Revista Electrónica de Metodología Aplicada
2011, Vol. 16 nº 1, pp. 13-34

Eventos por Variable en Regresión Logística y Redes Bayesianas para Predecir Actitudes Emprendedoras

Jorge López Puga, Juan García García

Universidad de Almería

RESUMEN

Pese a que la regresión logística es una de las técnicas estadísticas de análisis mas usadas en ciencias sociales no está carente de ciertas limitaciones. El reducido tamaño de la muestra y la presencia de casos perdidos son algunas de las situaciones que han sido identificadas como problemáticas para la regresión logística. En este trabajo hemos comparado la regresión logística dicotómica y el clasificador simple de Bayes en su habilidad para predecir la tendencia emprendedora manipulando el número de eventos por variable. Una muestra de estudiantes universitarios (N = 1230) respondió a cinco escalas (motivación, actitud emprendedora, obstáculos, carencias y preparación percibida) que fueron utilizadas como variables predictoras de la tendencia emprendedora y a un conjunto de tres preguntas relativas a la tendencia emprendedora que fueron consideradas como variables de respuesta. Nuestros resultados indican que el número de eventos por variable afecta más a la regresión logística en términos del área bajo la curva ROC comparado con las redes bayesianas. Así pues, proponemos que las redes bayesianas podrían considerarse como otra alternativa más, junto a las ya existentes, para superar las debilidades de la regresión logística en determinadas condiciones de ejecución.

Palabras clave: eventos por variable, regresión logística dicotómica, redes bayesianas, actitudes emprendedoras, curva ROC, tamaño muestral.

ABSTRACT

Although logistic regression is one of the most commonly used data analysis techniques in social sciences it is also true that it has some limitations. A reduced sample size and the presence of missing data are some of the problems logistic regression can't cope with. In this work we compare the success of dichotomic logistic regression model and the Bayes simple classifier to predict entrepreneurship after manipulating the sample size. A sample of university undergraduate students (N = 1230) was asked to fill in five scales (motivation, attitude towards business creation, obstacles, deficiencies and training needs) whose scores were used as predictors and three questions referred to entrepreneurship tendency were considered as outcomes. Our results show that the Receiver Operating Characteristic (ROC) curve is affected by the number of events per variable in both techniques but logistic regression seems to be more vulnerable. We propose to use Bayesian networks as an additional alternative to surpass the weaknesses of logistic regression.

Keywords: events per variable, dichotomic logistic regression, Bayes nets, entrepreneurship, Receiver Operating Characteristic curve, sample size.

Contacto:
e-mail: jpuga@ual.es, jgarciag@ual.es.

1.- Introducción

La regresión logística surgió durante la década de 1960 como alternativa al procedimiento de estimación de los mínimos cuadrados ordinarios (OLS, *Ordinary Least Squares*) tradicionalmente usado en el modelo de regresión lineal (Long, 1997; Ortega y Cayuela, 2002) y su objetivo era estimar la probabilidad de ocurrencia de un evento como función de un conjunto de variables predictoras (Walker y Duncan, 1967). No obstante, su uso comenzó a popularizarse a partir de los años setenta cuando comenzó a implementarse en el software estadístico (Peng y So, 2002). El nacimiento de la regresión logística, por otro lado, estuvo condicionado por un contexto histórico dominado por una crisis en las herramientas estadísticas orientadas a la predicción ya que las herramientas de predicción discreta como los antecedentes de las Redes Neruales Artificiales, los perceptrones, mostraban serias limitaciones para hacer predicciones ante problemas relativamente simples (Cowell, Dawid, Lauritzen, y Spiegelhalter, 1999; Delgado, 2003; Quinlan, 1991; SPSS y Recognition Systems, 1997).

Pese a que el análisis de regresión logística tiene una serie de ventajas frente a, por ejemplo, el análisis de regresión lineal (por ejemplo, no ha de cumplir supuestos distribucionales en los datos), existen algunos puntos en los que el análisis de regresión logística tiene problemas (por ejemplo, el nivel de separación entre los valores de la variable resultado, problemas numéricos derivados de la frecuencia observada en las combinaciones entre los niveles de la/s variable/s predictora/s y la predicha, la colinealidad, el tamaño de muestra, el número de casos perdidos o la categorización de las variables). El tamaño de la muestra es posiblemente uno de los aspectos más preocupantes en el diseño de estudios que implican un análisis de regresión logística (p. e., Bull, Mak, y Greenwood, 2002; Firth, 1993; Harrel, Lee, Matchar, y Reichert, 1985; Jovel, 1995; Long, 1997; Ortega y Cayuela, 2002; Silva y Barroso, 2004). Ello se debe al propio procedimiento de estimación del modelo. Dado que el método de estimación del modelo de regresión logística es máximo-verosimil, a medida que aumenta el tamaño de la muestra las estimaciones son más estables y fiables.

La discusión relativa al efecto que tiene el tamaño de la muestra sobre los resultados del análisis de regresión logística ha girado en torno a la estimación de los parámetros del modelo así como en la determinación de las relaciones entre variables. En este sentido, se ha identificado tres tipos de errores que pueden producirse en la estimación de modelos logísticos (Irala, Fernández-Crehuet, y Serrano, 1997; Ortega y Cayuela, 2002; Peduzzi, Concato, Kemper, Holford, y Feinstein, 1996): el *sobreajuste* o *error tipo I* se produce cuando se mantienen muchas variables que sólo aportan «ruido» en el modelo final; el *sub-ajuste* o *error tipo II* se da cuando no han sido incluidas en el modelo final variables relevantes; y el *ajuste paradójico* o *error tipo III* que aparece cuando un factor concreto ha sido asociado en la dirección opuesta al efecto real que tiene la variable predictora sobre la variable resultado.

Lilienfeld y Pyne (1984) observaron que el tamaño de la muestra influía en la fiabilidad de las estimaciones de los parámetros, en la exactitud de la estimación de β y en la exactitud del contraste de hipótesis $H_0 : \beta = 0$. En concreto, encontraron que la desviación típica para las estimaciones de β en muestras simuladas se reducía a medida que aumentaba el tamaño de la muestra. La exactitud de esta estimación también se veía afectada a medida que se reducía el tamaño de la muestra, aunque en este caso la dirección del sesgo dependía de la distribución muestral de las observaciones (esto es, si los datos se distribuían exponencial, normal o binomialmente). En cuanto a la exactitud del contraste de hipótesis, también

encontraron que a medida que aumentaba la muestra los resultados eran más válidos. En lo que respecta al porcentaje de varianza de la variable resultado explicada por el conjunto de variables predictoras, DeMaris (2002) encontró que las muestras que contenían menos de 200 observaciones solían mostrar problemas en las estimaciones de los parámetros pero no observó ningún patrón significativo en relación al coeficiente de determinación.

Con el objetivo de minimizar el impacto del tamaño de la muestra, Whittermore (1981) desarrolló un procedimiento para calcular tamaños muestrales óptimos en el análisis de regresión logística usando una estimación máximo-verosimil asintótica de la matriz de covarianza a partir de la matriz de Hess, la cual es válida como representante de la matriz de información de Fisher cuando la probabilidad de respuesta es baja. Con posterioridad, Hsieh (1989) publicó un conjunto de tablas de tamaños muestrales para regresión logística simple y múltiple basados en la propuesta de Whittermore. Tanto Whittermore (1981) como Hsieh (1989) mostraron que sus tablas son útiles para estudios epidemiológicos en los que la prevalencia de los factores es alta o baja y que, aunque las tablas no son del todo precisas para ciertos factores de riesgo, son razonablemente apropiadas para distribuciones en las que los factores de riesgo se distribuyen exponencial y normalmente.

Recientemente se han desarrollado métodos más sencillos orientados a estimar el tamaño muestral necesario para estudios que implican análisis de regresión teniendo en cuenta la tasa de prevalencia de las variables predictoras, la potencia estadística así como su nivel de medida (Hsieh, Bloch, y Larsen, 1998). La recomendación general que se hace respecto al tamaño de la muestra es que exista un ratio entre el número de observaciones y el número de parámetros mayor que diez (Bull et al., 2002). Otro aspecto relacionado con el tamaño de la muestra que ha acaparado gran interés es el número de eventos por variable (EPV); esto es, la relación que existe entre el número de veces en que la variable resultado tiene resultados positivos y el número de variables del modelo de regresión (Concato, Peduzzi, Holford, y Feinstein, 1995; Harrel et al., 1985; G. King y Zeng, 2001a, 2001b; Peduzzi, Concato, Feinstein, y Holford, 1995; Peduzzi et al., 1996). En concreto, se estima que en la medida en que los EVP son mayores o iguales a diez, el porcentaje de sesgo relativo en la estimación de los coeficientes así como sus errores absolutos se reduce, la varianza de las estimaciones se minimiza, se reduce la tasa de ajustes paradójicos y aumenta la potencia estadística (Peduzzi et al., 1995, 1996).

