ELECCIÓN DE UNA ASIGNATURA OPTATIVA: UNA APLICACIÓN DEL ANÁLISIS DISCRIMINANTE Y LA REGRESIÓN LOGÍSTICA

Olga Valencia García
Departamento de Economía
Universidad de Burgos

RESUMEN

Los actuales planes de estudios de la enseñanza universitaria ofrecen a los alumnos la posibilidad de elegir entre un número considerable de asignaturas optativas, con el objeto de alcanzar un mayor grado de especialización en su formación. Para realizar una mejor planificación de la oferta de este tipo de asignaturas, es interesante para los centros universitarios contar con la mayor cantidad de información posible tanto del nivel previsible de demanda, como del perfil de los alumnos que optarían por cada materia.

En este trabajo se ha realizado un ensayo sobre la posibilidad de predecir el grado de aceptación de una asignatura concreta. En particular, se ha utilizado una asignatura optativa del área de Estadística, incluida en el plan de estudios de una diplomatura. La información, suministrada por los propios alumnos, ha sido tratada mediante análisis discriminante, técnica frecuentemente utilizada en investigación comercial para la previsión y caracterización de la demanda de un producto. De esta forma, se ha pretendido detectar las variables relevantes en la elección de esta materia y determinar, en función de ellas, las posibilidades de que un alumno la escoja. Alternativamente, se ha probado la regresión logística como método para conseguir el mismo objetivo.

1. INTRODUCCIÓN

La diversidad de asignaturas optativas de los planes de estudios universitarios confieren a los alumnos cierta capacidad de elección. Frecuentemente la decisión de matricularse en una materia no responde a cuestiones objetivas como la orientación personal y/o profesional, sino que además pueden intervenir otros aspectos de distinta naturaleza. A menudo los profesores se preguntan cuáles son las motivaciones reales de los alumnos que han optado por la asignatura que imparten. Tanto para ellos, como para los responsables de los centros universitarios, sería interesante conocer estas motivaciones y con ellas, el perfil de los alumnos que escogen determinadas materias, así como poder realizar una previsión de los alumnos que se decantarán por una disciplina concreta.

En este trabajo se ha realizado una prueba, un pequeño ensayo sobre esta cuestión, tomando la asignatura optativa de Estadística II, que se ofrece en el plan de estudios de la Diplomatura en Relaciones Laborales de esta universidad, junto con otras disciplinas de contendidos muy diferentes.

En primer lugar se presenta la metodología del estudio, con una breve alusión a la matriz de información, y algunas explicaciones sobre las técnicas estadísticas empleadas: el análisis discriminante y la regresión logística. Posteriormente se detallan los principales resultados de los análisis y su interpretación.

2. METODOLOGÍA

2.1 LA MATRIZ DE INFORMACIÓN

Los datos han sido proporcionados por los alumnos a través de la respuesta a un cuestionario en el que se incluyeron 17 variables que, en principio, parecían tener algún tipo de influencia en la decisión. La selección de estas variables se realizó teniendo en cuenta las opiniones de un grupo de alumnos, con los que se mantuvo una reunión previa a la elaboración del cuestionario, con el objetivo de recoger sus propuestas. Entre estas variables, hay algunas relacionadas con características personales de los alumnos,

otras relativas a la valoración previa que hacen de la asignatura cuando se plantean su elección, y un tercer tipo que se refiere a cuestiones prácticas para poder cursarla.

La muestra con la que se ha trabajado, compuesta por 78 individuos, no se ha extraído con criterios aleatorios, sino que se trata de una muestra de conveniencia en la que se incluyen alumnos que han elegido y que no han elegido la asignatura, (los dos grupos considerados en el análisis), pero que no representa correctamente a todos ellos. Nuestro propósito no ha sido llegar a conclusiones definitivas, sino realizar un primer ensayo para probar el método, ensayo que tendrá que ser mejorado con un muestreo más adecuado.

