MODELO PREDICTIVO DE RIESGO DE MOROSIDAD PARA CRÉDITOS BANCARIOS USANDO DATOS SIMULADOS

CLAUDIO ALFREDO LÓPEZ MIRANDA

Un problema grave en las instituciones de crédito son los préstamos no recuperados, por lo que al momento de la solicitud, es deseable un modelo matemático para predecir el riesgo de tener un cliente potencialmente moroso, ya que su capacidad económica no garantiza por si sola el pago. Este trabajo presenta, a través de datos simulados, una técnica estadística conocida como regresión logística para predecir la probabilidad de no recuperar un crédito, con base al perfil socioeconómico del cliente, considerando el número de dependientes económicos, número de impagos previos, su empleo, destino del crédito y salario mensual. El modelo predictor logra una tasa aciertos de 95% para clientes morosos y una tasa global del 70% de clasificaciones correctas. Los resultados sirven para apreciar los indicadores estadísticos conocidos, como la razón de ventajas por categorías. Por ejemplo, por cada crédito previo no pagado del cliente, se estima que el riesgo es 3.7 veces mayor y por cada $1,000.00 de incremento en el salario se estima que el riesgo (multiplicativo) disminuye por un factor del 80%.

DR. CLAUDIO ALFREDO LÓPEZ MIRANDA
Universidad de Sonora, Departamento de Matemáticas
Correo: claudio@gauss.mat.uson.mx

*Autor para correspondencia: Claudio Alfredo López Miranda
Correo electrónico: claudio@gauss.mat.uson.mx
Recibido: 12 de marzo de 2013
Aceptado: 11 de junio de 2013
ISSN: 2007-4530

INTRODUCCIÓN

La regresión logística ha sido aceptada durante las últimas décadas como modelo de predicción-clasificación desde sus orígenes en investigaciones epidemiológicas y clínicas (1). Hoy en día se emplea comúnmente, aunque no limitada a, investigaciones biomédicas, negocios y finanzas, criminología, ingeniería, salud pública, política, bilogía de la vida salvaje y psicología. Las aplicaciones van desde créditos no recuperados (2), votación en elecciones políticas (5), asignación de una beca, hasta predicción de cáncer de próstata (4), riesgo de contagio de VIH (4), etcétera.

La regresión logística es una técnica de estadística multivariada basada en los principios de regresión lineal, pero a diferencia de ésta, la regresión logística utiliza una variable dependiente categórica binaria o dicotómica (en vez de una variable cuantitativa continua) cuyo resultado o valor de respuesta es 0 o 1; por ejemplo, en créditos bancarios se asigna 1 si los datos observados provienen de un cliente moroso y 0 en caso contrario. Debido a que la morosidad es una variable dicotómica que depende como variable de respuesta de otros indicadores tales como el salario y las cargas familiares, entre otros, es factible aplicar la regresión logística como modelo clasificador. Además, el modelo puede utilizarse para estimar la probabilidad de que un cliente no devuelva el crédito; a dicha probabilidad se le conoce como riesgo de morosidad. El status de morosidad se categorizará dependiendo si la probabilidad estimada, evaluada con los datos del cliente, es mayor que un punto de corte, el cual debe prefijarse por cada institución particular con base en su experiencia, por ejemplo: Pc = 0.30, o bien usar una técnica formal para seleccionarlo (3).

EL MODELO DE REGRESIÓN LOGÍSTICA

El modelo de regresión logística surge cuando queremos estimar la probabilidad de un evento dicotómico de Si (1) o No (0), en función de un conjunto de variables predictoras comúnmente llamados factores de riesgo, que pueden ser discretas o continuas, categóricas (nominales u ordinales), cualitativas o cuantitativas. En ocasiones la variable de respuesta *Y*, o variable dependiente, tiene más de dos categorías, para la que existen otras técnicas no tratadas aquí (1) y (3). Estos modelos son especialmente útiles cuando de manera natural o controlado experimentalmente, la probabilidad de ocurrencia del evento de interés (etiquetado con un 1) sigue una forma-S o **sigmoidal**, por ejemplo, en casos reportados de estudios clínicos, (1) y (3), el riesgo *p* de padecer hipertensión típicamente se comporta en la forma-S conforme aumenta la edad *x* (*i.e.*, hasta cierta edad el riesgo es mínimo, luego conforme avanza la edad el riesgo crece paulatinamente hasta estabilizarse en un valor cercano a uno después de ciertos años); así mismo ocurre típicamente o de manera aproximada con el riesgo de no pagar un crédito conforme aumenta la deuda previa del solicitante (forma-S creciente) o al aumentar su ingreso (forma-S decreciente). Cuando el comportamiento no corresponde de manera aproximada a una forma-S, se recomienda utilizar otras técnicas distintas a la regresión logística, por ejemplo, árboles de clasificación o los modelos de riesgo aditivo (4).