Una de las estrategias que se ha seguido para paliar el problema del tamaño de la muestra ha sido el desarrollo de lo que se conoce como regresión logística exacta (Hirji, Mehta, y Patel, 1987; E. King y Ryan, 2002; Mehta y Patel, 1995) que se basa en cálculos condicionales y permutacionales. Otra técnica que se ha seguido para intentar solucionar los problemas que surgen de un reducido tamaño de muestra es la optimización de las funciones que hacen mínimo el error asintótico (Firth, 1993) en la estimación de los parámetros. Por ejemplo, (Bull et al., 2002) desarrollaron una función de penalización que reducía este tipo de problemas en la regresión logística multinomial.

El problema de la presencia de datos perdidos está muy relacionado con el problema del tamaño de la muestra descrito anteriormente. Ello es debido a que cuando existe un caso con un dato perdido en alguna de sus variables en una muestra, este caso suele ser excluido del análisis y, por consiguiente, el tamaño de la muestra se reduce. Aunque se han desarrollado procedimientos de imputación orientados a reducir el impacto de la tasa de

valores perdidos en las muestras (p. e. Allison, 2002; Hair, Anderson, Tatham, y Black, 1998), lo más recomendable es poner medidas para que se produzca la menor tasa posible de datos perdidos.

Por su parte, las redes bayesianas (también conocidas como redes causales probabilísticas, sistemas expertos bayesianos, sistemas expertos probabilísticos, redes causales, redes de creencia o diagramas de influencia) son herramientas estadísticas que representan un conjunto de incertidumbres asociadas con base en las relaciones de independencia condicional que se establecen entre ellas. Pertenecen al conjunto de técnicas orientadas a la modelización gráfica (Martínez y Rodríguez, 2003) y forman parte de la familia de los *Sistemas Estocásticos Altamente Estructurados* (Cowell et al., 1999). Siguiendo a Kadie, Hovel, y Horvitz (2001) diríamos que una red bayesiana es un conjunto de variables, una estructura gráfica conectando estas variables y un conjunto de distribuciones de probabilidad condicional. Este tipo de red codifica la incertidumbre asociada a cada variable por medio de funciones probabilísticas y, gracias al teorema de Bayes, esta incertidumbre es susceptible de ser modificada en base a observaciones (o evidencias) sobre el modelo. Aunque las redes bayesianas han sido utilizadas en psicología como modelos formales normativos para modelar procesos psicológicos (p. e., Conati, Gertner, y VanLehn, 2002; Conejo et al., 2004; Conejo, Millán, Pérez-de-la-Cruz y Trella, 2001; Gopnik y Schulz, 2004; Gopnik et al., 2004; Gopnik, Sobel, Schulz y Glymour, 2001; Glymour, 2001, 2003; Jurafsky, 1996; Narayan y Jurafsky, 1998, 2002; Krynski y Tenenbaum, 2007; López y García, 2009; Martin y VanLehn, 1995; Mislevy y Gitomer, 1996), no han sido ampliamente utilizadas como herramientas de análisis de datos en el contexto de la psicología (López, García, De la Fuente, y De la Fuente, 2007). Asimismo, las aplicaciones analíticas en nuestro contexto más cercano han estado relacionadas con el estudio de las actitudes (p. e., García, López, Cano, Gea y De la Fuente, 2006; López, 2009; López y García, 2007; López, García, Cano, Gea, y De la Fuente, 2009).

Varios trabajos sugieren que los modelos basados en estructura de red podrían tener ventajas frente a la regresión logística en términos predictivos (p. e., Ankarali, Canan, Akkus, Bugdayci, y Ali, 2007; Bartfay, Mackillop, y Pater, 2006; Eftekhar, Mohammad, Ardebili, Ghodsi, y Ketabchi, 2005; Finch y Schneider, 2007; García, López, De la Fuente, Cano, y Gea, 2007; Jaimes, Farbiarz, Alvarez, y Martínez, 2005; Kumar, Rao, y Soni, 1995; Terrin, Schmid, Griffith, D'Agostino, y Selker, 2003). No obstante, no hay evidencias robustas que muestren una mejor ejecución de las técnicas basadas en modelos de red frente a la regresión logística. Por su parte, los trabajos que comparan las redes bayesianas y la regresión logística no son numerosos según la bibliografía revisada. Por ejemplo, Lee, Abbott, y Johantgen (2005) hicieron notar que el uso de las redes bayesianas podrían representar ciertas ventajas frente al uso de la regresión logística. Entre las ventajas que suponen las redes bayesianas frente a la regresión logística cabe destacar la superación de ciertos supuestos estadísticos como el de aditividad, la facilidad para el manejo de una gran cantidad de predictores, la facilidad para identificar e interpretar los efectos de interacción y la facilidad para modelar relaciones no lineales entre variables (Lee et al., 2005). Por otro lado, en un trabajo reciente (López, Ruiz-Ruano, y García, 2008) se ha observado que las redes bayesianas son más eficientes en la predicción que la regresión logística bajo ciertas circunstancias.

Por otro lado, el estudio de las actitudes emprendedoras ha cobrado especial relevancia recientemente dadas las consecuencias socioeconómicas que conlleva (Corman, Lussier, y

Nolan, 1996). Más específicamente, el estudio de las actitudes o tendencias emprendedoras en la universidad tiene relevancia tanto desde un punto de vista práctico como teórico (Cano, García, y Gea, 2003; Ruiz, Rojas, y Suárez, 2008).

El origen del concepto de *emprendedor* se puede situar en el economista irlandés Richard Cantillon (Hayek, 1985), quien identifica a este agente social como el verdadero catalizador del desarrollo económico. Aunque, como señalaron Hébert y Link (1989), y pese a las críticas expuestas por Gartner (1988); podríamos conceptualizar al emprendedor como "una persona o grupo que pretende explotar una oportunidad económica" (McKenzie, Ugbah, y Smothers, 2007, p. 24). Como hizo notar Samuelson (1970), la función que caracterizaría a una persona emprendedora sería la introducción de un nuevo producto o servicio en el mercado más que inventarlo o crearlo. En este trabajo nos hemos centrado en concepto de emprendedor potencial (Huefner, Hunt, y Robinson, 1996; López et al., 2009, López y García, 2010), que se define como una persona que aún no ha creado su empresa pero que le gustaría hacerlo a corto o medio plazo. Más concretamente, dado que nuestra muestra está integrada por estudiantes universitarios, consideramos que los emprendedores potenciales son los estudiantes universitarios a los que les gustaría crear su empresa una vez finalizados sus estudios. Desde un punto de vista teórico, el estudio de las actitudes emprendedoras en estudiantes universitarios puede arrojar luz sobre la naturaleza del emprendedor potencial; mientras que desde un punto de vista práctico se podrían poner en marcha programas de intervención que optimizaran las posibilidades de éxito de las empresas creadas en el seno de las universidades.

Por actitud emprendedora vamos a entender *la predisposición aprendida de responder favorable, mixta o desfavorablemente ante la creación de una empresa* (Brehm, Kassin, y Fein, 2005; Feldman, 1998). Por otro lado, dado que la concepción multidimensional de la actitud ha sido la que más atención ha recibido desde un punto de vista clásico (Allport, 1935), vamos a tomar como modelo teórico de referencia el marco tridimensional de la actitud ((p.e., Ajzen y Fishbein, 2005; Deaux, Dane, y Wrightsman, 1993; Feldman, 1998; Franzoi, 2005; Morales, Rebolloso, y Moya, 1994) integrado por una dimensión cognitiva, otra conductual y un aspecto emocional. Más concretamente, vamos a considerar la *Teoría del Comportamiento Planeado* (TCP) como argumento de fondo, en términos de la relación que se establece entre actitud y comportamiento, sobre el que se basará nuestro análisis (Ajzen y Fishbein, 2005). El antecedente del la TCP se conoce como *Teoría de la Acción Razonada* (TAR) (Ajzen y Fishbein, 1980); no obstante, ambas teorías suponen que el comportamiento ha sido razonado, que las personas piensan sobre las consecuencias de sus actos y que toman decisiones orientadas a alcanzar unos resultados y a evitar otros. Los estudios destinados a estudiar la naturaleza de las actitudes emprendedoras no son muy abundantes en nuestro contexto. No obstante, se ha desarrollado una Escala de Actitudes hacia la Creación de Empresas (ACEMP) que ha mostrado tener un aceptable valor predictivo en diferentes contextos (Cano et al., 2003; García, Cano y Gea., 2005, 2006, 2007; López, 2009; López et al., 2009).

