CAPE
CUERPO ACADÉMICO DE PROBABILIDAD Y ESTADÍSTICA BUAP

Decimocuarta Semana Internacional de la Estadística y la Probabilidad
14-18 de junio de 2021

Aplicación de la Regresión Logística para el Análisis de Credit Scoring.

Alan Oswaldo Rosado Osorio [a], **Arlet Alondra Martínez Morales** [a], **Bulmaro Juárez Hernández** [a], **Jaime Daniel Orozco Jiménez** [a]
[a] *Facultad de Ciencias Físico Matemáticas, Benemérita Universidad Autónoma de Puebla, Avenida San Claudio y 18 sur, Colonia San Manuel, Ciudad Universitaria, C.P. 72570, Puebla, México*
alan.rosadoo@alumno.buap.mx, arlet.martinez@alumno.buap.mx, bjuarez@fcfm.buap.mx, jaime.orozcoj@alumno.buap.mx

Resumen

En este trabajo se presentan algunos conceptos básicos sobre Regresión Logística así como un caso de estudio enfocado en el área de Finanzas, especificamente, en una rama teórica conocida como Credit Scoring, en la que la aplicación de Regresión Logística tiene como objetivo el análisis y cálculo de las probabilidades involucradas en el otorgamiento o no de un crédito monetario a individuos de un determinado grupo, para de esta manera facilitar el proceso de decisión a las instituciones financieras involucradas en el proceso. El caso de estudio concluyó que, aunque la regresión logística es una herramienta importante para orientar a la institución financiera durante sus procesos de decisión, ésta solo es un apoyo más en la toma de decisiones para dichas instituciones.

Palabras clave: Regresión Logística, Credit Scoring.

1. Introducción

Cualquier institución financiera requiere de mecanismos eficientes que le permitan la consecución de la mejor cartera de clientes posibles, desde que son prospectos hasta ya consagrados. El credit scoring viene a coadyuvar al criterio humano en la difícil tarea de filtrar solicitudes de crédito. El CS, se refiere al uso de conocimiento sobre desempeño y características de un individuo que solicita un prestámo a una institución financiera para pronosticar si no representará un riesgo en el futuro para la compañía, es decir que tan factible es la posibilidad de otorgarle o no un crédito financiero y en caso de otorgarse, qué tan probable sería que cumpla con las obligaciones de pago [5]

En estadística, la regresión logística es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictoras. Es útil para modelar la probabilidad de un evento ocurriendo como función de otros factores.

Empleando la regresión logística un cliente puede clasificarse en un grupo de liquidez, rentabilidad, apalancamiento, o cualquier aspecto de interés para la institución financiera en función de las variables analizadas y, de esta forma, saber si le conviene o no a la institución financiera otorgar el crédito así como vislumbrar las posibles consecuencias de la decisión que tome [4].

La regresión logística se ha convertido en un paradigma muy usado para construir modelos predictores. La razón fundamental de esta popularización es debido al hecho de que los parámetros en los que se basa tienen una interpretación en términos de riesgo.

El análisis de regresión logística se enmarca en el conjunto de Modelos Lineales Generalizados (GLM) que usa como función de enlace la función logit. Las probabilidades que describen el posible resultado de un único ensayo se modelizan, como una función de variables explicativas, utilizando una función logística.

La regresión logística analiza datos distribuidos binomialmente de la forma:

$$Y_i \sim B(p_i, n_i), \text{ para } i = 1, ..., m.$$

Donde los números de ensayos Bernoulli n_i son conocidos y las probabilidades de éxito p_i son desconocidas.

El modelo es entonces obtenido en base a lo que cada ensayo (valor de i) y el conjunto de variables independientes pueden informar acerca de la probabilidad final. Estas variables explicativas pueden entenderse como un vector X_i k-dimensional con lo que el modelo toma la siguiente forma:

$$p_i = E\left(\frac{Y_i}{n_i}|X_i\right).$$

Asimismo, la función logit de las probabilidades binomiales desconocidas son modelizadas como una función de las X_i, tal que:

$$logit(p_i) = \ln(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 X_{1,i} + ... + \beta_k X_{k,i} + \epsilon_i,$$

donde los parámetros desconocidos β_j son estimados por Máxima Verosimilitud.