Para adentrarnos en el conocimiento de la técnica de regresión logística, utilizaremos como referencia los conceptos de la técnica de regresión lineal, ya que la regresión logística utiliza una metodología muy similar, excepto que no son necesarios supuestos de normalidad en la variable de respuesta ni en los residuos (lo cual es una de sus ventajas). Aunque ambas técnicas se basan en un ajuste lineal, la regresión logística utiliza primero una transformación de la variable de respuesta en unidades logarítmicas para deducir el modelo de regresión lineal final. El objetivo es estimar la probabilidad de ocurrencia $p(\boldsymbol{x})$ de un evento, dado un conjunto de covariables predictoras $\boldsymbol{x} = (x_1, x_2, \cdots, x_k)$, ajustadas o controladas a ciertos valores particulares. El modelo matemático de forma-S comúnmente utilizado es la conocida función logística (para más detalles ver las citas (1)-(4)):

$$p(\boldsymbol{x}) = \frac{e^{f(x)}}{1 + e^{f(x)}} = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 \quad + \quad e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}} = \frac{1}{1 + e^{-\beta_0 - \sum_{i=1}^{k} \beta_i x_i}}$$

Una transformación sencilla en términos de lo que se conoce como razón de ventajas, expresada como el cociente de la probabilidad $p(x)$ de ser moroso contra la probabilidad de ser cumplido 1- $p(x)$, nos lleva al modelo de regresión en unidad logarítmica o función **logit** (del inglés log-unit):

$$\text{logit} \Rightarrow ln\left[\frac{p(x)}{1 - p(x)}\right] = f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

El cual es ahora un modelo lineal para el que se deben estimar los parámetros βs mediante la técnica de máxima verosimilitud, (1)-(5), en vez de utilizar la técnica de mínimos cuadrados de la regresión lineal. Aunque, cabe mencionar que en el caso particular de la regresión lineal ambas técnicas producen las mismas estimaciones (1). El término entre corchetes en la ecuación anterior, se conoce como la razón de ventaja entre la probabilidad de ocurrencia de un éxito con respecto a la de fracaso, dado un valor de $\boldsymbol{x}$ particular. Esta razón de ventajas (***odds*** de la literatura en inglés) es muy importante ya que nos ayuda a entender la relación entre clientes sujetos o no a un factor de riesgo, como puede ser tener o no dependientes económicos. Por ejemplo, en el caso simple de una variable dependiente predictora dicotómica, x = tiene dependientes económicos, denotando con $x = 1$ para "si" y $x = 0$ para "no"; calculamos la razón de ventajas con ambos valores de x, al dividir ambos **odds** obtenemos lo que se conoce como el ***odds ratio OR*** (que representa la proporción en la que la razón de ventaja de estar expuesto a un factor supera a la de no estar expuesto). Por ejemplo, si el $OR \triangleq \frac{\boldsymbol{odds}(x=1)}{\boldsymbol{odds}(x=0)} = 2$, significa que un cliente con cargas familiares ($x = 1$) tiene 2 veces más riesgo de no pagar el crédito que un cliente que no las tiene ($x = 0$). La interpretación también es válida en estudios

clínicos o epidemiológicos, donde, por ejemplo, se puede decir que el riesgo de contraer el VIH es dos veces mayor en personas drogadictas que las que no lo son. Lo interesante del modelo logístico estriba en que las estimaciones de los coeficientes del modelo de regresión (las βs) sirven a su vez para estimar las ***odds ratios*** precisamente con la transformación logaritmo inversa, tal que $\widehat{OR}_i = e^{\widehat{\beta}_i}$, (esta igualdad es fácil verificarla directamente de la definición anterior del ***OR***). Así también para obtener los intervalos de confianza de las ***ORs***, se calculan primero los intervalos de confianza de los parámetros βs usando el estadístico Wald, el cual tiene una distribución aproximadamente normal bajo ciertas condiciones, no discutidas aquí ya que utilizaremos los resultaos directos del paquete SPSS, ver (1) y (3); después, mediante exponenciación de los límites de los intervalos obtenidos calculamos los intervalos para las ***ORs***.

MODELO DE RIESGO DE MOROSIDAD PARA CRÉDITOS BANCARIOS

Dada la dificultad de vinculación con el sector empresarial para utilizar datos reales, presentamos un caso de estudio simulado sobre créditos bancarios, el cual intenta ampliar los resultados y análisis reportados para regresión logística en (2), ya que los autores comparan la capacidad predictora de un modelo empírico logístico frente a uno de análisis discriminante (http://www3.unileon.es/pecvnia/pecvnia01/01_175_199.pdf). Nuestro objetivo es determinar que otros factores, además de los reportados en (2), caracterizan la morosidad en las entidades financieras, así como incluir variables adicionales de interés práctico como el *ingreso*, el cual sorprende no se contempla como variable significativa en (2).