Con lo expuesto anteriormente, el objetivo de este trabajo es comparar la ejecución de la regresión logística y las redes bayesianas para predecir la tendencia hacia la creación de empresas en función del número de eventos por variable. Se utilizará el modelo de regresión logística binaria y el clasificador ingenuo de Bayes (Bayes Naïve Classifier) (Martínez y Rodríguez, 2003) o clasificador simple de Bayes (Domingos y Pazzni, 1996). Aunque

recientemente se han desarrollado técnicas orientadas a incorporar la filosofía de la estadística bayesiana en el análisis de regresión logística binaria (p. e., Genkin, Lewis, y Madigan, 2005; Ortiz, Martín, Ureña, y García, 2005), este trabajo se centrará en comparar la regresión logística binaria clásica y el clasificador simple de Bayes porque ambas técnicas surgieron en la misma época, mientras que la versión bayesiana de la regresión logística es un fenómeno más reciente; y porque la regresión logística binaria clásica está más extendida que su homóloga bayesiana, con lo cual las comparaciones pueden ser más útiles desde un punto de vista práctico. En lo que respecta al clasificador ingenuo de Bayes, se ha seleccionado este modelo por ser el que más se parece desde un punto de vista formal a la regresión logística (Greiner, Su, Shen, y Zhou, 2005; Greiner y Zhou, 2002; Shen, Su, Greiner, Musilek, y Cheng, 2003). Se considera que tanto la regresión logística como la red bayesiana son herramientas de clasificación; esto es, funciones que asignan una etiqueta de clase a ejemplos, típicamente descritas o caracterizadas por un conjunto de atributos (Shen et al., 2003). Por eficiencia predictiva se entenderá, en términos generales, el grado en que una herramienta de clasificación proporciona respuestas correctas de manera frecuente (Greiner et al., 2005). No obstante, las dos técnicas comparadas en este trabajo producen diferentes tasas de clasificaciones correctas dependiendo del umbral de corte establecido para la clasificación. Así pues, utilizaremos curvas ROC para comparar los niveles de especificidad y sensibilidad para diferentes puntos de corte en las probabilidades proporcionadas por ambos modelos (p. e., DeMaris, 2002; Hanley y McNeil, 1982, 1983). Dado que el tamaño de la muestra es uno de los problemas que se han destacado del análisis de regresión logística porque las estimaciones de los coeficientes del modelo son sesgadas a medida que la muestra se hace más pequeña (p. e., Hsieh, 1989; Hsieh et al., 1998; E. King y Ryan, 2002; G. King y Zeng, 2001a, 2001b; Whittermore, 1981), esperamos encontrar que las redes bayesianas obtengan mayores áreas bajo la curva ROC en comparación con las obtenidas por la regresión logística binaria a medida que el tamaño de la muestra se hace más pequeño en términos de eventos por variable (Concato et al., 1995; Peduzzi et al., 1995, 1996). Por otro lado, dado que la tasa de respuestas positivas en la variable predicha o de respuesta es un factor relevante en las estimaciones de los modelos, se prevé que este parámetro influirá en los niveles de validez predictiva de ambas técnicas.

2.- Metodología

2.1. - Participantes

Una muestra de 1230 estudiantes universitarios participaron en el estudio. Los participantes fueron seleccionados con base en un muestreo por bloques con afijación proporcional por sexo y titulación (error muestral ± 3%, IC 95%, z = 1,96 y $p = q$ = 0,5) de los primeros y últimos cursos de todas las titulaciones que se imparten en la Universidad de Almería. Un total de 426 (34,6%) participantes fueron hombres y el resto (797, 64,8%) mujeres, mientras que la edad de los participantes estuvo comprendida entre los 17 y los 56 años, M = 22,45 y DT = 4,46.

2.2.- Materiales

Para la recogida de información se utilizó un cuestionario de tres folios tipo A4 impresos a doble cara y grapados por su esquina superior izquierda. En la primera página del

formulario aparecieron los membretes de las entidades patrocinadoras del estudio. A continuación aparecía una breve presentación de finalidad del estudio y un requerimiento formal para colaborar en el mismo. También se indicó que los datos recogidos con el cuestionario serían tratados confidencialmente y se facilitó información de contacto con los responsables de la investigación por si alguien tenía alguna duda. Para terminar se agradeció la colaboración en la investigación.

En el reverso de la portada aparecieron los espacios dedicados a recoger los datos sociodemográficos y una pregunta referida a la preferencia laboral. En la siguiente página se mostraron varias preguntas relativas a la intención de crear una empresa estándar o sin ánimo de lucro y una pregunta para evaluar el grado en que los participantes consideraban más o menos fácil la creación de una empresa respecto al pasado. Las tres primeras preguntas de esta segunda página son las variables sobre las que se han basado los modelos de regresión logística y de red bayesiana en términos predictivos. Esto es, cada una de estas variables fueron consideradas (una a una) como variables predichas o de respuesta, en los modelos de regresión logística y como nodos clase o divergentes, en los modelos de red bayesiana. Las preguntas fueron: ¿Considera deseable crear una empresa propia? ($N_{Sí}$ = 913) ¿Ha pensado seriamente, como una opción real a corto/medio plazo, montar su propia empresa? ($N_{Sí}$ = 463) y ¿Ha iniciado en algún momento acciones encaminadas a montar su propia empresa? ($N_{Sí}$ = 109) De este modo, la primera pregunta hace referencia a la deseabilidad de creación empresas (DES), la segunda pregunta se referiría a la tendencia a corto y medio plazo hacia la creación de empresas (TEM) mientras que la tercera pregunta se referiría a la dimensión conductual (CON). Estas tres preguntas fueron respondidas por los participantes en términos de «sí» o «no». Para responder a cada pregunta los participantes tuvieron que marcar la respuesta deseada marcando en una casilla de verificación que aparecía a la izquierda de cada posible alternativa de respuesta.

Bajo este conjunto de preguntas relativas a la creación de empresas aparecieron cinco escalas cuyas puntuaciones fueron utilizadas como variables predictoras en los modelos de regresión logística y como variables de divergencia en los modelos de red bayesiana. En todas ellas los ítems se puntuaron en una escala tipo Likert con cinco alternativas (excepto en la Escala de Actitudes hacia la Creación de Empresas que se puntuó en una escala de cuatro alternativas) y se corrigieron dividiendo el sumatorio de todos los ítems por el número de ítem de cada escala tras invertir los ítems expresados en sentido negativo. En primer lugar, aparecieron escalas de Carencias Formativas Percibidas (12 ítems, α = 0,89; M = 2,38; DT = 0,66), Preparación Percibida (4 ítems, α = 0,85; M = 2,08; DT = 0,76), Motivación para Crear un Negocio (11 ítems, α = 0,88; M = 3,26; DT = 0,57) y Obstáculos Percibidos ante la Creación de una Empresa (17 ítems, α = 0,83; M = 3,17; DT = 0,58) donde todos los ítems estuvieron expresados en sentido positivo. En la tercera página apareció la Escala de Actitudes hacia la Creación de Empresas (ACEMP) (Cano et al., 2003; García et al., 2005, 2006, 2007; López et al., 2009). La escala ACEMP consta de 29 ítems relativos a la frecuencia con que se realizan ciertos comportamientos o pensamientos de los cuales 13 ítems son inversos respecto a una interpretación positiva de la puntuación, a más cantidad de puntuación mayor cantidad de actitud emprendedora (α = 0,75; M = 2,77; DT = 0,31). La escala está integrada por 10 facetas (Creatividad, Perseverancia, Capacidad de Organización, Independencia, Confianza en si Mismo, Riesgo Calculado, Tolerancia a la Incertidumbre,

Competitividad, Negociación y Locus de Control) que constituyen la actitud general hacia tareas emprendedoras.