2.2 TÉCNICAS ESTADÍSTICAS UTILIZADAS

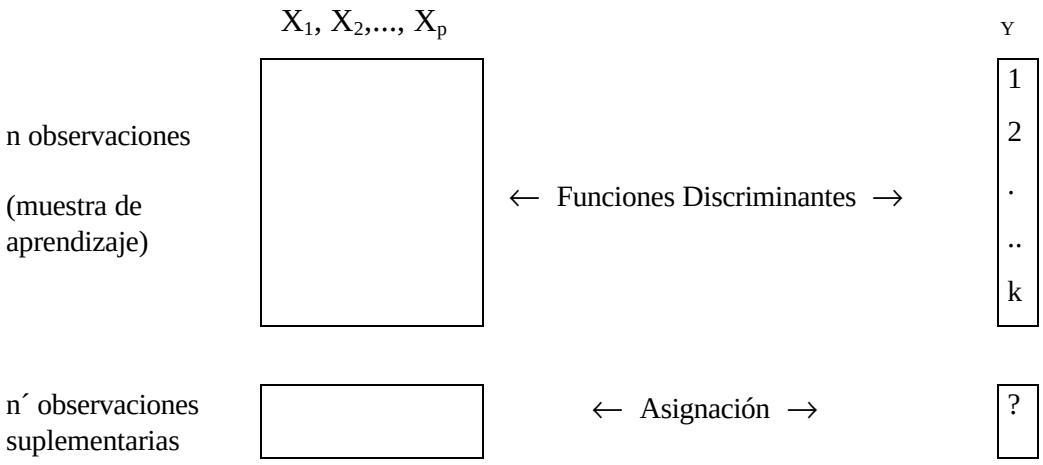
2.2.1 ANÁLISIS FACTORIAL DISCRIMINANTE (AFD)

Dada una variable dependiente cualitativa y un conjunto de variables independientes, cuantitativas o cualitativas, el Análisis Discriminante es una técnica que permite obtener funciones lineales de las variables independientes, a partir de las cuáles se puede clasificar a los individuos en uno de los grupos o clases establecidas por los valores de la variable dependiente.

Si tenemos “n” individuos descritos por un conjunto de “p” variables (X_1, X_2,..., X_p) y repartidos en “k” grupos definidos a priori por una variable nominal con “k” modalidades, el objeto del análisis discriminante se puede explicar en dos pasos sucesivos, uno de naturaleza descriptiva y otro de tipo predictivo:

1. Como otros métodos factoriales, el Análisis Factorial Discriminante (AFD) da lugar al cálculo de ejes principales. Estos ejes son las combinaciones lineales de las variables explicativas (X_1, X_2,..., X_p) que mejor separan los k grupos, y se denominan funciones discriminantes.
2. Una vez obtenidas estas funciones y las puntuaciones de los individuos en ellas, el análisis permite clasificar los individuos en uno de los grupos preexistentes, a través de distintos criterios que tratan de minimizar los errores de clasificación.

Algunos autores, como Lebart y otros (1995), sugieren la utilización de una muestra de aprendizaje mediante la que se elaboren las funciones discriminantes, y otra muestra de individuos nuevos o suplementarios, que son asignados a los grupos según su puntuación en estas funciones, y que permiten valorar la eficacia de estas funciones.



Funciones discriminantes

A partir de la matriz de información de "n x p" datos, en la que cada individuo puede considerarse como un punto en un espacio p-dimensional, el AFD pretende extraer un nuevo espacio de dimensión inferior tal que, al proyectar la nube de puntos sobre dicho espacio, los puntos-individuo del mismo grupo estén lo más próximos posible entre sí, y lo más alejados posible de los individuos de otros grupos. Los ejes de este nuevo espacio se denominan funciones discriminantes.

Basándose en la descomposición de la varianza total en varianza dentro de clases (intra) y varianza entre clases (inter), el criterio para extraer este espacio es encontrar combinaciones lineales de las variables explicativas, ortogonales entre sí, que maximicen el cociente entre la varianza inter y la varianza intra (o de forma equivalente, entre la varianza inter y la varianza total). Esto supone diagonalizar la matriz T^{-1} x E, donde T representa la matriz de covarianzas total, sin distinguir grupos, y E la matriz de covarianza entre grupos.

Reglas de asignación de los individuos

Una vez determinadas las funciones discriminantes, es necesario contar con un criterio que permita asignar a un grupo a los nuevos individuos descritos por las variables explicativas. El objetivo es por tanto, encontrar una regla de clasificación que minimice los errores de asignación que se pueden cometer. Para elaborar esta regla de clasificación se pueden utilizar elementos de la inferencia paramétrica si se conocen las distribuciones de las poblaciones, o bien otro tipo de procedimientos no paramétricos.