Nuestra propuesta incluye las cuatro variables reportadas con significancia estadística, a saber, la variable dependiente dicotómica Y, *estatus de la morosidad* (Y = 1 para moroso, Y = 0, no moroso); y como variables independientes o predictoras: *Destino del crédito* (variable categórica: 0 = "traspaso de negocio"; 1 = "compra de automóvil"; 2 = "otros" el cual incluye compra de vivienda, remodelación, pago de deudas, compra de negocio o activos para actividad profesional y otros fines); *Nuevo residente en la localidad* (1 = si, 0 = no); *Número de impagos anteriores* (0, 1, o 2) como variable discreta de razón. A diferencia de los autores de (2) y buscando tener un modelo más realista, incluimos tres variables adicionales consideradas de interés práctico, pero que fueron excluidas por los autores en (2) basados en la poca significancia estadística. Es bien sabido que la capacidad de pago está íntimamente relacionada con el *Ingresos del cliente*, el *número de dependientes económicos* (variable discreta con valores 0, 1, 2, 3, 4, 5 o más) y su *tipo de empleo*, por lo que son agregadas al modelo.

A continuación utilizaremos la técnica de regresión logística progresiva del paquete SPSS (5) para analizar tanto la significancia estadística mediante la prueba Wald como la prueba de de Hosmer-Lemeshow, para analizar qué tan bien se ajustan el modelo a los datos. Una vez que se obtiene la estimación de parámetros, se puede usar el modelo para medir su capacidad predictora, la interpretación de las ***ORs*** e intervalos de confianza. Para más detalles de justificación y uso de variables ver (2).

Para seleccionar la muestra hipotética de clientes, es importante considerar los tipos de entidades financieras como bancos, cajas de ahorros, cooperativas de crédito. Se recomienda un muestreo por conglomerados en dos etapas donde las unidades primarias sean las entidades y las unidades últimas los clientes (nuestro estudio presupone tal diseño de muestreo). Además, utilizaremos 100 casos en vez de los 72 reportados en (2); es importante notar que para la variable *empleo,* con nueve categorías el tamaño de muestra resulta muy pequeño para confiar estimar sus ***ORs***, por lo que se debe tener cuidado con los resultados en casos reales, verificando que los resultados sean consistentes antes de asumirlos como válido, ya que para un usuario con poca experiencia podría pasar desapercibido (4). Asimismo, se debe utilizar un tamaño de muestra suficiente para garantizar que todas las categorías tienen una frecuencia de casos observados distinta de cero, para evitar divisiones con denominadores nulos al calcular los **odds**, en particular la categoría de referencia. Si este no fuera el caso, se recomienda fundir varias categorías adyacentes o similares en algún sentido práctico. Por cierto, esto es una de las posibles desventajas al aplicar regresión logística (4).

GENERACIÓN DE LA BASE DE DATOS MEDIANTE SIMULACIÓN

La base de datos utilizada (ver apéndice) fue simulada mediante el paquete SPSS. Cabe mencionar que dicha base sirve como conjunto de entrenamiento y prueba para generar el modelo, por lo que la variable morosidad es generada desde un principio para luego contrastarla con el modelo clasificador-pronosticador, una vez que se estiman las probabilidades o riesgo de morosidad. En este sentido, se da por sentado que la base de datos captura una supuesta relación intrínseca entre la variable de respuesta

y los valores de las covariables, los cuales usaremos para generar el modelo inmerso.

Las variables fueron generadas de acuerdo a los siguientes criterios:

y = Morosidad: Se genera una muestra de números aleatorios con distribución Bernoulli de parámetro p = 0.26 que corresponde a la probabilidad de morosidad reportada en (2).

x_1 = *Destino:* Números aleatorios con distribución binomial de parámetros $n = 2$ y p = 0.4. Donde 0 = Traspaso de negocio, 1 = compra de automóvil 2 = otras inversiones (casa, local, activos, pago de deudas y otros gastos). Es posible utilizar un esquema discreto distinto a la binomial más realista, sin embargo, no hay pérdida de generalidad usando la binomial.

x_2 = *Nuevo Residente:* Variable 0 - 1; muestra de números aleatorios con distribución Bernoulli con p = 0.35, suponiendo de manera empírica que el 35% son nuevos residentes.

x_3 = *Número Impagos:* Variable aleatoria (v. a.) discreta con valores enteros 0, 1 y 2; más de tres impagos suponemos rechazada automáticamente la solicitud de crédito. Se genera mediante una binomial (n = 2, éxito = número de impagos en su dos operaciones anteriores, con probabilidad general de impago p = 0.08). Dicho valor del 8% sólo es de referencia para generar la base, no confundir con la probabilidad de riesgo que buscamos predecir.

x_4 = *Salario:* v. a. continua con distribución normal (en miles de pesos) con media 15 y desviación estándar 3. "Promedios" más o menos realistas de ingresos mensuales actuales para un trabajador de clase media-alta.

x_5 = *Empleo:* v. a. con 9 categorías, generadas usando Excel con una distribución multinomial de acuerdo a la siguiente tabla de probabilidades propuestas.