2.3.- Procedimiento

Los cuestionarios fueron administrados entre los meses de diciembre de 2005 y mayo de 2006 a estudiantes de los primeros y últimos cursos de los estudios que se imparten en la Universidad de Almería. El cuestionario fue administrado de forma grupal en horas lectivas con el consentimiento del profesorado y la participación voluntaria del alumnado. Para solicitar la colaboración por parte de los profesores/as y con el fin de que donaran parte de su horario lectivo para que su alumnado pudiese cumplimentar el cuestionario, se contactó con ellos/as con un correo electrónico que fue el mismo en todos los casos. Antes de la administración se dieron unas instrucciones generales (marcar la alternativa deseada en cada ítem y que no había respuestas correctas o incorrectas, que la tarea consistía en responder según sus opiniones) y se pidió a los participantes que leyesen detenidamente la carta de presentación que aparecía en la primera hoja del formulario. Durante la cumplimentación del cuestionario el administrador del mismo resolvía las dudas que pudiesen surgir. La cumplimentación del cuestionario tenía una duración de entre 20 y 30 minutos. No se dio ninguna recompensa por la colaboración excepto el agradecimiento verbal.

2.4.- Análisis de datos

Dado que había cinco variables predictoras que podían ser candidatas para formar parte del mejor modelo sustantivo y estadístico, y teniendo en cuenta el principio de parsimonia respecto al número de variables; llevamos a cabo una comparación de todas las posibles combinaciones de variables predictoras que se podrían dar en esta situación. En total se estimaron 93 modelos, un tercio de cada uno para cada variable resultado. No fueron introducidos términos de interacción para no tratar la complejidad de este caso. Aunque podríamos haber tratado este tema, estudios posteriores tendrán que dirigir investigaciones a solventar este tópico. En nuestro caso, los resultados que se obtengan aquí, en cualquier caso, se pueden tomar como un indicativo conservador del caso más complejo que incluiría a los términos de interacción (Peduzzi et al., 1995). Una vez seleccionado el conjunto de variables que mejor predecía cada una de las variables resultado se seleccionaron 21 muestras aleatorias independientes de la base de datos original. El criterio para seleccionar los tamaños muestrales fue la tasa de eventos por variable que se estableció en 5, 10, 15, 20, 25, 30, 100 y 200 análogamente a como lo hicieron Peduzzi et al. (1995, 1996) y Concato et al. (1995). Los tamaños muestrales establecidos para los modelos DES fueron 20, 40, 61, 81, 101, 121, 404 y 808; para los modelos TEM fueron 53, 106, 159, 213, 266, 319 y 1066; mientras que para los modelos CON fueron 169, 339, 508, 677, 846 y 1016. La diferencia en el número condiciones de tamaño muestral se explica por la tasa diferencial de respuestas positivas en las variable resultado como se indicó en la sección de materiales cuando se describieron las variables de respuesta.

La estimación de los modelos de regresión logística se llevó a cabo utilizando el paquete estadístico PASW en su versión 18.0 (*SPSS Inc.*) en su formato no condicionado, sin incluir ningún término de interacción y configurando el nivel de confianza de las estimaciones al 95%. No se utilizó ningún método de selección secuencial de variables predictoras y por tanto se configuró el programa para que se introdujesen en los modelos de regresión logística las variables definidas previamente por el diseño del estudio. Los parámetros de las redes

bayesianas fueron estimados utilizando el método de máxima verosimilitud corregido con la ecuación de la sucesión de Laplace (Morales, 2006; Ng y Jordan, 2002). Tanto la estimación paramétrica como la posterior actualización de probabilidades, así como la estimación de los parámetros de bondad de ajuste se llevaron a cabo con Netica 4.02 (*Norsys Sofware*). Las estructuras gráficas fueron creadas manualmente con base al diseño del estudio. Para cada modelo se estimaron sus correspondientes índices de bondad de ajuste y para todos se obtuvieron los valores de sensibilidad, especificidad, proporción de falsos positivos, proporción de falsos negativos, valor predictivo positivo, valor predictivo negativo y proporción general de clasificaciones correctas.

El área bajo la curva ROC (θ) se estimó usando la técnica no paramétrica propuesta por Hanley y McNeil (1982) que se basa en el estadístico *U* de Mann-Whitney. Para comprobar si existían diferencias estadísticamente significativas entre las curvas ROC generadas por la regresión logística y la red bayesiana se utilizó el procedimiento propuesto por Hanley y McNeil (1983). Dado que se trata de muestras pareadas, se tuvo en cuenta la correlación entre ambas áreas bajo la curva.

3.- Resultados

Atendiendo a criterios estadísticos y sustantivos se seleccionaron tres conjuntos de variables diferentes como predictores de las tres variables de respuesta. La puntuación de la escala ACEMP, el nivel de preparación percibida y el número de obstáculos percibidos fueron seleccionados como predictores de la tendencia conductual (CON) hacia la creación de empresas. A éstas tres se les añadió la puntuación de la escala motivacional para predecir la deseabilidad de crear una empresa (DES) y, por último, para predecir la respuesta a la pregunta sobre la tendencia temporal hacia la creación de empresas (TEM) se seleccionaron la puntuación de la escala ACEMP, el grado de carencias percibidas y la preparación percibida.

Como se puede observar en la Figura 1, la manipulación del tamaño de la muestra beneficia globalmente a las redes bayesianas. Aunque en ambas técnicas se produce una reducción del área bajo la curva ROC a medida que se reduce el porcentaje de respuestas positivas en la variable de respuesta, el área bajo la curva ROC que producen las redes bayesianas es mayor que la obtenida con la regresión logística. Así, mientras que el área bajo la curva ROC para el componente de deseabilidad de la tendencia emprendedora estimada en la regresión logística es de 0,7761 (z = 6, 86, p < 0, 001), el de la red bayesiana es 0,8344 (z = 9, 15, p < 0,001); el área para el componente temporal estimada usando la regresión logística es 0,6932 (z = 5, 96, p < 0, 001) mientras que el de la red bayesiana es 0,7945 (z = 11, 12, p < 0, 001); y, por último, la estimación del área bajo la curva ROC para la regresión logística es de 0,6337 (z = 2, 81, p = 0, 002) cuando la estimada para la red bayesiana es 0,7235 (z = 4, 91, p < 0, 001) en términos del componente conductual.

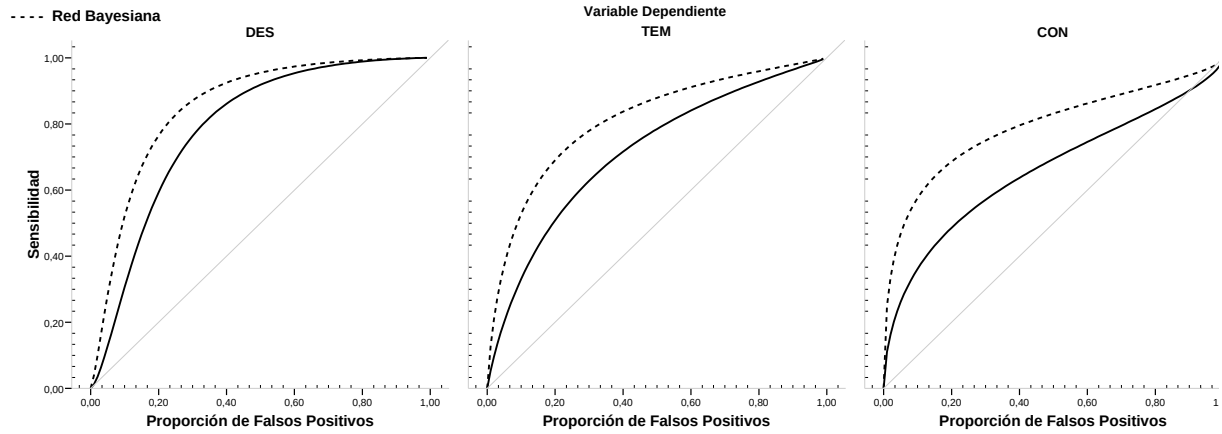

Figura 1. Curvas ROC promedio en función de la variable resultado.