Una regla simple y geométrica de asignación es elegir la clase a la cuál el centro de gravedad es el más próximo al punto-individuo. La métrica generalmente utilizada en estas aplicaciones es la Distancia de Mahalanobis. Sin embargo, esta aproximación es puramente geométrica y no tiene en cuenta las probabilidades a priori de las diferentes clases. Utilizando la Teoría de Bayes, que tiene en cuenta esas probabilidades a priori, es posible obtener una regla que permite clasificar a los individuos en cada uno de los "k" grupos, a partir de sus puntuaciones discriminantes (valores obtenidos en la función discriminante). La probabilidad de que un individuo i, con puntuaciones discriminantes $D(d_{i1}, d_{i2}...)$, pertenezca a un grupo j, se denota por $P(G_j/D)$ y se obtiene mediante la fórmula de Bayes:

$$p\ (Gj / D) = \frac{p(D / G_j)p(G_j)}{\sum_{j=1}^{k} p(D / G_j)p(G_j)}$$

$P(G_j)$ es la probabilidad a priori de que un individuo pertenezca al grupo "j", es decir, la probabilidad que le corresponde si no se dispone de ningún tipo de información previa sobre el mismo. Estas probabilidades se asignan en función del tamaño de los grupos en la muestra o bien se parte de probabilidades idénticas para todos los grupos.

$P(D/G_j)$ son las probabilidades condicionadas. Si las puntuaciones discriminantes están normalmente distribuidas para cada uno de los grupos y podemos estimar los parámetros de los mismos, también es posible calcular la probabilidad de obtener una determinada puntuación discriminante bajo el supuesto de pertenencia a un grupo.

$P(G_j/D)$ son las probabilidades a posteriori, es decir, las probabilidades de pertenencia a un grupo, contando con la información proporcionada por su puntuación discriminante.

Lógicamente, un individuo será asignado al grupo para el que la probabilidad a posteriori sea máxima, es decir, clasificado en Gj si:

$P(G_j/D) = máx \{P(G_1/D), P(G_2/D),...,P(G_k/D)\}$

2.2.2. REGRESIÓN LOGÍSTICA (RL)

La regresión logística es una herramienta alternativa para conseguir los mismos objetivos expuestos: predecir si un suceso ocurrirá o no, así como identificar las variables útiles en hacer tal predicción.[1] Su utilización es adecuada cuando se tiene una variable dependiente con únicamente dos categorías, es decir, dicotómica, y se pretende averiguar la probabilidad de ocurrencia de una u otra a partir de una serie de variables independientes, que pueden ser cuantitativas o cualitativas. Así, la probabilidad de que un individuo pertenezca a la segunda categoría, se formula del modo siguiente:

$$p = 1 / (1+ e^{-z})$$

siendo Z una combinación lineal de las "p" variables independientes:

$Z = B_0 + B_1X_1 + B_2X_2 + ...+ B_pX_p$

Lógicamente, la probabilidad de pertenecer a la primera categoría será q=1-p.

De acuerdo con su carácter de probabilidad, la función anterior sólo adopta valores comprendidos entre 0 y 1, y es expresión de una relación no lineal. El criterio utilizado para la estimación de los parámetros es el de máxima verosimilitud, y para su aplicación se emplean métodos de cálculos iterativos.

Una expresión alternativa para el modelo de regresión logística es:

$$p/q = e^{Bo} \ (e^{B1}) \ X_1 ... \ e(^{Bp}) \ X_p$$

[1] En realidad, el análisis discriminante ha de cumplir una serie de supuestos como por ejemplo la distribución normal multivariada de las variables predictoras. La regresión logística se basa en supuestos menos rigurosos, ya que no requiere la normalidad mutivariante aunque sus soluciones son más estables si esto se cumple.

lo que supone que cuanto mayor sea el coeficiente B_i de una variable, mayor es el cociente entre la probabilidad de pertenecer al segundo grupo, respecto a la de pertenecer al primero

La clasificación de los individuos en uno u otro grupo (o categoría), se realiza a partir de la probabilidad estimada de pertenecer al segundo grupo. Si para un individuo esa probabilidad es igual o mayor que un punto de corte predeterminado, habitualmente 0,5, será clasificado en dicho grupo, y en caso contrario, se asignará al primer grupo. El porcentaje de casos correctamente clasificados será un índice de la efectividad del modelo.