De igual forma, el modelo tiene una formulación equivalente dada por:

$$p_i = \frac{1}{1 - e^{-(\beta_0 + \beta_1 x_{1,i} + ... + \beta_k x_{k,i})}}.$$

2. Usos de la Regresión Logística

La regresión logística puede usarse para tratar de correlacionar la probabilidad de una variable cualitativa binaria (valores reales ”0ז ”1çon una variable escalar x). El objetivo es que la regresión logística aproxime la probabilidad de obtener ”0”(no ocurre cierto suceso) o ”1”(ocurre el suceso) con el valor de la variable explicativa x [2].

La probabilidad aproximada del suceso se calculará mediante una función logística del tipo:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{e^{(\beta_0 + \beta_1 x)} + 1} = \frac{1}{e^{-(\beta_0 + \beta_1 x)} + 1}.$$

Propiedades de la Regresión Logística

Sea $f(z) = \dfrac{1}{1 + e^{-z}}$ la función del modelo de regresión logística, obsérvese que:

- $\lim_{z \to \infty} f(z) = \lim_{z \to \infty} \dfrac{1}{1 + \exp^{-z}} = 1.$
- $\lim_{z \to -\infty} f(z) = \lim_{z \to -\infty} \dfrac{1}{1 + \exp^{-z}} = 0.$

Como puede observarse en la Figura 1.

3. Variables involucradas en la Regresión Logística

Como en la metodología empleada para la estimación del modelo logístico no intervienen variables cualitativas, la solución es la creación de variables dicótomicas conocidas como variables dummy, que toman sólo dos valores, generalmente uno o cero, para indicar la presencia o ausencia de una característica o para indicar si una condición es verdadera o falsa.

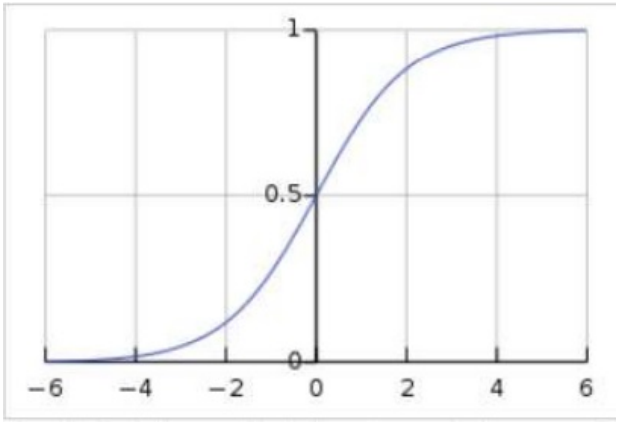

Figura 1. Función de Regresión Logística.

Las k - 1 variables dicotómicas se denotan por X_1, X_2, ..., X_{k-1}. A cada categoría o clase de la variable nominal le corresponde un conjunto de valores de los X_k con el cual se identifica dicha clase.

La manera más usual de definir estas k - 1 variables es la siguiente: Si el sujeto pertenece a la primera categoría, entonces las k - 1 variables dummy valen 0: (X_1 = X_2 = ... = X_{k-1} = 0); si el sujeto se encuentra en la segunda categoría, (X_1 = 1 y X_2 = ... = X_{k-1} = 0); y así sucesivamente hasta llegar a la última categoría, para la cual X_{k-1} = 1 y las restantes valen 0.

4. Risk Ratio y Odds Ratio

Risk Ratio. El riesgo relativo (Risk Ratio o RR) mide la fuerza de asociación entre la exposición y el evento; indica la probabilidad de que se desarrolle el evento de interés en los individuos expuestos o con cierta característica de riesgo en relación con los individuos que no estén expuestos a ese factor.

Interpretación

- Si $RR = 1$ indica que no hay asociación entre la presencia del factor y la ocurrencia del evento.
- Si $RR > 1$ indica que la asociacíon es positiva, es decir, la presencia del factor se asocia a mayor ocurrencia del evento.
- Si $RR < 1$, la asociación es negativa.

Odds Ratio. Es la única medida de asociación directamente estimada a partir de un modelo logístico. Su cálculo es mediante el cociente entre la probabilidad de que ocurra el evento y la probabilidad de que no ocurra, es decir, indica cuanto más probable es la ocurrencia del evento que su no ocurrencia.

$$OR(x) = \frac{P(C=1|x)}{1-P(C=1|x)}.$$

Interpretación

Si se tiene un OR de $1/3$ se interpreta diciendo que para una variable X la probabilidad de que el evento de interés ocurra es una tercera parte de que el evento de interés no ocurra [3].

