FACULTAD DE CIENCIAS ECONÓMICAS

FAMAF
Facultad de Matemática, Astronomía, Física y Computación

UNC Universidad Nacional de Córdoba

DIPLOMATURA

CIENCIA DE DATOS, INTELIGENCIA ARTIFICIAL Y SUS APLICACIONES EN ECONOMÍA Y NEGOCIOS

Regresión logística

Introducción

Técnicas de clasificación supervisada: tenemos una muestra de elementos bien clasificados que sirve de modelo para la clasificación de las observaciones siguientes

- Se tiene un conjunto de elementos de los cuales se conoce la pertenencia a un grupo preestablecido, representado por una variable categórica Y.
- Se definen un conjunto de variables clasificadoras (predictoras) designado por el vector X
- Regla de asignación: función discriminate.
- Se desea determinar que variables caracterizan a cada categoría de la variable dependiente o clasificar un nuevo elemento, con valores de las variables conocidas, en uno de los grupos.

Ejemplo: tenemos información sobre una muestra de empresas que cotizan en bolsa y se desea determinar qué ratios contables influyen sobre la decisión de distribuir dividendos. Además, una vez construido el modelo será posible emplearlo para predecir si la empresa distribuirá no dividendos en función de indicadores económico-financieros seleccionados.

Introducción

En el análisis de regresión lineal, trabajamos con variables dependientes de tipo métrico. Las variables no métricas se incluyeron en el análisis, pero dentro del grupo de variables independientes. No obstante, pueden presentarse, en la práctica, situaciones donde la variable dependiente sea de naturaleza cualitativa (no métrica).

Modelos de respuesta no métrica

De respuesta dicotómica: la variable dependiente puede tomar sólo dos valores posibles, mutuamente excluyentes (tener o tener una característica).

De respuesta múltiple: la variable dependiente puede tomar más de dos valores posibles, mutuamente excluyentes.

Regresión logística binaria

Tenemos una o varias variables predictoras (x_1,x_2, …., x_k) y una variable dependiente (y) que asume los valores 0 o 1, por ejemplo:

Y= ¿tiene empleo?, donde 0= tiene, 1= no tiene → $P(Y_i = 1)$ = probabilidad de desempleo (éxito)
Y= Solvencia, donde 1=solvente, 0= insolvente → $P(Y_i = 1)$ = probabilidad de solvencia (éxito)

¿Podemos ajustar un modelo de regresión lineal?

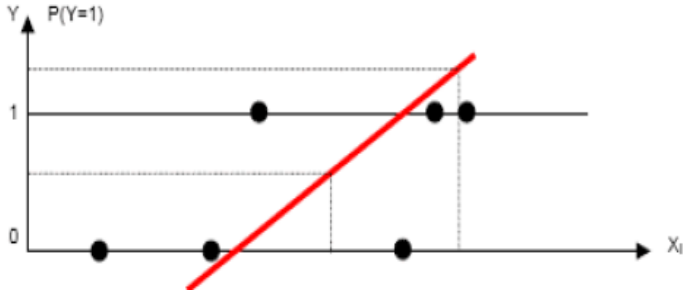
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

En un modelo de regresión logística lo que estamos es estimando es la probabilidad de Y_i dados los valores de las X_i

→ $P(Y_i = 1/X_i) = P_i$

Regresión logística binaria

- Si estimamos P_i con el modelo lineal, dadas las variables predictoras $(x_1, x_2, \ldots, x_k)$ no hay garantías de que $\widehat{p_i}$ esté entre 0 y 1 (podemos obtener probabilidades mayores que 1 o negativas). Esto no es un problema para clasificar la observación, pero sí para interpretar el resultado de la regla de clasificación.

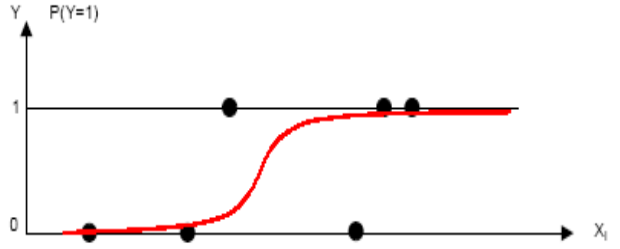


- Como los únicos valores posibles de Y son 0 y 1, se puede demostrar matemáticamente que la distribución de los errores aleatorios no es normal ni se cumple la homocedastidad.

Si queremos que el modelo construido para discriminar nos proporcione directamente la probabilidad de éxito, debemos transformar el modelo de regresión lineal para garantizar que la respuesta prevista esté entre cero y uno.

$$P_i = F(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki})$$

F debe ser una función no decreciente, acotada entre cero y uno.