3. RESULTADOS E INTERPRETACIÓN DEL AFD

La aplicación del análisis a la matriz de información proporciona en primer lugar, las variables que contribuyen en mayor grado a discriminar a los alumnos entre matriculados y no matriculados en la asignatura, la función discriminante, obtenida como combinación lineal de estas variables, y una serie de medidas para valorar su eficacia. En segundo lugar, se presentan los resultados de la clasificación efectuada sobre los individuos, según sus puntuaciones discriminantes. La comparación entre el grupo real de pertenencia y el pronosticado permite completar la evaluación del poder discriminante de la función.

Obtención y Evaluación de la FUNCIÓN DISCRIMINANTE

De los 78 alumnos que han contestado el cuestionario, 14 han sido excluidos de esta primera parte del análisis debido a que presentan valores ausentes en alguna variable discriminante[2], por lo que quedan 64 casos considerados válidos.

Una vez calculadas las medias y desviaciones típicas de cada variable para el total de los individuos y para cada grupo por separado, se realizan los contrastes de comparación de medias entre los grupos, para detectar aquellas características en las que se obtienen promedios significativamente distintos. Los valores del estadístico F, que corresponden al análisis de la varianza de un factor para cada una de las variables independientes consideradas, así como los niveles de significación de este estadístico, aparecen en la Tabla 1.

Como puede apreciarse, tan sólo se encuentran diferencias significativas a un nivel del 5%, en 6 de las 16 variables inicialmente consideradas, en las que se registran además, los valores más bajos de la λ de Wilks.

La selección de las variables para configurar la función discriminante, única en este caso al existir sólo dos grupos, se ha realizado a través de un procedimiento por pasos, que combina la posibilidad de ir introduciendo y eliminando variables del modelo, según su nivel de aproximación al criterio de selección que se establezca. El criterio utilizado en este caso es el de la λ de Wilks, que mide las desviaciones dentro de cada grupo respecto a las desviaciones totales sin distinguir grupos. De esta forma, en cada paso se selecciona la variable para la que, junto con las variables previamente seleccionadas, el valor de la λ de Wilks sea mínimo, lo que implica una buena discriminación.

[2] La falta de datos se produce fundamentalmente en la variable "Adecuación de las fechas de exámenes de la asignatura al horario de clases o trabajo", ya que muchos no recuerdan haber tenido en cuenta esta cuestión.

TABLA 1

Pruebas de igualdad de las medias de los grupos

	Lambda de Wilks	F	gl1	gl2	Sig.
EDAD	,890	7,666	1	62	,007
ESTADISTICA I SI/NO	,895	7,303	1	62	,009
ADECUACION FECHA EXAMEN	,993	,420	1	62	,519
FORMACIÓN MATEMÁTICA PREVIA	1,000	,000	1	62	,998
ADECUACION HORARIO	,927	4,918	1	62	,030
HORAS ESTUDIO DIARIAS	,992	,518	1	62	,475
CONOCIMIENTO PROFESOR	,996	,219	1	62	,641
VALORACION RENDIMIENTO	,971	1,865	1	62	,177
TRABAJA SI O NO	,953	3,044	1	62	,086
VALORACION UTILIDAD	,976	1,528	1	62	,221
OTROS ESTUDIOS AFINES	,793	16,174	1	62	,000
DIFICULTAD ESTADISTICA I	,995	,337	1	62	,564
VALORACION DIFICULTAD	,831	12,616	1	62	,001
AFINIDAD CON GUSTOS	,728	23,165	1	62	,000
AREA PROFESIONAL PREFERIDA	,939	3,998	1	62	,050
% DE ASISTENCIA A CLASE	,999	,050	1	62	,824

