UNSAAC

Universidad Nacional de
San Antonio Abad del Cusco

Análisis Multivariado 1

- PROFESOR: ARTURO ZUÑIGA

Regresión Logística Binaria

Introducción I

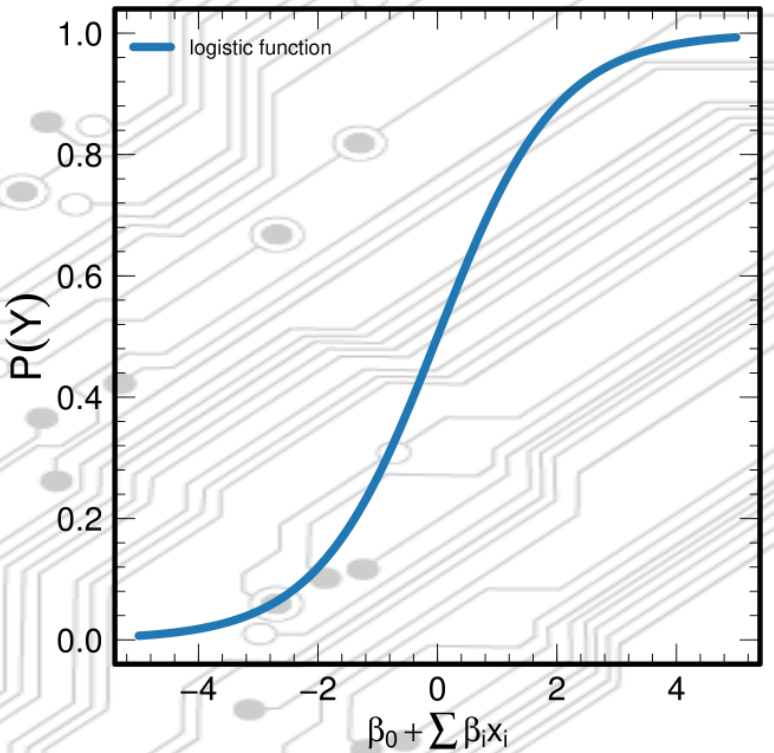
En vez de modelar directamente la respuesta Y , los modelos de regresión logística modelan la probabilidad de que Y pertenezca a una categoría en particular.

Para la data de nuestro ejercicio al final, la regresión logística modela la probabilidad de que un cliente incumpla con el pago de la tarjeta de crédito (moroso).

Por ejemplo, la probabilidad de que sea moroso dado balance puede ser escrita como