En términos diferenciales globales, la mayor diferencia estadísticamente significativa entre las curvas producidas por la regresión logística y la red bayesiana se observa con la variable predicha del componente temporal de la actitud emprendedora, donde las redes bayesianas mejoran en un 10,14% respecto a la regresión logística (z = 5, 76, p < 0, 001); o lo que es lo mismo, cuando la tasa de respuestas positivas y negativas está más equilibrada. Por otro lado, cuando la tasa de respuestas positivas de la variable de respuesta es baja (como sucede con el componente conductual de la tendencia emprendedora) las redes bayesianas también muestran una mejora estadísticamente significativa del 8,68% en la predicción en comparación con la regresión logística (z = 2, 41, p = 0, 008). Sin embargo, no se observan diferencias estadísticamente significativas en términos globales entre las redes bayesianas y la regresión logística cuando la tasa de respuestas positivas es muy elevada en la variable resultado como pasa en el caso del componente de deseabilidad en la tendencia emprendedora (z = 1, 31, p = 0, 095).

Si hacemos un análisis detallado teniendo en cuenta el tamaño de la muestra en términos de eventos por variable y nos fijamos en las áreas bajo las curvas ROC se observa que cuando la tasa de respuestas positivas en la variable de respuesta es muy alto, como pasa con el componente de deseabilidad de la tendencia emprendedora, las redes bayesianas muestran una ejecución casi perfecta cuando el número de eventos por variable es relativamente bajo, con tamaños muestrales inferiores a 100 casos. Sin embargo, cuando el número de eventos por variable es relativamente grande (100 o 200) que suponen muestras entre 400 y 800 casos, la ejecución de las redes bayesianas y la regresión logística es prácticamente la misma (Tabla 1) y no existen diferencias estadísticamente significativas entre las áreas ROC generadas por ambas técnicas (Tabla 2). Otra observación importante es, como se puede apreciar en la Tabla 1, que las curvas ROC generadas por la regresión logística cuando el tamaño de muestra es de 20 a 40 (con 5 y 10 eventos por variable respectivamente) no alcanzan a ser estadísticamente diferentes de 0,5 con un nivel de confianza del 95%.

Técnica	*VR*	*EPV*	θ	DT_θ	*z*	*p*	IC_{inf}	IC_{sup}
RL	DES	5	0,7157	0,1343	1,61	0,054	0,4525	0,9789
	DES	10	0,6149	0,1280	0,90	0,185	0,3640	0,8657
	DES	15	0,8191	0,0624	5,11	*	0,6968	0,9414
	DES	20	0,7316	0,0588	3,94	*	0,6164	0,8468
	DES	25	0,6192	0,0707	1,69	0,046	0,4807	0,7577
	DES	30	0,6579	0,0615	2,57	0,005	0,5373	0,7785
	DES	100	0,8409	0,0248	13,75	*	0,7923	0,8894
	DES	200	0,7770	0,0202	13,70	*	0,7374	0,8166
	TEM	5	0,7197	0,0897	2,45	0,007	0,5438	0,8956
	TEM	10	0,7297	0,0542	4,24	*	0,6235	0,8360
	TEM	15	0,6442	0,0463	3,12	0,001	0,5535	0,7350
	TEM	20	0,7167	0,0386	5,61	*	0,6411	0,7924
	TEM	25	0,6602	0,0356	4,50	*	0,5905	0,7299
	TEM	30	0,7025	0,0315	6,42	*	0,6407	0,7644
	TEM	100	0,6944	0,0175	11,10	*	0,6601	0,7288
	CON	5	0,6555	0,0908	1,71	0,043	0,4775	0,8335
	CON	10	0,5478	0,0677	0,71	0,240	0,4152	0,6804
	CON	15	0,6065	0,0548	1,94	0,026	0,4990	0,7139
	CON	20	0,6605	0,0452	3,55	*	0,5719	0,7491
	CON	25	0,6602	0,0405	3,95	*	0,5808	0,7396
	CON	30	0,6191	0,0365	3,27	0,001	0,5477	0,6906
RB	DES	5	0,9854	0,0365	13,29	*	0,9138	1,0000
	DES	10	0,9891	0,0308	15,88	*	0,9287	1,0000
	DES	15	0,9889	0,0134	36,57	*	0,9627	1,0000
	DES	20	0,9417	0,0375	11,79	*	0,8683	1,0000
	DES	25	0,9732	0,0146	32,45	*	0,9447	1,0000
	DES	30	0,8584	0,0529	6,78	*	0,7547	0,9621
	DES	100	0,8060	0,0274	11,17	*	0,7523	0,8597
	DES	200	0,7778	0,0203	13,71	*	0,7381	0,8175
	TEM	5	0,9299	0,0397	10,82	*	0,8520	1,0000
	TEM	10	0,9356	0,0218	19,94	*	0,8927	0,9784
	TEM	15	0,8225	0,0352	9,18	*	0,7536	0,8914
	TEM	20	0,8656	0,0271	13,51	*	0,8126	0,9187
	TEM	25	0,8315	0,0254	13,05	*	0,7817	0,8814
	TEM	30	0,7758	0,0278	9,92	*	0,7213	0,8303
	TEM	100	0,7460	0,0156	15,75	*	0,7154	0,7766
	CON	5	0,7950	0,0793	3,72	*	0,6395	0,9504
	CON	10	0,7343	0,0663	3,53	*	0,6043	0,8643
	CON	15	0,7570	0,0490	5,24	*	0,6609	0,8530
	CON	20	0,7230	0,0426	5,24	*	0,6395	0,8065
	CON	25	0,7222	0,0373	5,97	*	0,6492	0,7953
	CON	30	0,6943	0,0344	5,65	*	0,6269	0,7617

Nota. RL: regresión logística, RB: red bayesiana, IC: intervalo de confianza al 95% para θ, *$p <$ 0,001, bilateral.

Tabla 1. Áreas bajo la curva ROC en función del número de eventos por variable (EPV), de la variable resultado (VR) y del tipo de técnica

VR	*EPV*	r_p	r_a	*r*	$DT_{\theta-\theta'}$	$z_{\theta-\theta'}$	$p_{\theta\neq\theta'}$
DES	5	0,501	0,496	0,410	0,12	2,18	0,015
DES	10	0,200	0,001	0,080	0,13	2,90	0,002
DES	15	0,456	0,115	0,210	0,06	2,78	0,003
DES	20	0,335	0,550	0,380	0,06	3,72	*
DES	25	0,052	0,214	0,070	0,07	4,98	*
DES	30	0,248	0,205	0,190	0,07	2,74	0,003
DES	100	0,590	0,795	0,630	0,02	1,55	0,061
DES	200	0,714	0,792	0,700	0,02	0,05	0,481
TEM	5	0,464	0,199	0,280	0,09	2,41	0,008
TEM	10	0,685	0,308	0,420	0,05	4,18	*
TEM	15	0,742	0,457	0,540	0,04	4,43	*
TEM	20	0,667	0,472	0,510	0,03	4,38	*
TEM	25	0,537	0,626	0,540	0,03	5,61	*
TEM	30	0,708	0,700	0,660	0,02	2,97	0,002
TEM	100	0,884	0,843	0,840	0,01	5,41	*
CON	5	0,736	0,700	0,670	0,07	2,00	0,023
CON	10	0,667	0,531	0,550	0,06	2,93	0,002
CON	15	0,689	0,664	0,630	0,04	3,35	*
CON	20	0,877	0,667	0,730	0,03	1,93	0,027
CON	25	0,826	0,689	0,710	0,03	2,08	0,019
CON	30	0,795	0,703	0,710	0,03	2,78	0,003

Nota. r_p*:* correlación entre las respuestas de las técnicas para los casos con respuesta positiva, r_a: correlación entre las respuestas de las técnicas para los casos con respuesta negativa, *r*: correlación entre las áreas bajo las curva, $*p < 0{,}001$, bilateral.

Tabla 2. Comparación de las áreas bajo la curva ROC en función de la variable resultado (VR) y el número de eventos por variable (EPV)

Cuando analizamos las curvas ROC generadas por los modelos de red bayesiana y regresión logística que incluyen como variables de respuesta al componente cognitivo y conductual de la actitud hacia la creación de empresas se observa que las diferencias son menores. Sin embargo, en todos los casos la curva ROC generada por la red bayesiana deja bajo sí un área mayor que la que produce la regresión logística. Otra observación importante es que el área bajo la curva ROC se va reduciendo sensiblemente en la curvas generadas por las redes bayesianas a medida que se aumenta en número de eventos por variables. Este cambio no es tan evidente en el caso de la regresión logística. Así pues, se pueden hacer dos apreciaciones básicas en relación a la evolución del área bajo la curva ROC en ambas técnicas a medida que aumenta el número de eventos por variable. En primer lugar, el área bajo la curva ROC tiende a descender a medida que el número de eventos por variable va aumentando cuando se utiliza la red bayesiana mientras que el área para la regresión logística tiende a mantenerse más estable. Además, el área bajo la curva ROC producida por la regresión logística tiende a ser inferior a la producida por la red bayesiana excepto en el caso de variables resultado con una alta tasa de respuestas positivas donde ambas técnicas funcionan similarmente cuando la tasa de eventos por variable está en torno al 100.