Tabla 1. Probabilidades por categorías para la variable *Empleo*

Número de categoría	Probabilidad	Categoría de empleo
1	0.20	Cuenta propia
2	0.20	Empelado ejecutivo
3	0.24	Empleado indefinido
4	0.14	Empleado eventual
5	0.02	Cuenta propia y empleado ejecutivo
6	0.01	Cuenta propia y empleado indefinido
7	0.05	cuenta propia y empleado eventual
8	0.04	Jubilado o prejubilado
9	0.10	Desempleado

x_6 = *Cargas familiares*: variable discreta generada como una v.a. binomial con n = 5 p = 0.20, que representa la probabilidad de tener un hijo. Muchos casos resultaron con 0, 1 ó 2 hijos; lo que concuerda para clientes jóvenes solteros (29%, carga = 0); en su mayoría casados 47% (carga = 1), con un hijo (carga = 2) y no más de dos hijos (carga = 3).

Se debe notar que *cargas familiares* se genera como variable discreta pero simbólicamente se emplea en el modelo como variable continua (en escala de razón) para analizar el aumento del riesgo por cada dependiente económico adicional (incremento unitario), en vez de comparar categorías por cantidad de hijos; o con cargas o sin cargas (dicotómica). Así también el *número de impagos* se emplea internamente como variable continua. Además, la variable *destino del crédito* es de varias categorías y sólo fueron significativas "traspaso de negocio" o "compra de automóvil". Es importante mencionar que los autores en (2) tomaron las categorías como variables independientes, lo cual no resulta congruente al estudiar el modelo. Las categorías de remodelación o compra de vivienda, compra de local o activos para actividad profesional, pagar deudas y otros gastos no fueron variables significativas y se agrupan en la categoría de "otros".