- Si $OR = 1$ indica que no hay asociación entre la presencia del factor y la ocurrencia del evento.
- Si $OR > 1$ indica que la asociacíon es positiva, es decir, la presencia del factor se asocia a mayor ocurrencia del evento.
- Si $OR < 1$, la asociación es negativa.

La equivalencia entre RR y OR, se da en la Tabla 1.

Tabla 1. Equivalencia entre RR y OR.

Risk Ratio	Odds Ratio
0.1	0.1/0.9=0.11
0.2	0.2/0.8=0.25
0.3	0.3/0.7 = 0.43
0.4	0.4/0.6=0.67
0.5	0.5/0.5=1
0.6	0.6/0.4=1.50
0.7	0.7/0.3 = 2.33
0.8	0.8/0.2 = 4.00
0.9	0.9/0.1 = 9.00

5. Bondad de Ajuste en Regresión Logística

Una vez construido el modelo de regresión logística, tiene sentido comprobar que tan bueno es el ajuste de los valores predichos por el modelo a los valores observados. En un test global de bondad de ajuste se tiene el siguiente juego de hipótesis:

$$H_0 : p_j = \frac{1}{1+e^{-\beta_0+\sum_{i=1}^{k}\beta_i X_i}}.$$

vs

$$H_1 : p_j \neq \frac{1}{1+e^{-\beta_0+\sum_{i=1}^{k}\beta_i X_i}}.$$

En regresión logística existen varias medidas de ajuste global para comparar los valores predichos y los valores observados, por ejemplo, tests basados en patrones de las covariables como el test basado en la devianza D y el estadístico χ^2 de Pearson.

Devianza (D)

Se define como el doble logartimo del estadístico de verosimilitud y permite comparar los valores de la predicción con los valores observados en dos momentos:

- El modelo sin variables independientes, sólo con la constante (modelo referencia).
- El modelo con las variables predictoras introducidas.

El valor de la devianza debiera disminuir sensiblemente entre ambos momentos descritos e, idealmente, aproximarse a 0 cuando el modelo predice bien.

$$D = 2\sum_{j=1}^{J}(ln(\frac{\bar{p}_j}{\hat{p}_j} + (n_j - y_j)ln(\frac{1-\bar{p}_j}{1-\hat{p}_j}))).$$

El estadístico así construido tiene distribución asintótica Chi-Cuadradada, con grados de libertad dados por la diferencia entre la dimensión del espacio paramétrico y la dimensión del espacio bajo la hipótesis nula.

La hipótesis nula será rechazada cuando el p-valor de contraste sea menor que el nivel α fijado.

Estadístico Chi-Cuadrada de Pearson

El estadístico χ^2, con distribución de probabilidad del mismo nombre, sirve para someter a prueba hipótesis referidas a la distribución de frecuencias. Esta prueba contrasta frecuencias observadas con las frecuencias esperadas de acuerdo con la hipótesis nula.

$$\chi^2 = \sum_{j=1}^{J}\frac{(Y_j - n_j\hat{p}_j)^2}{n_j\hat{p}_j(1-\hat{p}_j)}.$$

Tiene la misma distribución asintótica que la devianza. Con lo cual, la hipótesis nula será rechazada cuando el p-valor sea menor que el nivel de significancia fijado.

6. Caso de Estudio.

El Sistema Financiero público y privado, busca priorizar el otorgamiento de créditos siempre que cumplan con el sistema de acreditación o calificación establecido por cada institución financiera. Lamentablemente el acceso no ha sido lo suficientemente aprovechado, ya que por razones administrativas no cumplen con los requisitos básicos para acceder a ellos; desde ese punto de vista, las decisiones de otorgamiento de crédito se encuentran determinadas por una serie de variables, tales como: registro de impagos, número de años de funcionamiento de la organización, ingresos por ventas, posesión de garantías, destino de los recursos, tipo de interés, etc. con las que determina si un individuo o compañía es apto para recibir un crédito sin que la institución financiera arriesgue en demasía sus recursos ante la posibilidad de no pago.