Por último, como se puede observar en la Tabla 1, las mayores diferencias se observan cuando el número de eventos por variable es inferior a 100 para el caso de la dimensión de

deseabilidad de la tendencia hacia la creación de empresas, cuando el número de eventos por variable es inferior a 30 en el componente temporal de la tendencia y cuando el número de eventos positivos es igual o inferior a 25 en el componente conductual. En relación a la variable predicha, se puede apreciar que el componente de deseabilidad en la creación de empresas es el que produce una mayor ejecución diferencial, mientras que el componente temporal es el que muestra un porcentaje de diferencia más bajo.

4. – Discusión

El hallazgo más relevante de este estudio, como pensábamos inicialmente, es que la manipulación del tamaño de la muestra, o más concretamente del número de eventos por variable, tiene un efecto sobre la habilidad predictiva más pronunciado sobre la regresión logística que sobre las redes bayesianas. En términos generales, la habilidad predictiva de la regresión logística se ve más amenazada, en comparación con las redes bayesianas, a medida que la muestra se hace más pequeña. El hecho de que ambas técnicas converjan en su tasa predictiva en muestras grandes pone de manifiesto la naturaleza asintótica de la teoría clásica (p. e., Firth, 1993; Freedman y Pee, 1989) frente a la perspectiva bayesiana (p. e., Alonso y Tubau, 2002; Cowell et al., 1999; De la Fuente, García, y De la Fuente, 2002; Heckerman, 1995; Serrano, 2003). Estos resultados podrían tomarse en cuenta de una manera prescriptita y, como señala López (2009), se podría plantear el uso de las redes bayesianas como otra alternativa más en situaciones con pequeño tamaño muestral donde se deseen desarrollar modelos probabilísticos.

También se ha puesto de manifiesto, como hipotetizábamos al principio, que el área bajo la curva ROC se reduce a medida que disminuye la tasa de respuestas positivas en la variable predicha aunque, en todos los casos, la red bayesiana sigue comportándose mejor, en términos predictivos, que la regresión logística. Estos resultados son consistentes con la idea de que en muestras balanceadas se necesita menor muestra para que funcione bien la regresión logística mientras que se necesita un mayor tamaño muestral cuando hay una baja prevalencia en la variable resultado (Hsieh et al., 1998; Whittermore, 1981).

Estos resultados son también consistentes con los encontrados por Concato et al. (1995) quienes observaron que los coeficientes de modelo de regresión logística tienden a ser sesgados cuando la variable predicha tiene valores reducidos de eventos por variable en el modelo. Así pues, además de producirse un aumento en la varianza de las estimaciones del modelo de regresión logística, de producirse ajustes paradójicos y de reducirse la potencia estadística (Peduzzi et al., 1995, 1996); la reducción del número de eventos por variable tiende a ser más crítico para la regresión logística que para las redes bayesianas en términos de validez predictiva. Una debilidad de nuestro trabajo que deberán afrontar futuras investigaciones está relacionado con las tasas de eventos por variable. Aunque hemos utilizado siete condiciones a este respecto, algunas de ellas han mostrado no provocar ningún efecto considerable. En futuras investigaciones se debería de profundizar sobre otros factores que puedan interactuar con estos parámetros en la validez predictiva de las redes bayesianas y la regresión logística.

Los resultados aquí obtenidos pueden ser de utilidad como criterio de toma de decisiones ante la elección de una u otra técnica teniendo en cuenta las condiciones de los datos y del modelo deseado. Dado que una de las desventajas más llamativas de la regresión

logística es el trabajar con muestras pequeñas (p. e., Bull et al., 2002; Long, 1997; Ortega y Cayuela, 2002; Silva y Barroso, 2004), el uso de las redes bayesianas puede considerarse como una alternativa más a las opciones ya disponibles para subsanar tal debilidad. Por tanto, el uso de una red bayesiana para la construcción de modelos probabilísticos puede tener relevancia como alternativa a la regresión logística clásica junto a la regresión logística exacta (p. e., Hirji et al., 1987; Mehta y Patel, 1995) o a las funciones de minimización del error asintótico en la estimación (p. e., Bull et al., 2002). Sin embargo, nuestro trabajo podría criticarse por no haber utilizado datos reales en los que no se ha manipulado directamente factores críticos (como la tasa de respuestas positivas en la variable predicha) y sería recomendable que futuros estudios utilizasen simulaciones Monte Carlo para intentar replicar los resultados que aquí se han obtenido.

En cuanto al aspecto sustantivo, los resultados de este estudio sirven para caracterizar el perfil del emprendedor potencial entre los estudiantes universitarios (Huefner et al., 1996). En este sentido, hemos creado tres modelos diferentes que sirven para predecir los tres componentes (de deseabilidad, temporal y conductual) de la tendencia emprendedora en estudiantes universitarios utilizando dos técnicas estadísticas diferentes que han mostrado niveles razonables de ajuste predictivo. La figura del emprendedor potencial es especialmente importante cuando se pretende optimizar las probabilidades de éxito de los futuros creadores de empresas (López, 2009; López et al., 2009; López y García, 2010). En términos aplicados, una buena definición de emprendedor potencial puede ser útil para diseñar más y mejores programas de intervención en universidades y otros centros de estudios superiores (p. e., Cano et al., 2003).

El primer elemento a destacar de los tres modelos de las dimensiones de la tendencia emprendedora es que en todos los casos aparece la puntuación en la escala ACEMP como variable relevante. Este hecho tiene repercusiones tanto a nivel práctico como a nivel teórico. Desde un punto de vista práctico, la validez de la escala ACEMP se ve reforzada, así como su potencialidad de uso como herramienta para la detección precoz de emprendedores potenciales (Cano et al., 2003; García et al., 2005, 2006, 2007; López et al., 2009). Por su parte, el hecho de que la puntuación en nuestra escala de actitudes emprendedoras sea un elemento relevante ante la predicción de la tendencia a crear empresas, se puede considerar un elemento útil ante la definición de emprendedor potencial. Por ello, sería deseable explorar en detenimiento las dimensiones que la componen para evaluar cuales son los factores más útiles en este sentido (Cano et al., 2003).

Los resultados obtenidos en este estudio también son consistentes con otros trabajos orientados a caracterizar al emprendedor ya que también hemos encontrado que la preparación percibida es un predictor importante de la tendencia a crear empresas (p. e., Genescá y Capelleras, 2004; Genesca y Veciana, 1984; Gómez, Mira, y Martinez, 2007; HayGroup y SAP AG, 2003; Rogoff y Lee, 1996; Sánchez, 2003; Veciana, 1989). Así pues, los programas de intervención social destinados a potenciar y valorar positivamente a las personas emprendedoras cobran relevancia en el contexto universitario (p. e., Díaz, 2003; Peñas y Quijano, 2008).

5.- Referencias

Ajzen, I., y Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Englewood, NJ: Prentice-Hall.

Ajzen, I., y Fishbein, M. (2005). The influence of attitudes on behavior. En D. Albarracín, B. T. Hohnson, y M. P. Zanna (Eds.), *The hadnbook of attitudes* (pp. 173–221). Mahwah, NJ: Erlbaum.

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: SAGE University Papers.

Allport, G. W. (1935). Attitudes. En C. Murchison (Ed.), *Handbook of social psychology* (pp. 798–844). Worcester, MA: Clark University Press.

Alonso, D., y Tubau, E. (2002). Inferencias bayesianas: una revisión. *Anuario de Psicología, 33*, 25–47.

Ankarali, H., Canan, A., Akkus, Z., Bugdayci, R., y Ali, M. (2007). Comparison of logistic regression model and classification tree: An application to postpartum depression data. *Expert Systems with Applications, 32*, 987–994.