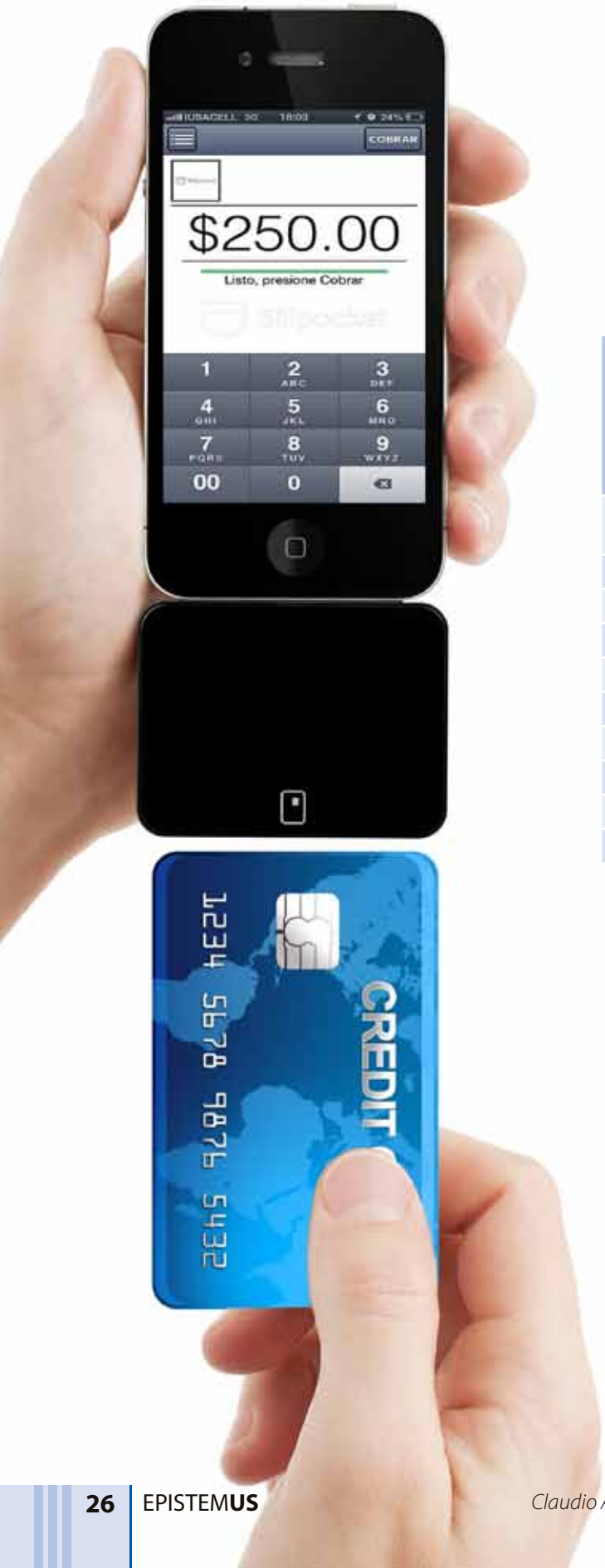
DISCUSIÓN DE RESULTADOS

La base de datos del apéndice fue procesada con SPSS (5); las variables con más de dos categorías (*empleo* y *destino*) fueron codificadas con "celda de referencia" (3), usando 1 para dicha categoría y cero para el resto; la categoría de referencia fue la última y surge cuando todas las demás categorías toman el valor cero. Lo primero es una evaluación general del modelo mediante la prueba *–2Log de la Verosimilitud*, la cual se utiliza para decidir si es el modelo es confiable estadísticamente hablando para predecir la variable de respuesta. El SPSS utiliza como hipótesis nula que el modelo es significativo y, como hipótesis alternativa lo contrario, asumiendo que el estadístico *-2Log* tiene una distribución chi cuadrada con k –2 grados de libertad. El estadístico *-2Log* resultó 82,880 con 14 grados de libertad y un nivel de significancia de 0.246, por lo tanto, no rechazamos la hipótesis nula de que el modelo logístico en general es significativo como modelo predictor. Analicemos ahora el ajuste de los datos.

La tabla 2 contiene la bondad de ajuste, la cual resulta muy significativa 55.9%. Lo anterior se aprecia en la tabla 3 que muestra la diferencia entre las frecuencias observadas y esperadas muy aproximadas, agrupando los valores de probabilidad estimada en 10 clases percentiles. Para descripción de la prueba de bondad ver (1) y (4).

Tabla 2. Prueba de bondad de ajuste Hosmer y Lemeshow.

	gl	Sig.
6,792	8	0.559



La tabla 4 muestra la clasificación de los casos observados contra los casos pronosticados, dichos resultados se pueden obtener comparando en la tabla del apéndice la primera columna (clase observada), contra la última columna que representa la probabilidad *p* de ser clasificado como moroso si dicho valor es mayor que el punto de corte utilizado, Pc = 0.15. El modelo tuvo una tasa total de pronósticos correctos del 70%, es decir, los pronósticos tanto de morosos como no morosos clasificados de manera correcta con respecto a su clase observada.

Tabla 3. Prueba de Hosmer y Lemeshow para frecuencias observadas y esperadas.

		Si el crédito fue pagado o no = No moroso		Si el crédito fue pagado o no = Moroso		Total
		Observado	Esperado	Observado	Esperado	
Paso	1	10	9.819	0	0.181	10
	2	10	9.556	0	0.444	10
	3	9	9.319	1	0.681	10
	4	10	9.146	0	0.854	10
	5	10	8.852	0	1.148	10
	6	7	8.226	3	1.774	10
	7	7	7.521	3	2.479	10
	8	5	6.909	5	3.091	10
	9	6	6.117	4	3.883	10
	10	6	4.536	4	5.464	10

Tabla 4. Clasificación-predicción del modelo.

Observado	Pronosticado		Porcentaje correcto
	No moroso	Moroso	
No moroso	51	29	63.8
moroso	1	19	95.0
Porcentaje global			70.0

La tasa de clasificación de verdaderos negativos (no morosos pronosticados como no morosos) fue del 63%; mientras que la tasa más relevante en este caso, clasificar correctamente a los clientes morosos (verdaderos positivos) fue de 95%, demostrando gran capacidad predictiva del modelo. Cabe mencionar que el Pc se eligió con base en la experiencia de la simulación, sin embargo, formalmente se puede seleccionar usando la llamada curva de características de operación o curva **ROC** (*Receiver Operating Characteristic*), consultar (1) y (3).

Tabla 5. Estadísticas para el modelo de regresión logística.