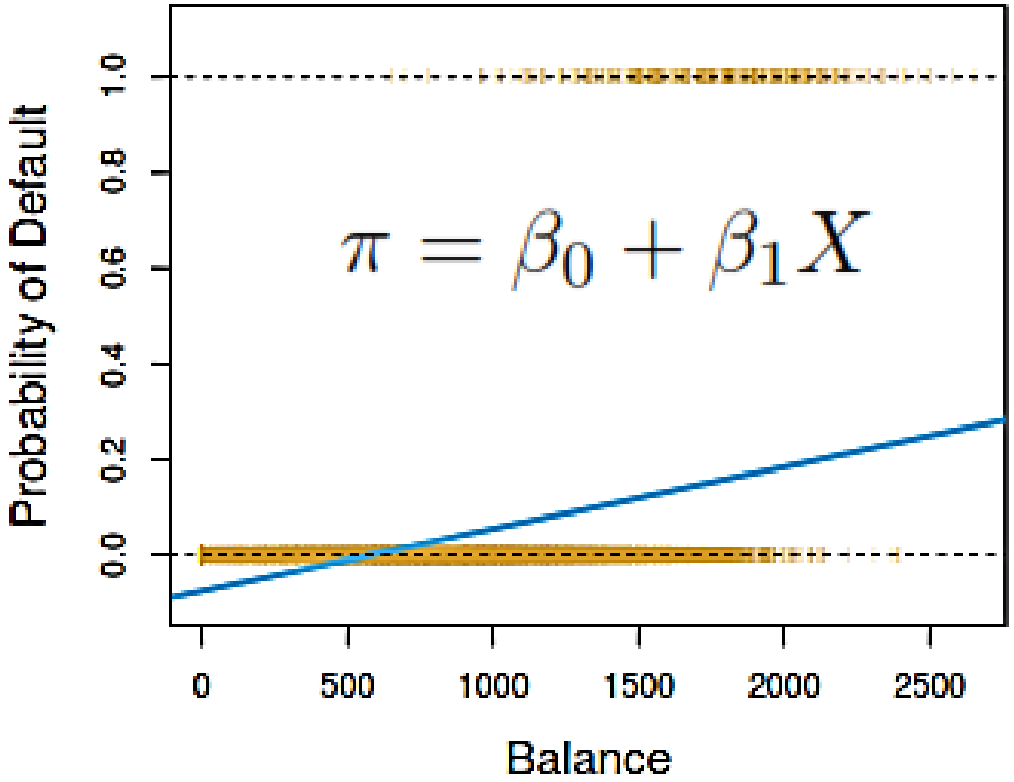
$$Pr(default = Yes \mid balance)$$

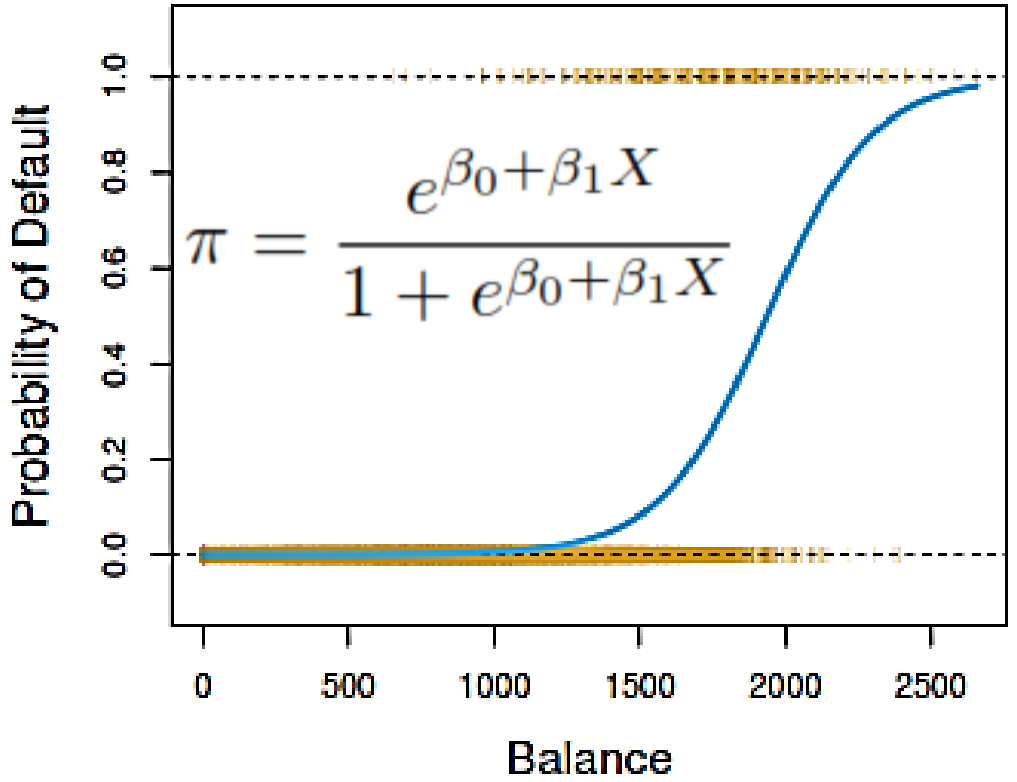
Los valores de $Pr(default = Yes \mid balance)$, pueden abreviarse como π, se encontrarán en el rango entre 0 y 1.



Introducción II

¿Cómo debería de modelarse la relación entre $\pi = Pr(Y = 1|X)$ y X?





Regresión logística I

En la regresión logística, es usada la función logística

$$\pi = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

La grafica del lado derecho de la figura anterior muestra el ajuste a un modelo de regresión logística para el conjunto de datos Default.

Se puede observar que el modelo logístico captura mejor el rango de probabilidades que el modelo de regresión lineal mostrado en el lado izquierdo.

La probabilidad ajustada promedio en ambos casos es 0.0333, la cual es la misma que la proporción total de morosos en la data.