Ahora bien, el hecho de que la λ de Wilks tome un valor mínimo no implica que éste sea suficientemente pequeño, por lo que hay que analizar si la variable a la que corresponde ese valor mínimo, y que por ello es la candidata a ser seleccionada, aporta una información significativa. En este sentido, el criterio se completa con el estadístico F de entrada, que evalúa la disminución que se produciría en la λ de Wilks si la variable correspondiente fuera seleccionada. Así, una disminución significativa supone una

selección adecuada. Paralelamente, respecto a las variables incluidas, un estadístico F de salida valora el incremento que se produciría en la λ de Wilks si la variable correspondiente fuera eliminada. De este modo, un incremento no significativo pone de manifiesto que la información que se pierde es escasa, y por tanto, que la variable en cuestión puede ser eliminada.

Adicionalmente, este método de selección por pasos permite evitar el riesgo de que alguna de las variables independientes sea una combinación lineal de las otras, y se obtengan estimaciones no fiables de los coeficientes de la función. Para ello se calcula un valor de Tolerancia de cada variable X_j respecto al resto de variables $X_1,...X_{j-1}$, X_{j+1}, $...X_p$, que se define como el complemento a 1 del cuadrado del coeficiente de correlación múltiple entre X_j y las restantes variables. Con ello, para que una variable candidata a ser seleccionada en un paso, pueda serlo, la tolerancia con las variables incluidas en la ecuación debe ser superior a un cierto valor mínimo.

Finalmente, señalar que para eliminar la posibilidad de que el método de selección por pasos se convierta en un proceso cíclico, se establece un límite para el número de pasos, normalmente el doble del número de variables independientes.

En nuestro caso, el proceso se ha detenido en el sexto paso, dado que ninguna variable fuera de la ecuación cumplía el requisito de entrada, y ninguna variable introducida cumplía el requisito de salida. El programa ha introducido una variable en cada paso, y no ha eliminado ninguna. Así se han seleccionado 6 variables para conformar la función discriminante que, por orden de inclusión, son las siguientes: Afinidad con los gustos, Posibilidad de continuar estudios universitarios afines, Adecuación del horario de la asignatura al horario de clases o trabajo, Valoración previa de la dificultad de la asignatura, Grado de dificultad de la asignatura llave de Estadística I, y Edad.

Una vez seleccionadas las variables que se incluyen en la función discriminante, se calculan las puntuaciones discriminantes de los individuos, sin más que sustituir en dicha función los valores observados en esas variables. A partir de las puntuaciones

discriminantes, existen una serie de estadísticos que permiten valorar la eficacia de la función (Tabla 2):

1. El autovalor mide las desviaciones de las puntuaciones discriminantes entre los grupos respecto a las desviaciones dentro de los grupos. En consecuencia cuanto mayor sea este valor, mejor será la función discriminante.[3]

2. La correlación canónica cuantifica las desviaciones de las puntuaciones discriminantes entre los grupos respecto a las desviaciones totales sin distinguir grupos (proporción de varianza atribuible a la diferencia entre los grupos). Es una medida de asociación entre las puntuaciones discriminantes y los grupos, con un valor, 0.763, bastante elevado en este caso.

3. La λ de Wilks para las puntuaciones discriminantes representa la proporción de varianza total de las puntuaciones discriminantes no explicada por las diferencias entre grupos (cociente entre desviaciones dentro de grupos y desviaciones totales sin distinguir grupos). Como se deduce de la tabla, un 58,2% de la varianza es explicada por las diferencias entre grupos.

A partir de la λ de Wilks, se construye un estadístico que sigue una distribución χ^2 , con el que se contrasta la hipótesis de que los grupos procedan de una población en la que no existan diferencias entre las puntuaciones discriminantes.[4] En nuestro caso rechazamos este supuesto, ya que el valor de χ^2 es elevado y las diferencias entre los grupos son estadísticamente significativas.

4. Los coeficientes estandarizados correspondientes a cada variable incluida en la función discriminante, se calculan a partir de las variables independientes tipificadas. Eso

[3] Teniendo en cuenta que el autovalor asociado a una función se interpreta como la parte de la variabilidad total de la nube proyectada sobre el conjunto de todas las funciones, atribuible a esa función concreta, y que en este caso al existir dos grupos sólo hay una función, ésta acumula el 100% de la varianza.