	B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95% para EXP(B)	
							Inferior	Superior
Destino			1.602	2	0.449			
Destino (1)	0.897	0.782	1.316	1	0.251	0.408	0.088	1.888
Destino (2)	0.715	0.687	1.086	1	0.297	0.489	0.127	1.878
Nuevo residente	1.096	0.687	2.543	1	**0.111**	0.334	0.087	1.285
No. de impagos	1.307	0.568	5.294	1	**0.021**	3.696	1.214	11.255
Salario	0.217	0.098	4.872	1	**0.027**	0.805	0.664	0.976
Cargas	0.207	0.334	0.385	1	0.535	1.231	0.639	2.370
Empleo			1.882	8	0.984			
Empleo (1)	0.483	1.052	0.211	1	0.646	0.617	0.078	4.850
Empleo (2)	0.507	1.084	0.219	1	0.640	0.602	0.072	5.044
Empleo (3)	0.540	1.104	0.240	1	0.625	0.583	0.067	5.069
Empleo (4)	0.318	1.080	0.087	1	0.769	1.374	0.166	11.400
Empleo (5)	19.286	25577.4	0.000	1	0.999	0.000	0.000	
Empleo (6)	0.939	1.567	0.359	1	0.549	2.557	0.119	55.122
Empleo (7)	18.062	40192.9	0.000	1	1.000	0.000	0.000	
Empleo (8)	0.527	1.595	0.109	1	0.741	0.590	0.026	13.452
Constante	2.402	1.719	1.952	1	0.162	11.043		

Nota. El número entre paréntesis en el nombre de cada variable indica su categoría.

La tabla 5 contiene en la columna B la estimación de los coeficientes βs del modelo y el estadístico de prueba en la columna Wald, el cual es un indicador para probar la significancia (Sig.) estadística de cada variable (1), en este caso se utilizó un nivel de significancia del 5%. Se observa en la sexta columna que las variables significativas fueron *número de impagos previos* y el *salario*; mientras que la variable *nuevo residente* es significativa al 12%. La variable *empleo* no resulta significativa en ninguna de sus nueve categorías. Cabe mencionar que se hicieron pruebas eliminando estas variables del modelo, sin embargo, su desempeño fue más pobre, por lo que se mantuvieron. La tabla 5 también incluye las estimaciones de interés de las ***OR*s** y cuya interpretación se da más abajo; además, se incluyen sus intervalos de confianza con propósitos ilustrativos; para una discusión más amplia ver (1), (3) y (4). Observe que los coeficientes βs negativos implican un OR menor a 1 como indicador de disminución del riesgo a medida que aumenta la variable predictora; un coeficiente β positivo mucho mayor que cero no influirá en el cálculo de la probabilidad, mientras que un valor negativo mucho menor que cero producirá por si sola una probabilidad estimada aproximadamente cero. Lo anterior debido al cambio de signo en la función **logit**.

Si usamos los coeficientes estimados de la tabla 5, columna B, y los valores de las variables para cada caso en la base de datos simulada, podemos estimar la probabilidad o riesgo de morosidad de cada cliente mediante el modelo de regresión logística. Por ejemplo, los cálculos para el caso 1 muestran el valor de predicción de riesgo de morosidad siguiente:

Caso	Morosidad	Destino	Nuevo residente	Número de impagos	Salario	Empleo	Cargas	Predicción
X_1	0	0	0	0	18.047	2	1	0.0623
X_2	1	0	1	0	12.040	1	1	0.0773

Predicción de riesgo de morosidad para el caso 1:

$$P(\boldsymbol{X_1}) = \frac{1}{1 + \exp\left[-(2.402 - 0.897 \times Destino(1) - 0.217 \times Salario - 0.507 \times Empleo(2) + 0.207 \times Cargas)\right]}$$

$$P(\boldsymbol{X_1}) = \frac{1}{1 + \exp\left[-(2.402 - 0.897 \times 1 - 0.217 \times (18.047) - 0.507 \times 1 + 0.207 \times 1)\right]} = 0.0623$$

Observe que el caso 1 representa datos de un cliente no moroso, cuyo destino del crédito fue para compra de automóvil, no es nuevo residente, no tiene impagos previos, tiene un salario elevado y con una carga familiar, el cual podríamos decir que es un perfil típico de un buen cliente. Observe que la probabilidad estimada por el modelo resultó menor que el Pc = 0.15, por lo que este primer caso es clasificado de manera correcta como cliente no moroso. Es necesario aclarar que en la variable *destino* la codificación de categoría de referencia es la última, en este caso la tercera categoría "otros"; así, la categoría 0 representa a "traspaso de negocio" mientras que la categoría 1 representa "compra de automóvil", esto explica por qué se usó el se usó el valor uno para indicar la variable destino(1) en el cálculo anterior. Las probabilidades restantes se calculan de manera similar y se muestran en la columna 8 del apéndice. Debido a que la probabilidad estimada para los primeros dos casos es menor que el Pc = 0.15, estos serán pronosticados no morosos, sin embargo, según los datos del apéndice, observe que el segundo caso corresponde en realidad a un cliente moroso, por lo que la clasificación del modelo para este caso es incorrecta, de hecho es el único "caso extremo-moroso" presente en los datos simulados.