El presente estudio consiste en un análisis cuyo objetivo es establecer un modelo que permita calcular la probabilidad de que los integrantes de un grupo de personas obtengan un crédito bancario en función de la nota que han obtenido

según la calificación otorgada por los parámetros (variables) de la institución bancaria.

La variable crédito está codificada como 0 si no se tiene el crédito y 1 si se tiene.

Los resultados que se presentarán de aquí en adelante han sido obtenidos a través del procesamiento de la información del caso de estudio en el software estadístico R.

Obtención, Ingreso y Representación de la Información.

La información a analizar corresponde a dos matrices con un total 200 datos cada una, clasificadas a través de 2 variables de interés para hacer el análisis, la variable *crédito* y la variable $califbancaria$, esta información se presenta en la Tabla 2.

Tabla 2. Datos sobre la variable *crédito* y la variable *calificación bancaria*.

Crédito	0	1
Individuos	151	49

La Figura 2 muestra la distribución de los datos del caso de estudio así como la correlación existente entre las dos variables de interés, califbancaria y crédito, se puede observar que para la mayoría de los individuos se cumple la relación de que a mayor *calificación bancaria*, mayor presencia de *crédito*, y reciprocamente, a menor califación bancaria, menos personas con obtención de crédito. La presencia de datos extremos en ambas categorías de análisis podría explicarse mediante la inclusión de más variables en el modelo logístico.

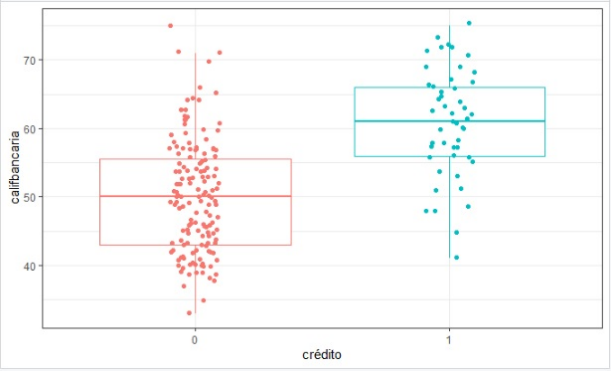

Figura 2. Distribución de los Datos.

Obtención del Modelo de Regresión Logística.

El coeficiente estimado para la intersección es el valor esperado del logaritmo de odds de que una persona haya obtenido el crédito bancario teniendo una mala calificación bancaria, que en este caso resulta ser igual -9.793942. De forma que la probabilidad estimada de que una persona haya obtenido el crédito bancario teniendo una mala calificación bancaria, es:

$$\pi(0) = \frac{1}{e^{-(-9.793942+0.1563404\cdot 0)}} = 5.578543261e^{-5}.$$

Acorde al modelo, el logaritmo de los odds de que una persona tenga el crédito bancario está positivamente relacionado con la calificación que el banco le haya otorgado (coeficiente de regresión = 0.1563404)

Esto significa que, por cada unidad que se incrementa la variable $califbancaria$, se espera que el logaritmo de odds de la variable *crédito* se incremente en promedio 0.1563404 unidades.

Aplicando la inversa al logaritmo de los odds de la variable crédito se tiene que $e^{0.1563404}) = 1.169$. Esto significa que, por cada unidad que se incrementa la variable $califbancaria$, los odds (probabilidades) de obtener el crédito se incrementan en promedio 1.169 unidades.

Interválos de Confianza.

Obsérvese que, a través del uso del software R, así como de la instrucción "confint(object = modelo, level = 0.95)", se obtienen los interválos de confianza al $95\,\%$ para el Modelo de Regresión Logística, esto es, los intervalos de confianza que se obtiene para el intercepto y el coeficiente estimado de la variable *califbancaria* son:

	LI	LS
Intercepto	-12.9375208	-7.0938806
CalifBancaria	0.1093783	0.2103937

Los intervalos de confianza presentados en la tabla anterior, permiten determinar un rango de valores en los que los resultados obtenidos por el modelo logístico serían confiables y por tanto útiles para su análisis e interpretación.

Interpretación de los Odds Ratio como Probabilidades.