Bartfay, E., Mackillop, W. J., y Pater, J. L. (2006). Comparing the predictive value of neural network models to logistic regression models on the risk of death for small-cell lung cancer patients. *European Journal of Cancer Care, 15*, 115–124.

Brehm, S. S., Kassin, S., y Fein, S. (2005). *Social psychology* (6 ed.). New York: Houghton Mifflin.

Bull, S. B., Mak, C., y Greenwood, C. M. T. (2002). A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics and Data Analysis, 39*, 57–74.

Cano, C. J., García, J., y Gea, A. B. (2003). *Actitudes emprendedoras y creación de empresas en los estudiantes universitarios*. Almería: Servicio de Publicaciones de la Universidad de Almería / Consejo Social de la Universidad de Almería.

Conati, C., Gertner, A., y VanLehn, K. (2002). Using bayesian networks to manage uncertainty in student modeling. *Modeling and User-Adapted Interaction, 12*, 371–417.

Concato, J., Peduzzi, P., Holford, T. R., y Feinstein, A. R. (1995). Importance of events per independent variable in proportional hazards analysis. I. Background, goals, and general strategy. *Journal of Clinical Epidemiology, 48*, 1495–1501.

Conejo, R., Guzmán, E., Millán, E., Trella, M., Perez-de-la-Cruz, J. L., y Ríos, A. (2004). Siette: A web-based tool for adaptive testing. *International Journal of Artificial Intelligence in Education, 14*, 29–61.

Conejo, R., Millán, E., Perez de la Cruz, J. L., y Trella, M. (2001). Modelado del alumno: un enfoque bayesiano. Inteligencia Artificial, *Revista Iberoamericana de Inteligencia Artificial, 12*, 50–58.

Corman, J., Lussier, R., y Nolan, K. G. (1996). Factors that encourage entrepreneurial start-ups and existing firm expansion: a longitudinal study comparing recession and expansion periods. *Academy of Entrepreneurship Journal, 1*, 43–55.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., y Spiegelhalter, D. J. (1999). *Probabilistic netwoks and expert systems*. Harrisonburg, VA: Springer.

De la Fuente, E. I., García, J., y De la Fuente, L. (2002). Estadística bayesiana en la investigación psicológica. *Metodología de las Ciencias del Comportamiento, 4*, 185–200.

Deaux, K., Dane, C. F., y Wrightsman, L. S. (1993). *Social psychology in the 90s* (6 ed.). Pacific Grove, CA: Brooks/Cole.

Delgado, M. L. (2003). *Aplicación de las redes neurales artificiales a la estadística*. Madrid: Muralla / Hespérides.

DeMaris, A. (2002). Explained variance in logistic regression. A Monte Carlo study of proposed measures. *Sociological Methods & Research, 31*, 27–74.

Díaz, J. C. (2003). *La creación de empresas en extremadura. Un análisis institucional*. Tesis doctoral no publicada, Departamento de Economía Financiera y Contabilidad, Universidad de Extremadura.

Domingos, P., y Pazzni, M. (1996). Beyond independence: conditions for the optimality of the simple bayesian classifier. En L. Saitta (Ed.), *Proccedings of the 13th international conference on machine learning* (pp. 105–112). Bari, Italia: Morgan Kaufman.

Eftekhar, B., Mohammad, K., Ardebili, H. E., Ghodsi, M., y Ketabchi, E. (2005). Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data. *BMC Medical Informatics and Decision Making, 5*, 3.

Feldman, R. S. (1998). *Social psychology* (2 ed.). Upper Saddle River, NJ: Prentice Hall.

Finch, H., y Schneider, M. K. (2007). Classification accuracy of neural networks vs. discriminant analysis, logistic regression, and classification and regression trees. *Methodology, 3*, 47–57.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika, 80*, 27–38.

Franzoi, S. L. (2005). *Social psychology* (4 ed.). New York: Mc Graw Hill.

Freedman, L. S., y Pee, D. (1989). Return to a note on screening regression equations. *The American Statistician, 43*, 279–282.

García, J., Cano, C. J., y Gea, A. B. (2005). Actitudes emprendedoras en estudiantes universitarios y empresarios. Evidencias de validez de un instrumento. *Iberpsicología, 10 (8)*, art. 12.

García, J., López, J., Cano, C. J., Gea, A. B., y De la Fuente, E. I. (2006, Septiembre). *Aplicación de las redes bayesianas al modelado de las actitudes emprendedoras*. Comunicación presentada en el IV Congreso de Metodología de Encuestas. Pamplona.

García, J., López, J., De la Fuente, L., Cano, C. J., y Gea, A. B. (2007, Febrero). *Modelos de ecuaciones estructurales y redes bayesianas. Una perspectiva confirmatoria aplicada a las actitudes emprendedoras*. Comunicación presentada en el X Congreso de Metodología de las Ciencias Sociales y de la Salud. Barcelona.

Gartner, W. B. (1988). "Who is an entrepreneur?" Is the wrong question. *American Journal of Small Business, 12* (4), 11–32.

Genescá, E., y Capelleras, J. L. (2004). Un análisis comparativo de las características de las microempresas en España. *Universia Business Review, 2*, 72–93.

Genesca, E., y Veciana, J. M. (1984). Actitudes hacia la creación de empresas. *Información Comercial Española, 611*, 147–155.

Genkin, A., Lewis, D. D., y Madigan, D. (2005). BBR: Bayesian logistic regression software. Descargado el 7 de Marzo de 2009, desde http://www.stat.rutgers.edu/ madigan/BBR/.

Glymour, C. (2001). *The mind's arrows. Bayes nets and graphical causal models in psychology*. Cambridge, MA: MIT Press.

Glymour, C. (2003). Learning, prediction and causal Bayes nets. *Trends in Cognitives Sciences, 7*, 43–48.

Gómez, J. M., Mira, I., y Martínez, J. (2007). Condicionantes de la actividad emprendedora e instituciones de apoyo desde el ámbito local: el caso de la provincia de Alicante. *Revista de Empresa, 20*, 20–31.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., y Danks, D. (2004). A theory of causal learning in children: causal and bayes nets. *Psychological Review, 111*, 3–32.

Gopnik, A., Sobel, D. M., Schulz, L., y Glymour, C. (2001). Causal learning mechanisms in very young children: two, three, and four-years-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology, 37*, 620–629.

Gopnik, A., y Schulz, L. (2004). Mechanisms of theory formation in young children. *Trends in Cognitives Sciences, 8*, 371–377.

Greiner, R., Su, X., Shen, B., y Zhou, W. (2005). Structural extension to logistic regression: discriminative parameter learning of belief net classifiers. *Machine Learning, 59*, 297–322.

Greiner, R., y Zhou, W. (2002). Structural extension to logistic regression: discriminative parameter learning of belief net classifiers. *Proceedings of the Eighteenth Annual National Conference on Artificial Intelligence, Aug, 2002*, 167–173.

Hair, J. F., Anderson, R. E., Tatham, R. L., y Black, W. C. (1998). *Multivariate data analysis*. Englewood Cliffs, NY: Prentice Hall.

Hanley, J. A., y McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*, 29–36.

Hanley, J. A., y McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology, 148*, 839–843.

Harrel, F. E., Lee, K. E., Matchar, D. B., y Reichert, T. A. (1985). Regression models for prognostic prediction: advantages, problems, and suggested solutions. *Cancer Treatment Reports, 69*, 1071–1077.

Hayek, F. A. (1985). Richard Cantillon. *The Journal of Libertarian Studies, 7*, 217–247.

HayGroup, y SAP AG. (2003). *Factbook, recursos humanos*. Navarra: Aranzadi.

Hébert, R. F., y Link, A. (1989). In search of the meaning of entrepreneurship. *Small Business Economics, 1*, 39–49.

Heckerman, D. (1995). *A tutorial on learning with bayesian networks* (Rep. Téc. MS-TR-95-06). Redmon, WA: Microsoft Research.

Hirji, K. F., Mehta, C. R., y Patel, N. R. (1987). Computing distributions for exact logistic regression. *Journal of the American Statistical Association, 82*, 1110–1117.

Hsieh, F. Y. (1989). Sample size tables for logistic regression. *Statistics in Medicine, 8*, 795–802.

Hsieh, F. Y., Bloch, D. A., y Larsen, M. D. (1998). A simple method of sample size calculation for linear and logistic regresion. *Statistics in Medicine, 17*, 1623–1634.