El valor 3.696 en la columna Exp (B) nos dice que por cada crédito impagado (incremento unitario) el riesgo de no devolver el crédito aumenta 3.696 veces. Asimismo, vemos que por cada $ 1,000 de incremento (unitario) en el salario, el riesgo (multiplicativo) de no pagar el crédito disminuye un 80.5%; es decir, si un cliente con un salario determinado presenta un riesgo del 50%, si mantenemos fijas todas las variables pero incrementamos el salario en $1,000, el riesgo (multiplicativo) se reduce aproximadamente al 40%; aún más, si el salario aumenta en incrementos de $ 5,000, el riesgo se reduce hasta un $e^{5\times(-0.217)} \times 100 \approx 34\%$. También vemos una correlación negativa en el renglón de nuevo residente, por lo tanto, el $\widehat{OR} = 0.334$ significa que los nuevos residentes (categoría 1) tienen una disminución del riesgo por un factor de 33.4% comparado con un residente antiguo.

CONCLUSIONES

Se mostró que un modelo de regresión logística para pronosticar la morosidad de un crédito bancario se desempeña adecuadamente como modelo clasificador sobre una base de datos simulada; en consecuencia, se asume factible que este tipo de modelos sea adecuado sobre casos reales o similares. El estadístico *– 2Log de Verosimilitud* mostró que en general el modelo es un buen predictor; además, se mostró que se ajusta bien a los datos, tanto en las frecuencias observadas contra las esperadas. Se obtuvo un excelente desempeño del 95% para predecir particularmente posibles clientes morosos. La tasa de predicción global del modelo fue buena de acuerdo a la literatura estándar, ya que se ajusta a un 70% de clasificaciones correctas. Las variables más significativas para pronosticar el riesgo de morosidad son *número de impagos previos*, *salario* y *nuevo residente*. Sorprende que la variable número de dependientes económicos o *carga familiar* resultó no significativa. Es importante recalcar que el modelo fue contrastado con los datos simulados, la metodología aquí presentada se puede ver como referencia para un trabajo futuro más realista, en tal caso, se debe tener cuidado en validar primero la base de datos mediante el estudio de las distribuciones adecuadas, así como validación final del modelo, o mejor aún, usar una base de datos reales.

BIBLIOGRAFÍA

1) Hosmer, D.W. and Lemeshow, S. (2000). *Applied Logistic Regression.* Second edition. Wiley series in probability and statistics.
2) Mures, Q. J., García G., A. y Vallejo P., E. (2005). *Aplicación del análisis discriminante y regresi*ón logística en el estudio de la morosidad en las entidades financieras. Comparación de resultados. Pecnia, vol. 1, pp. 175–199. Recuperado el 25 de enero de 2013. http://www3.unileon.es/pecvnia/pecvnia01/01_175_199.pdf
3) Kleinbaum, D.G. and Klein, M. (2010). *Logistic Regression a Self-Learning Text*. Third edition. Springer.
4) Vittinghoff, E., Shiboski, S., Glidden, D. and McCulloch, Ch. (2004). *Regression methods in biostatistics: Linear, Logistic, Survival, and Repeated Measures models.* Springer.
5) Pérez L, C. (2009). *Técnicas de Análisis de datos con SPSS*. Pearson Prentice Hall.

APÉNDICE: Base de datos simulados para el modelo de riesgo para créditos bancarios

La siguiente tabla contiene la base de datos para 100 clientes obtenida mediante simulación para generación del modelo; dichos datos pueden ser recuperados del repositorio ftp.mat.uson.mx/~claudio o replicando el experimento descrito en la sección de simulación, por lo que pueden omitirse (se incluyen para propósitos de revisión). El significado de cada variable $x_1,\ldots, x_6$, en esta tabla es como se definió también en la sección de simulación.