La siguiente gráfica Figura 3, muestra la distribución del modelo logístico obtenido para el caso de estudio, asimismo se visualiza la fluctuación del gráfico con los posibles valores que podría tomar el modelo según los intervalos de confianza presentados previamente.

Test de Máxima Verosimilitud.

El test Likelihood ratio calcula la significancia de la diferencia de residuos entre el modelo de interés y el modelo nulo. El estadístico sigue una distribución chi-cuadrado con grados de libertad equivalentes a la diferencia de grados de libertad de los dos modelos [1].

- Diferencia de residuos: 55.6368.

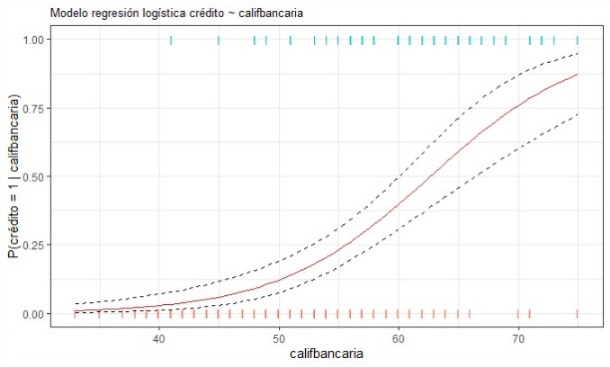

Figura 3. Modelo Logístico con Banda de Confianza.

- Grados de libertad: 1.
- p-value: $8.71759108087093e^{-14}$.

Con los resultados obtenidos en este caso, y dado que el p-value obtenido en el modelo es menor que el valor α del intervalo de confianza establecido, se puede concluir que el modelo si es significativo.

Evaluación del Modelo.

La Figura 4 muestra la concordancia del modelo de regresión logística generado (predicciones) respecto a los datos observados. De esta forma, el modelo es capaz de clasificar correctamente:

$$\frac{140+22}{140+22+27+11} = 81\,\% \text{ de los datos.}$$

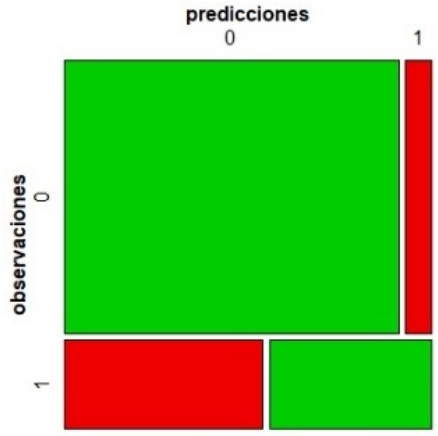

Figura 4. Observaciones vs Predicciones.

Asimismo, con los datos obtenidos en el caso de estudio, se obtuvieron fórmulas generales aplicables al modelo logístico en caso de la inclusión de nuevas observaciones.

$$logit(crédito) = -9.793942 + 0.1563404 * califbancaria.$$

$$P(Crédito) = \frac{e^{-9.793942+0.1563404*califbancaria}}{1+e^{-9.793942+0.1563404*califbancaria}}.$$

7. Conclusiones.

El modelo logístico obtenido para predecir la probabilidad de que una persona obtenga el crédito bancario en función de la calificación otorgada por el banco es:

- Significativo acorde al Ratio de Verosimilitud (p-value = $8.717591e^{-14}$)
- El p-value del predictor *califbancaria* es significativo ($p - value = 1.03e^{-09}$).

Existe una relación inversa entre el riesgo y el puntaje del score; a mayor puntaje menor riesgo de no pago del cliente. En resumen, el modelo de CS solo es un indicativo, la determinación fundamental radica en el recurso humano de las instituciones.

Referencias

[1] Albert, A. and J.A. Anderson. *On the Existence of Maximun Likelihood Estimates in Logistic Models; Biometrika, 71, 1-10, 1984.*

[2] Kleinbaum, D. and M. Klein. *Logistic Regression; Springer, 2da Edición, 2002.*

[3] Menard, S. *Coefficients of Determination for Multiple Logistic Regression Analysis; The American Statistician, 51, 17-24, 2000.*

[4] Orgler, Yair E. *A credit scoring model for commercial loans; Journal of Money Credit and Baking, 1971.*

[5] Schreiner, Mark. *A simple poverty scorecard for Mexico; National Household Income, 2011*