Regresión logística binaria

- Una función *F* que cumple esta condición es la función de distribución logística, dada por:

$$P_i = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+ \ldots+\beta_k X_{ki})}} = \frac{e^{(\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+ \ldots+\beta_k X_{ki})}}{1 + e^{(\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+ \ldots+\beta_k X_{ki})}}$$

- La función *logit* representa en una escala logarítmica la diferencias entre las probabilidades entre las probabilidades del éxito y del fracaso.

- Al estar en términos de una función lineal de las variables explicativas facilita la estimación (por MV) y la interpretación del modelo.

- Se aplica en una amplia gama de situaciones donde las variables explicativas no tienen distribución conjunta normal multivariante.

- Se puede tomar como F a otras funciones como la distribución normal estándar (modelo *probit*) pero no tiene las ventajas de interpretación del modelo logístico.

Estimación de los parámetros

- El modelo se estima mediante la minimización de la función de **máxima verosimilitud,** que es un planteamiento análogo a evaluar cuánta información queda por explicar después que el modelo se ha estimado:

$$LL = \sum_{i=1}^{N} [Y_{i.}.\ln(P(Y_{i.})) + (1 - Y_{i.}).\ln(1 - (P(Y_{i.}))\ ]$$

La función toma valores cercanos a cero cuando la probabilidad predicha implica clasificar correctamente y valores muy grandes en el caso de desacierto.

Mayor valor de *LL* implicará menor capacidad del modelo estimado para clasificar correctamente.

Contraste de significatividad global

Razón de Máxima Veroslimilitud (Likelihood Ratio)

$H_o)\ \beta_1 = \beta_2 = \ldots\ldots = \beta_K = 0$
$H_1)\ Al\ menos\ un\ \beta_i \neq 0$

Se calcula la función de máxima verosimilitud LL de un modelo sin variables explicativas, sólo con el intercepto (LL(0)) y del modelo que estamos estimando (LL(M)). Si la segunda es significativamente menor concluimos que al menos una variable es significativa (su β es distinto de cero).

$$\chi^2 = -2LL(M) - \big(-2LL(0)\big) = 2LL(0) - 2LL(M)$$

A esta diferencia se la conoce como **Razón de Máxima Verosimilitud** y sigue una distribución Chi- cuadrado con grados de libertad equivalentes a la diferencia de grados de libertad de los dos modelos comparados. -2LL suele denominarse *deviance.*

Si el test resulta significativo, implica que el modelo es útil, pero no que sea el mejor. Podría ocurrir que alguno de sus predictores no fuese necesario.

Contraste para los coeficientes individuales

Test de Wald

$$H_o)\ \beta_j = 0$$
$$H_1)\ \beta_j \neq 0$$

Permite analizar la contribución individual de los regresores en la explicación de la variable dependiente.

$$W_j = \frac{b_j}{S_{b_j}}$$

Este test sigue una distribución Normal y a partir de ella se establecen sus valores críticos.

Cuando el coeficiente estimado b_j es grande, el error tiende a crecer en exceso, lo que puede llevar a no rechazar la Ho cuando debería rechazarse (error tipo II) porque la contribución de la variable es significativa (Menard, 1995).

Interpretación de los coeficientes

En la regresión múltiple, los coeficientes están afectados por las unidades de las variables, por lo que a través de la estandarización se puede analizar su importancia relativa.Este papel en la regresión logística, lo tiene, lo que se conoce como *odd ratio*.

Se define un *odd* de un acontecimiento como la razón entre su probabilidad de ocurrencia y la de no ocurrencia.

$$odd = \frac{P(Y=1)}{P(Y=0)}$$

Odd ratio (OR)

Mide el efecto que tiene sobre el odd, un incremento unitario en la variable independiente (permaneciendo constante el resto).

$$odd\ ratio\ (OR) = \frac{odd_1}{odd_0} = e^{\beta_i}$$

OR > 1 Se interpreta que el factor incrementa las chances de que ocurra el evento (probabilidad de ocurrencia frente a la de no ocurrencia)

OR < 1 Se interpreta que el factor reduce las chances de que ocurra el evento.

OR = 1 No hay efecto del factor

Evaluación del ajuste del modelo

Pseudo R^2 de McFadden

No es un equivalente directo del coeficiente de determinación empleado en regresión lineal que indica que parte de la varianza total es explicada por las variables independientes.

$$R^2_{MF} = \frac{-2LL(0) - (-2LL(M))}{-2LL(0)}$$

Es la proporción de reducción en la *deviance* que supone el modelo propuesto respecto al modelo base.

Varía entre 0 y 1. Tiene valor 0 si el modelo no mejora al modelo nulo, y valor de 1 si se ajusta perfectamente a los datos.

