

TEIJEIRO, E. (1991). «Algunas técnicas multivariantes útiles para la presentación de los resultados de una encuesta». *Estadística Española*, 127, 305-324.

LOGISTIC REGRESSION. AN APPLICATION TO THE DEMAND FOR HIGHER EDUCATION

SUMMARY

Logistic Regression is included in the group of "Data Analysis Statistical Techniques", and it is used when we are interested in relating a qualitative dependent variable with some independent variables. This paper analyzes, using the Logistic Regression Method, what factors determine how students decide on being at University for three or four years.

Key words: Odds Ratio. Wald Statistic. Logistic Regression.

AMS Classification: 62J99 62P20 90A99