Con alguna manipulaciones básica

$$\frac{\pi}{1-\pi} = e^{\beta_0+\beta_1 X}$$

Regresión logística II

El valor $\frac{\pi}{1-\pi}$ es conocido como odds y puede tomar cualquier valor entre 0 y ∞.

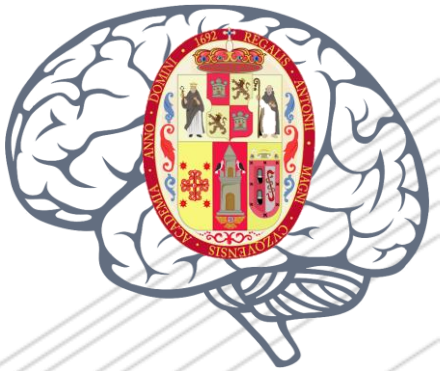
Los valores de odds cercanos a 0 o a ∞ indican probabilidades muy bajas o muy altas de ser morosos, de manera respectiva.

Tomando logaritmos a la anterior ecuación se obtiene

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

esta expresión es conocido como el logit.

Regresión logística III



- **Componente aleatorio**: Sean $Y_1, ..., Y_n$ v.a. dicotómicas independientes. Asumiendo que $y_i = 1$ tiene probabilidad π_i y $y_i = 0$ con probabilidad $1 - \pi_i$:

$$y_i \sim Bernoulli\,(\pi_i)$$

- **Componente sistemático**:

$$\eta_i = \beta_1 x_{1i} + ... + \beta_1 x_{pi} = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

donde η_i es denominado como predictor lineal y $x_i = (x_{i1}, \ldots, x_{ip})'$ es un vector de covariables, donde x_{i1} igual a 1 corresponde al intercepto.

- **Función de Enlace:**

$$g(\pi_i) = \eta_i$$

donde $g(.)$ es una función monótona y diferenciable.

Enlaces comunes

- Enlace Logit

$$\log\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \beta_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

- Enlace Probit

$$\Phi(\pi(x))^{-1} = \beta_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

donde $\Phi(.)$ es la f.d.a. de la normal estándar.

- Enlace log-log complementario (cloglog)

$$\log\{-\log(1-\pi(X))\} = \beta_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Definición del modelo

El modelo de regresión logística múltiple esta expresado por:

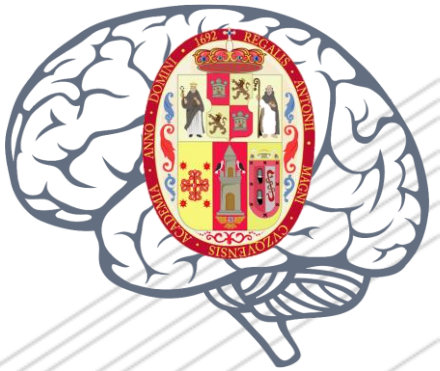
$$\log\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \beta_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

donde $x = \left(1, x_2, \ldots, x_p\right)^T$ contienen los valores observados de las variables explicativas. Entonces se tiene un modelo para el logaritmo de los odds (log_odds) $\log\left\{\frac{\pi(x)}{1-\pi(x)}\right\}$ usando la transformación con la función exponencial se obtiene

$$\frac{\pi(X)}{1-\pi(X)} = \exp(\beta_1) \cdot \exp(\beta_2 x_2) \cdot \ldots \cdot \exp(\beta_p x_p)$$

lo cual implica que los efectos de las covariables afectan los odds en un forma exponencial multiplicativa.

Interpretación de los coeficientes



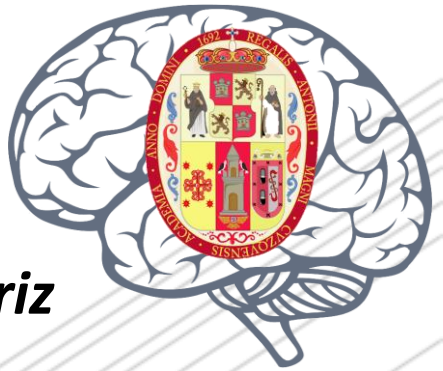
De acuerdo a esto :

- $\beta_i > 0$: $P(y_i = 1)/P(y_i = 0)$ se incrementa.
- $\beta_i < 0$: $P(y_i = 1)/P(y_i = 0)$ disminuye.
- $\beta_i = 0$: $P(y_i = 1)/P(y_i = 0)$ se mantiene constante.