Capacidad discriminante del modelo

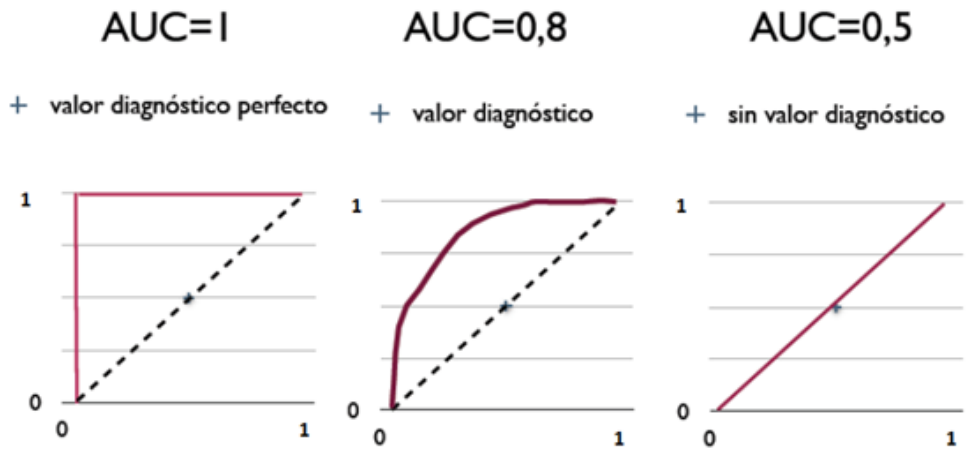
Matriz de confusión

Tabla que compara los valores reales y predichos. Se toma un criterio, por ejemplo, cuando la probabilidad es mayor a 0,50 se asigna al grupo 1 y cuando es menor al grupo 0. Cuanto más elementos haya correctamente clasificados, mayor es la precisión del modelo (porcentaje de concordancia)

Curva ROC (AUC)

Compara para diferentes puntos de corte de la probabilidad la tasa de clasificaciones correctas (positivos correctamente predichos / Positivos reales) y la tasa de falsos positivos (Falsos positivos /Negativos reales).

El estadístico AUC (área debajo de la curva) posee un valor comprendido entre 0,5 y 1, donde 1 representa un valor diagnóstico perfecto y 0,5 es una prueba sin capacidad discriminatoria diagnóstica.



Ejemplo

Se quiere establecer un modelo que permita calcular la probabilidad de obtener un crédito en función de un scoring determinado por el banco. La variable crédito está codificada como 0 si el cliente no obtuvo el crédito y 1 si accedió al crédito.

$$\widehat{p_i} = \frac{e^{(-9{,}7939+0{,}1563x_i)}}{1+e^{(-9{,}7939+0{,}1563x_i)}}$$

Si el scoring del cliente es de 55, por ejemplo, la probabilidad que acceda al crédito es:

$$\widehat{p_i} = \mathrm{P}(Y=1) = \frac{e^{(-9{,}7939+0{,}1563\,.\,55)}}{1+e^{(-9{,}7939+0{,}1563\,.\,55)}}$$

$\widehat{p_i} = \mathrm{P}(Y=1) =$ **0,2319**

Ejemplo

$$OR = e^{0,1563} = 1,17$$

El coeficiente estimado indica que el scoring está positivamente relacionado con las chances de obtener el crédito.

Por cada unidad en que se incrementa el scoring, las chaces de obtener el crédito aumentan en 1,17. Las chances estimadas de que un cliente obtenga el crédito se incrementan en (OR-1)*100 = (1,17-1)*100 = 17% por cada incremento unitario del scoring.

El intervalo de confianza para *OR* será:

$$(e^{0,1061}; e^{0,2065}) = (\ 1,11\ ;\ 1,23)$$

Ejemplo Telecom

Variable	Descripción	Valores
Género	Género del cliente	Femenino Masculino
Pareja	Indica si el cliente tiene pareja	Si No
Hijos	Indica si el cliente tiene hijos o no	Si No
Meses	Número de meses que el cliente ha permanecido en la empresa	
Servicio	Si el cliente tiene servicio telefónico o no	Si No
Lineas	Si el cliente tiene varias líneas o no	Si No Sin servicio telefónico
Internet	Proveedor de servicios de Internet del cliente	DSL fibra óptica No

Ejemplo Telecom

Variable	Descripción	Valores
Contrato	Plazo del contrato que tiene el cliente	Mensual Anual Bianual
Factura	Indica si el cliente tiene facturación electrónica	Si No
Pago	Método de pago del cliente	Cheque electrónico Cheque enviado por correo Débito en cuenta Tarjeta de crédito
Impormensual	Importe cobrado a la cliente mensualmente.	
Importotal	Importe total cargado al cliente	
Cancelación	Indica si el cliente canceló el servicio	Si No