Huefner, J. C., Hunt, H. K., y Robinson, P. B. (1996). A comparison of four scales predicting entrepreneursihp. *Academy of Entrepreneurship Journal, 1*, 56–80.

Irala, J., Fernández-Crehuet, R., y Serrano, A. (1997). Intervalos de confianza anormalmente amplios en regresión logística: interpretación de resultados de programas estadísticos. *Revista Panamericana de Salud Pública, 1*, 230–234.

Jaimes, F., Farbiarz, J., Alvarez, D., y Martínez, C. (2005). Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room. *Critical Care, 9*, 150–156.

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

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science, 20*, 137–194.

Kadie, C. M., Hovel, D., y Horvitz, E. (2001). *MSBNx: a component-centric toolkit for modeling and inference with bayesian networks* (Rep. Téc. MSTTR- 2001-67). Redmon, WA: Microsoft Research.

King, E., y Ryan, T. P. (2002). A preliminary investigation of maximum likelihood logistic regression versus exact logistic regression. *The American Statistician, 56*, 163–170.

King, G., y Zeng, L. (2001a). Explaining rare events in international relations. *International Organization, 55*, 693–715.

King, G., y Zeng, L. (2001b). Logistic regression in rare events data. *Political Analysis, 9*, 137–163.

Krynski, T. R., y Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General, 136*, 430–450.

Kumar, A., Rao, V. R., y Soni, H. (1995). An empirical comparison of neural network and logistic regression models. *Marketing Letters, 6*, 251–263.

Lee, S. M., Abbott, P., y Johantgen, M. (2005). Logistic regression and Bayesian networks to study outcomes using large data sets. *Nursing Research, 2*, 133–138.

Lilienfeld, D. E., y Pyne, D. A. (1984). The logistic analysis of epidemiologic prospective studies: investigation by simulation. *Statistics in Medicine, 3*, 15–26.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: SAGE Publications.

López, J. (2009). *Modelos predictivos en actitudes emprendedoras: análisis comparativo de las condiciones de ejecución de las redes bayesianas y la regresión logística.* Tesis doctoral no publicada, Facultad de Psicología, Universidad de Almería.

López, J. y García, J. (2010). Technological potential entrepreneurs and optimism. En I. Gómez, D. Martí, e I. Candel. (Eds.), *ICERI2010 Proceedings CD* (pp. 456-461). Valencia: International Association of Technology, Education and Development.

López, J., García, J., Cano, C. J., Gea, A. B., y De la Fuente, L. (2009, Septiembre). *A definition of potential entrepreneur from a probabilistic point of view*. Comunicación

presentada en el XI Congreso de Metodología de las Ciencias Sociales y de la Salud. Málaga.

López, J., García, J., De la Fuente, L., y De la Fuente, E. I. (2007). Las redes bayesianas como herramienta de modelado en psicología. *Anales de Psicología, 23*, 307–316.

López, J., Ruiz-Ruano, A. M., y García, J. (2008, Noviembre). *Relationship between self-assessment and marks in higher education: linear, logistic and bayesian analysis*. Comunicación presentada en la International Conference of Education, Research and Innovation (ICERI 2008). Madrid.

López, J., y García, J. (2007). Valores, actitudes y comportamiento ecológico modelados con una red bayesiana. *Medio Ambiente y Comportamiento Humano, 8*, 159–175.

López, J., y García, J. (2009). Asimetría en el razonamiento causal bayesiano bajo incertidumbre. *Boletín de Psicología, 95*, 43–58.

Martin, J., y VanLehn, K. (1995). Sutdent assessment using bayesian nets. *International Journal of Human-Computer Studies, 42*, 575–591.

Martínez, I., y Rodríguez, C. (2003). Modelos gráficos. En Y. del Águila et al. (Eds.), *Técnicas estadísticas aplicadas al análisis de datos* (pp. 217–257). Almería: Servicio de Publicaciones de la Universidad de Almería.

McKenzie, B., Ugbah, S., y Smothers, N. (2007). "Who is an entrepreneur" is still the wrong question? *Academy of Entrepreneurship Journal, 13*, 23–43.

Mehta, C. R., y Patel, N. R. (1995). Exact logistic regression: theory and examples. *Statistic in Medicine, 14*, 2143–2160.

Mislevy, R. J., y Gitomer, D. H. (1996). The role of probability-based inference in an intelligent tutoring system. *User-Mediated and User-Adapted Interaction, 128*, 253–282.

Morales, J. F., Rebolloso, E., y Moya, M. (1994). Actitudes. En J. F. Morales (Ed.), *Psicología social* (pp. 495–524). Madrid: McGraw-Hill.

Morales, M. E. (2006). *Modelización y predicción en estadística universitaria.* Tesis doctoral no publicada, Facultad de Ciencias Experimentales, Universidad de Almería.

Narayan, S., y Jurafsky, D. (1998, Agosto). *Bayesian models of human sentence processing*. Comunicación presentada en la XX Annual Meeting of the Cognitive Science Society. Madison.

Narayan, S., y Jurafsky, D. (2002). A bayesian model predicts human parse preference and reading times in sentence processing. *Advances in Neural Information Processing, 14*, 59–65.

Ng, A. Y., y Jordan, M. I. (2002). On discriminative vs. generative classifiers: a comparation of logistic regresion and naive Bayes. *Advances in Neural Information Processing Systems, 14,* 841–848.

Ortega, M., y Cayuela, A. (2002). Regresión logística no condicionada y tamaño de muestra: una revisión bibliográfica. *Revista Española de Salud Pública, 76*, 85–93.

Ortiz, A. J., Martín, M. T., Ureña, L. A., y García, M. A. (2005). Detección automática de SPAM usando regresión logística bayesiana. *Procesamiento del Lenguaje Natural, 35*, 127–133.

Peduzzi, P., Concato, J., Feinstein, A. R., y Holford, T. R. (1995). Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *Journal of Clinical Epidiemiology, 48*, 1503–1510.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., y Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidiemiology, 49*, 1373–1379.

Peñas, M. J., y Quijano, J. (2008, Abril). *¿Es posible fomentar el carácter emprendedor desde la universidad? Un diseño de la asignatura «Empresa Familiar».* Comunicación presentada en el Congreso Internacional de Emprendedores Ciudad de Salamanca. Salamanca.

Peng, C. Y. J., y So, T. S. H. (2002). Logistic regression analysis and reporting: a premier. *Understanding Statistics, 1*, 31–70.

Quinlan, P. (1991). *Connectionism and psychology: a psychological perspective on new connectionist research*. Hertfordshire: Cambridge University Press.

Rogoff, E. G., y Lee, M. S. (1996). Does firm origin matter? An empirical examination of types of small business owners and entrepreneurs. *Academy of Entrepreneurship Journal, 1*, 1–17.

Ruiz, J., Rojas, A., y Suárez, A. (2008). Actitudes de los estudiantes universitarios de Andalucía ante la creación de empresas. Cádiz: Servicio de Publicaciones de la Universidad de Cádiz.

Samuelson, P. A. (1970). *Economics* (8ª ed.). New York: McGraw-Hill.

Sánchez, M. L. (2003). *El perfil psicológico del autoempleado.* Tesis doctoral publicada en edición electrónica, Facultad de Psicología, Universidad Complutense de Madrid.

Serrano, J. (2003). *Iniciación a la estadística bayesiana*. Madrid: Muralla / Hespérides.

Shen, B., Su, X., Greiner, R., Musilek, P., y Cheng, C. (2003, Noviembre). *Discriminative parameter learning of general bayesian network classifiers*. Comunicación presentada en la 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI-03). Sacramento, California.

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

SPSS, y Recognition Systems. (1997). *Neural conection 2. 0. User's guide*. Chicago, IL: SPSS y Recognition Systems.

Terrin, N., Schmid, C. H., Griffith, J. L., D'Agostino, R., y Selker, H. P. (2003). External validity of predictive models: A comparison of logistic regression, classification trees, and neural networks. *Journal of Clinical Epidemiology, 56*, 721–729.

Veciana, J. M. (1989). Características del empresario en España. *Papeles de Economía Española, 39*, 19–36.

Walker, S. H., y Duncan, D. B. (1967). Estimation of the probability of an event as function of several independent variables. *Biometrika, 54*, 167–179.

Whittermore, A. S. (1981). Sample size for logistic regression with small response probability. *Journal of American Statistical Association, 76*, 27–32.