A partir de lo anterior podemos dar una interpretación a los parámetros del modelo: β_o es el valor del logit cuando las variables predictoras son nulas. β_j es la variación del logit cuando x_j se incrementa en una unidad y las demás variables se mantienen constantes.

Alternativamente, podemos también interpretar a e^{β_j} como la variación porcentual del riesgo relativo cuando x_j se incrementa en una unidad y las demás variables se mantienen constantes.

Evaluación de las Clases Predichas



Un método común para describir la performance de la clasificación es la ***matriz de confusión***.

Esta es un simple tabulación cruzada para las clases observadas y predichas.

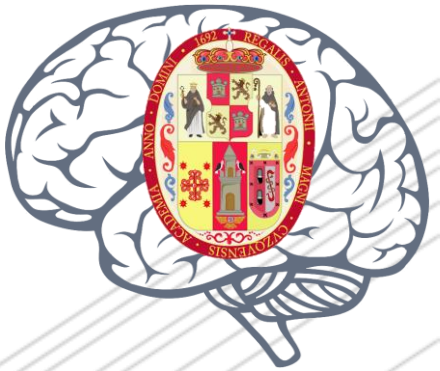
Predichos	Observados	
	Eventos	No Eventos
Eventos	TP	FP
No Eventos	FN	TN

La métrica más simple es el ratio de la exactitud total.

$$acc(d) = \frac{TP + TN}{P + N}$$

o, siendo pesimistas, el ratio de error: $1 - acc(d)$

Cuando la clasificación es binaria



Para dos grupos existen estadísticas adicionales que pueden ser relevantes cuando una clase es interpretada como un evento de interés.

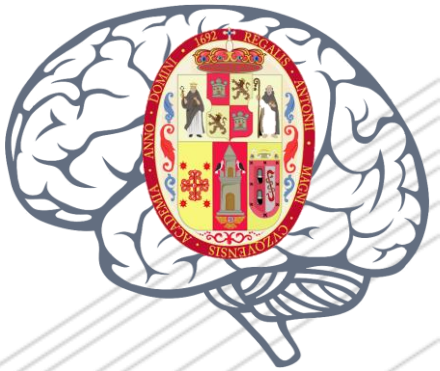
La sensibilidad o recuperación de un modelo es el ratio en que el evento de interés es predicho correctamente para todas las muestras que contienen el evento.

$$sensibilidad(d) = \frac{TP}{TP + FN}$$

Cuan bueno es el modelo prediciendo a los éxitos.

La sensibilidad es usualmente considerada como el ratio de verdaderos positivos dado que mide la precisión en los eventos de la población.

Cuando la clasificación es binaria



De manera inversa, la especificidad es definida como el ratio de observaciones que no son los eventos que son predichos como no eventos (ratio de verdaderos negativos).

$$Especificidad(d) = \frac{TN}{FP + TN}$$

La falsa alarma o ratio de falsos positivos es definido como uno menos la especificidad.

$$Falarm(d) = 1 - Especificidad(d) = \frac{FP}{FP + TN}$$

Equilibrio entre sensibilidad y especificidad

Usualmente, es de interés in tener una medida única que refleja los ratios de falsos positivos y los de falsos negativos. El índice de Youden (Youden, 1950) definido como:

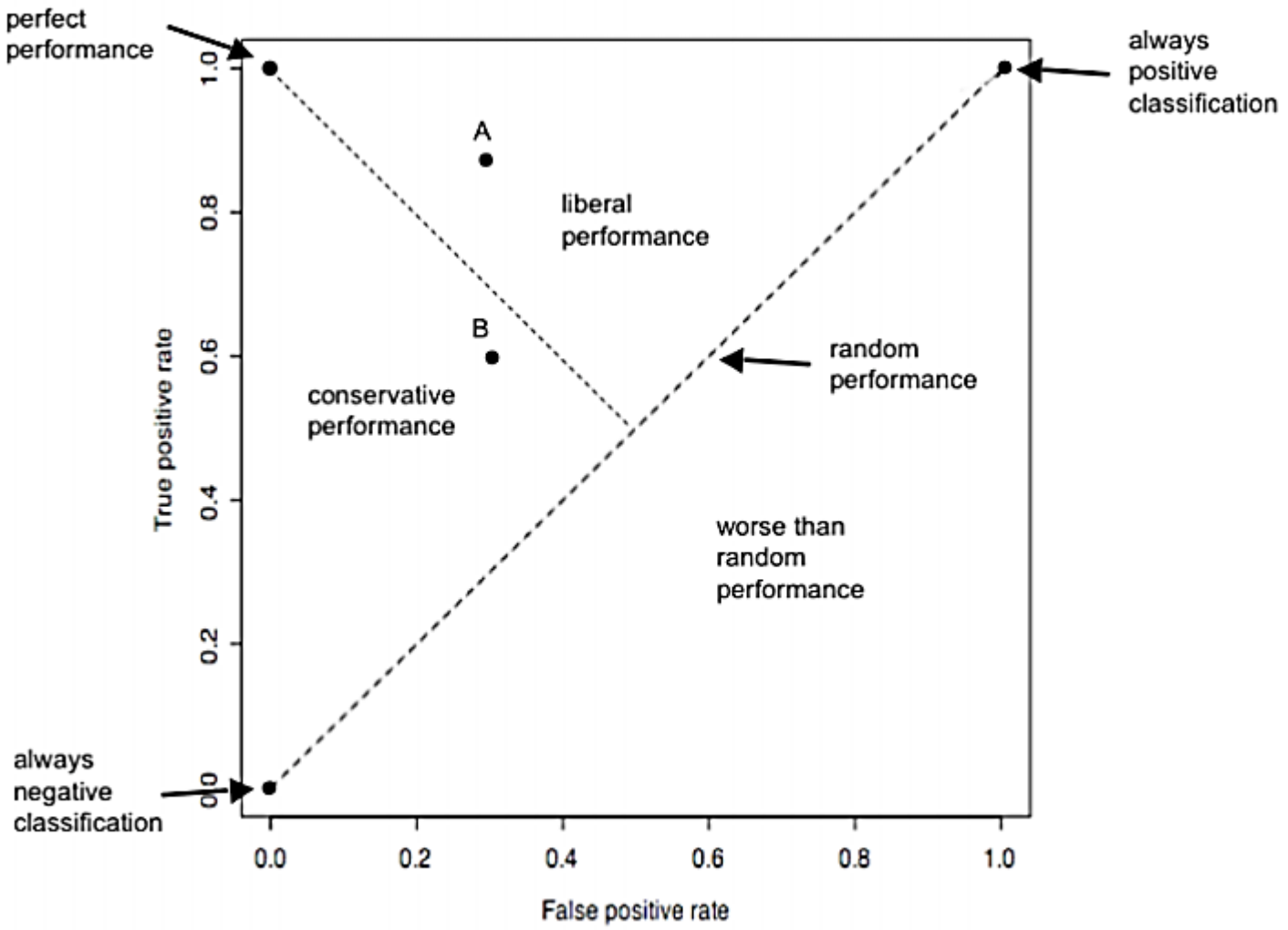
$$J = sensibilidad + Especificidad - 1$$

mide las proporciones de observaciones correctamente predichas tanto para los eventos como para los no eventos. En algunos contextos esto puede ser un método apropiado para resumir la magnitud de ambos tipos de errores.

Curvas ROC

La curva ROC (receiver operating characteristic) es una de las técnicas más comunes para evaluar la combinación de la sensibilidad y la especificidad dentro de un único valor.

Un aspecto que se suele pasar por alto al evaluar la sensibilidad y la especificidad es que son medidas condicionales.



Caso: Creditos de Banco Alemán

En el archivo ***credit.dta*** se encuentran datos referidos a 1000 créditos otorgados por un banco alemán publicado por Fahrmeir, Hamerle, y Tutz (1996). Cada cliente está asociado a una respuesta binaria y definida como:

Y=1 Cliente no es solvente

y = 0 Cliente es solvente

además de un conjunto de variables predictoras:

- ***Acc1:*** 1 = no posee cuenta corriente, 0 = cuenta corriente buena o mala
- ***Acc2***: 1 = cuenta corriente buena, 0 = cuenta corriente mala o no posee cuenta corriente
- ***duration:*** Duración del crédito en meses
- ***amount:*** Monto del crédito en miles de euros

El interés es predecir si un individuo va a incumplir en el pago de su tarjeta de crédito, sobre la base de su ingreso anual y el balance mensual de la tarjeta.

Gracias