[4] $\chi^2_{K(G-1)} = -[(n-1)- (K+G)/2] \text{ Ln } (\lambda \text{ de Wilks})$, siendo "n" el número de individuos, "G" el número de grupos, y "K" el número de variables discriminantes.

hace que la función tenga un término independiente nulo, y que los coeficientes puedan ser comparados entre sí, lo que permite deducir en este caso, que la "Afinidad" con la asignatura es la que más influye en la discriminación, seguida por la predisposición a continuar con "Otros estudios afines", y así podríamos establecer un orden hasta la variable "Edad", la que menos discrimina entre los alumnos, dentro de las variables discriminantes.

5. Se recogen también los promedios de cada grupo en la función, lo que corresponde a las medias de las puntuaciones discriminantes para cada uno de los grupos. Dado que estas medias no son muy similares, la función sirve para efectuar la discriminación, como ya ha confirmado el estadístico χ^2

TABLA 2

Autovalores

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	1,395[a]	100,0	100,0	,763

a. Se han empleado las 1 primeras funciones discriminantes canónicas en el análisis.

Lambda de Wilks

Contraste de las	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	,418	51,529	6	,000

Coeficientes estandarizados de las funciones discriminantes canónicas

	Función
	1
EDAD	-,357
ADECUACION HORARIO	,365
OTROS ESTUDIOS AFINES	-,527
DIFICULTAD ESTADISTICA I	,434
VALORACION DIFICULTAD	-,490
AFINIDAD CON GUSTOS	,704

Funciones en los centroides de los grupos

MATRICULADO EN ESTII SI/NO	Función
	1
1	,643
2	-2,101

Funciones discriminantes canónicas no tipificadas evaluadas en las medias de los grupos

Resultados de la CLASIFICACIÓN

En consonancia con el segundo objetivo del Análisis Discriminante, se comentan los resultados de la clasificación efectuada conforme a las puntuaciones discriminantes de los individuos. Como se ha mencionado, el criterio de clasificación utilizado es el de maximizar las probabilidades a posteriori, basado en la teoría de Bayes.

En el proceso de clasificación, no se ha excluido ningún caso por lo que los individuos procesados han sido 78. De ellos, 7 individuos presentaban valores perdidos en una de las variables predictoras, pero los valores ausentes han sido reemplazados por la media, por lo que la clasificación se ha realizado sobre el total de los elementos de la muestra.

Una vez calculadas las probabilidades condicionadas y las probabilidades a posteriori para cada individuo, y comparando el grupo real de pertenencia con el grupo pronosticado por el análisis, el proceso de clasificación se resume en la Tabla 3:

TABLA 3

Resultados de la clasificación[a]

		MATRICULADO EN ESTII SI/NO	Grupo de pertenencia pronosticado		Total
			1	2	
Original	Recuento	1	48	3	51
		2	7	20	27
	%	1	94,1	5,9	100,0
		2	25,9	74,1	100,0

a. Clasificados correctamente el 87,2% de los casos agrupados originales.

Como se aprecia, el análisis clasifica correctamente 68 de los 78 casos considerados, lo que supone un porcentaje de acierto del 87,2%. Entre los alumnos que eligen la asignatura, el porcentaje de clasificados correctamente es superior (94,1%) al que se obtiene para el grupo que no la elige (74,1%).

Hay que tener en cuenta que la clasificación está muy determinada por las probabilidades a priori. En nuestro caso, las probabilidades previas para los grupos se han asignado en función del tamaño de éstos en la muestra, lo que significa que al primer grupo, -los que eligen la asignatura-, le corresponde una probabilidad a priori del 76,6 %, frente a un 23,4% para los no matriculados. En este sentido, conviene señalar que la muestra obtenida no es representativa dado que no se ha extraído con procedimientos aleatorios, sino que se trata de una muestra de conveniencia, en la que los alumnos matriculados en Estadística II aparecen “sobre-representados” y lo contrario puede decirse respecto al otro grupo. Este hecho se ha visto reforzado por la mayor proporción de “no respuesta” en algunas variables, para el segundo grupo, lo que disminuye su presencia en el análisis. El resultado de la clasificación se hubiera visto posiblemente muy modificado trabajando con una muestra con representación proporcional de los grupos, o bien asignando probabilidades previas idénticas.