CASO	y	x_1	x_2	x_3	x_4	x_5	x_6	p
1	0	0	0	0	18.047	2	1	0.062
2	1	0	1	0	12.040	1	1	0.077
3	1	2	0	0	16.237	3	0	0.159
4	1	0	1	2	11.670	8	0	0.491
5	0	2	0	0	11.674	8	1	0.389
6	0	0	0	0	17.786	9	0	0.087
7	0	1	1	0	12.591	1	1	0.082
8	0	1	1	0	19.416	3	2	0.023
9	0	1	1	0	18.149	8	1	0.025
10	0	1	0	0	18.223	1	1	0.073
11	0	1	0	1	18.504	2	1	0.210
12	1	0	0	0	13.499	1	3	0.217
13	0	0	0	0	17.870	1	2	0.080
14	0	1	1	0	15.187	1	2	0.059
15	1	2	0	1	15.064	2	1	0.535
16	0	0	1	1	16.303	1	2	0.131
17	0	1	1	0	11.882	2	2	0.111
18	0	1	0	0	14.164	2	0	0.131
19	0	1	0	0	19.224	3	1	0.056
20	0	2	1	0	14.898	2	1	0.097
21	0	2	1	2	16.009	2	3	0.637
22	0	1	0	1	14.273	2	1	0.401
23	0	2	1	0	17.359	9	1	0.095
24	0	2	1	0	14.648	3	2	0.120
25	0	2	0	0	14.670	3	2	0.288
26	0	2	0	0	18.162	2	1	0.137
27	0	1	1	0	14.665	9	0	0.070
28	0	0	0	0	12.902	6	0	0.412
29	0	0	1	0	14.998	9	1	0.067
30	0	0	0	0	16.723	2	3	0.118
31	0	1	1	0	9.532	9	1	0.219
32	0	0	0	0	19.690	1	1	0.046
33	0	0	0	0	19.003	4	1	0.110
34	0	1	1	0	18.157	7	1	0.000
35	0	1	1	0	12.634	9	0	0.104
36	0	2	0	0	13.190	3	1	0.312
37	1	1	0	0	10.977	2	1	0.270
38	0	1	1	0	15.974	4	0	0.072
39	0	1	0	0	8.435	2	1	0.391
40	1	0	0	1	19.495	4	1	0.291
41	0	1	0	0	10.081	1	2	0.362
42	1	2	1	1	12.900	1	1	0.387
43	0	1	1	0	11.973	1	0	0.077
44	1	1	0	0	11.791	2	3	0.319
45	1	1	0	0	9.493	3	1	0.331
46	0	1	0	0	16.614	3	0	0.079
47	0	0	0	0	21.478	2	0	0.025
48	0	0	1	0	16.410	9	0	0.041
49	0	1	0	0	17.417	1	1	0.086
50	0	0	1	0	11.378	2	1	0.086

CASO	y	x_1	x_2	x_3	x_4	x_5	x_6	p
51	0	1	0	0	15.738	1	1	0.119
52	0	2	0	0	13.236	9	1	0.435
53	0	1	0	0	20.430	1	0	0.038
54	0	0	0	0	18.996	2	2	0.063
55	0	1	1	0	15.202	4	0	0.084
56	1	2	1	0	13.897	4	0	0.200
57	1	0	0	0	15.180	6	2	0.393
58	0	2	0	1	14.107	4	0	0.724
59	0	1	1	0	18.942	1	1	0.022
60	0	2	1	0	11.621	1	2	0.217
61	0	1	1	0	14.255	4	1	0.122
62	1	2	0	0	8.636	3	1	0.549
63	0	1	0	0	16.943	1	0	0.078
64	0	0	1	0	14.780	3	0	0.034
65	0	1	0	0	9.368	3	0	0.292
66	0	1	0	0	10.685	4	1	0.473
67	0	0	1	0	16.899	3	3	0.040
68	0	1	0	0	11.260	1	1	0.263
69	1	1	0	0	15.588	9	1	0.184
70	1	1	0	0	15.846	4	1	0.227
71	0	2	0	0	15.867	3	1	0.202
72	0	0	0	0	12.786	5	1	0.000
73	0	1	0	0	18.653	6	0	0.195
74	0	1	0	0	16.254	2	1	0.105
75	0	0	1	0	17.215	1	2	0.033
76	0	1	1	0	14.737	8	2	0.062
77	0	2	0	0	14.829	1	1	0.252
78	0	1	1	2	13.515	1	1	0.500
79	0	1	0	0	9.134	4	0	0.506
80	1	1	0	0	13.026	4	0	0.305
81	0	0	1	0	20.691	9	1	0.020
82	1	1	0	0	11.900	9	2	0.382
83	0	0	0	0	10.958	9	0	0.295
84	0	2	1	0	17.839	3	0	0.043
85	0	1	1	0	16.289	3	1	0.036
86	0	0	0	0	9.525	4	3	0.594
87	0	0	1	1	9.693	3	1	0.328
88	0	0	1	1	12.044	1	0	0.201
89	1	1	1	1	11.075	1	1	0.314
90	0	1	1	0	21.458	5	0	0.000
91	0	1	1	0	20.250	4	3	0.054
92	0	0	0	0	15.661	4	2	0.239
93	0	0	1	0	11.040	3	2	0.108
94	0	2	0	0	13.975	1	1	0.288
95	0	1	0	0	22.888	8	2	0.033
96	1	2	1	2	16.564	2	0	0.455
97	1	2	0	0	13.065	2	1	0.327
98	0	2	0	0	16.596	2	0	0.154
99	0	1	1	0	14.478	2	1	0.055
100	0	0	0	1	14.315	2	0	0.310