4. RESULTADOS E INTERPRETACIÓN DE LA RL

Al igual que en el caso anterior, se han rechazado 14 casos por valores ausentes en alguna variable, con los el número de casos incluidos en el análisis es 64. De las 16 variables predictoras, tres de ellas son tratadas como categóricas: Área en que desearía desarrollarse profesionalmente, Trabaja (Sí o No), y Aprobada la asignatura llave de Estadística I (Sí o No), por lo que han sido recodificadas por el sistema. Interesa indicar que la codificación interna de la variable dependiente se realiza como "0"=SÍ se matricula, y "1"=NO se matricula en la asignatura de Estadística II, ya que esto condiciona la interpretación de los resultados.

Para elegir las variables del modelo de Regresión Logística, se ha utilizado también un método de selección por pasos, que permite introducir las variables que más información aportan a las probabilidades de pertenecer a cualquiera de los dos grupos establecidos (matriculados y no matriculados en Estadística II), admitiendo también la posibilidad de eliminar variables previamente seleccionadas. En particular, el procedimiento escogido, Forward Wald, emplea dos estadísticos: la puntuación eficiente de Rao para la selección, y el estadístico de Wald, para la eliminación de variables. El estadístico de Wald se calcula para cualquier variable independiente seleccionada en un paso y sirve para contrastar la hipótesis de que el parámetro asociado a esa variable en el modelo, es estadísticamente no nulo[5]. Si no es así, la variable en cuestión será posteriormente eliminada. La puntuación eficiente de Rao evalúa la significación de una variable no incluida en un paso, en el caso de que fuera seleccionada en el paso siguiente.

En cada uno de los pasos del proceso de selección, se ofrecen una serie de medidas para contrastar la bondad de ajuste del modelo de regresión logística:

- Estadístico $-2LL_o$, donde L es el logaritmo neperiano y L_o la verosimilitud. Teniendo en cuenta que un buen modelo es aquel en el que la probabilidad de los resultados

[5] El estadístico de Wald para cada variable es el cociente entre el cuadrado del coeficiente correspondiente a esta variable en el modelo, y el cuadrado del error estándar de dicho coeficiente. Se distribuye como una χ^2 con un grado de libertad si la variable es cuantitativa y con grados de libertad iguales al número de categorías menos uno, si la variable es cualitativa.

observados, dadas las estimaciones de los parámetros, es muy elevada, y que un modelo perfecto tendría una verosimilitud igual a 1, el valor de este estadístico debe ser lo más bajo posible.

- Estadístico Chi-Cuadrado del modelo. Es la diferencia en el valor del estadístico -$2LL_o$ entre los sucesivos pasos de construcción del modelo. Se distribuye como una chi-cuadrado con un número de grados de libertad igual a la diferencia de grados de libertad entre dos pasos concretos, que es igual al número de variables independientes del modelo. Sirve para someter a prueba la hipótesis de que los coeficientes del modelo, excepto la constante son nulos.
- Bondad de ajuste. Se construye comparando las probabilidades observadas con las predichas por el modelo[6] y contrasta la hipótesis nula de que el modelo es significativo

La clasificación de los individuos se realiza del modo siguiente: cuando la probabilidad estimada es inferior a 0,5, el individuo se asigna al primer grupo: los que SI eligen la asignatura, y cuando es superior a este valor, se clasifican en el segundo grupo, los No matriculados en ella.

En nuestro ejemplo, el proceso de selección se ha completado en 6 pasos. En los cinco primeros se han incluido, por este orden y siguiendo los criterios anteriormente mencionados, las siguientes variables: "Afinidad" con la asignatura, predisposición a continuar con "Otros estudios afines" , adecuación del "Horario" de asignatura, "Area" en que prefiere desarrollarse profesionalmente y nivel de "Asistencia a clase". En el último paso se excluye la variable relacionada con el "Area" profesional, por lo que al final, el sistema ofrece un modelo con cuatro variables. Sin embargo, este modelo con cuatro variables, que es significativo a nivel global (estadísticos -$2LL_0$,chi-cuadrado, bondad de ajuste), contiene una variable, -Asistencia a clase- cuyo coeficiente resulta no significativo según el estadístico de Wald. Dado que la inclusión de esta variable no aumenta la eficacia de la clasificación (% de casos correctamente clasificados), respecto

---

[6] Este estadístico es $Z^2 = \frac{\sum_{i=1}^{n}(E_i)^2}{p_i(1-p_i)}$, siendo E_i el residuo o diferencia entre la probabilidad observada y la estimada, pi la probabilidad estimada. Se distribuye como una χ^2 con n-k-1 grados de libertad.

al modelo con sólo las otras tres variables, se considera preferible éste último, que se presenta en la Tabla 4:

TABLA 4

Classification Table for ESTADII

The Cut Value is ,50

		Predicted		Percent Correct
		1	2	
Observed	1	47	2	95,92%
	2	5	10	66,67%
		Overall		89,06%

Variables in the Equation

Variable	B	S.E.	Wald	df	Sig	R	Exp (B)
AFINIDAD	-,8336	,2869	8,4424	1	,0037	-,3040	,4345
HORARIO	-,3986	,1794	4,9345	1	,0263	-,2052	,6713
OTROSEST	,8030	,2786	8,3090	1	,0039	,3009	2,2323
Constant	2,4085	2,1199	1,2909	1	,2559		

Los estadísticos de contraste global del modelo ofrecen resultados positivos: $-2LL_0$ es relativamente bajo, el chi-cuadrado es significativo, y lo mismo se puede decir de la bondad de ajuste. Asimismo, son significativos cada uno de los coeficientes de las variables según el estadístico de Wald. La variable Afinidad, mantiene una correlación parcial (R) negativa con la variable dependiente, lo que se interpreta como que a medida que aumenta el grado de afinidad, menor es el valor en la dependiente, es decir, más cerca se está del valor "0", que corresponde a elegir la asignatura. Lo mismo puede decirse respecto a la variable de adecuación del "Horario". En cambio, la posibilidad de realizar "Otros estudios afines" posee una correlación positiva, esto es, cuanto mayor es la propensión a realizarlos más próxima está la variable dependiente a adoptar el valor

“1”, y en consecuencia, mayor es la probabilidad de no cursar la asignatura. El cociente entre las probabilidades de No cursar la asignatura y Si cursarla, en relación con cada variable, aparecen en la última columna de la tabla, -Exp (B)-, en la que se observa como la cuestión de realizar otros estudios implica una probabilidad mucho mayor de No escoger la asignatura que de escogerla, y lo contrario es válido para las otras dos variables.

La clasificación tiene un nivel de acierto del 89,06%, dado que solamente 7 casos son asignados incorrectamente.

En conclusión, parece que ambos métodos pueden utilizarse para los objetivos mencionados al comienzo del trabajo. En nuestro caso concreto, los dos coinciden en las tres variables con mayor influencia en la decisión de elegir esta asignatura optativa: la Afinidad, la posibilidad de realizar Otros estudios afines, y la adecuación del Horario, incluyendo además el análisis discriminante, otras tres características, como son la valoración previa de su Dificultad, el grado de dificultad de la asignatura llave y la Edad. Según esto, los alumnos matriculados en esta materia son afines a ella, su horario les conviene, tienen menos intención de continuar con estudios universitarios afines, no juzgaron en principio que la asignatura fuera demasiado difícil, aunque la asignatura llave les pareciera algo más difícil que al otro grupo, y son algo más jóvenes.

Aunque sólo se ha pretendido realizar un primer ensayo sobre el tema, y evidentemente el modelo puede mejorarse, y sus conclusiones generalizarse con un muestreo más apropiado, parece que ahora se tiene una idea más clara sobre las motivaciones reales de los alumnos y que, conociendo estos aspectos, se podría predecir el número de personas matriculadas en esta opción, con un nivel de acierto relativamente elevado.

BIBLIOGRAFÍA

- FERRÁN, M. (1997): “SPSS para WINDOWS. Programación y análisis estadístico”. McGraw-Hill. Madrid.

- LEBART, L. y otros (1995): ”Statistique exploratoire multidimensionnelle”. Dunod. París.

- VISAUTA, B. (1998): “Análisis estadístico con SPSS para WINDOWS. Volumen II. Estadística multivariante”. McGraw-Hill. Madrid.