Regresión logística

Modelo de regresión logística

En el modelo de regresión lineal visto anteriormente la variable de interés era numérica, por ejemplo: salario.

Pero ¿qué sucede cuando se quiere explicar una variable categórica en función de ciertas variables observadas?
por ejemplo: ser peronista o antiperonista.

Modelo de regresión logística

En estos casos se utiliza el método de regresión logística o logit:

- Binomial (cuando la variable dependiente admite dos valores: 0, 1, sí, no)
- Multinomial (cuando la variable dependiente se representa en tres o más categorías: votante de la UCR, votante del peronismo, votante del PRO)

Los nombres se deben en cada caso a la distribución asociada a la estimación de la probabilidad.

Modelo de regresión logística con variable dependiente binaria o dummy

Ejemplo: puntaje crediticio.
Se busca estimar la probabilidad de repago de un crédito dadas ciertas variables explicativas consideradas relevantes para explicar esta probabilidad.

Variable dependiente: probabilidad de repago de un crédito
1 = se paga el crédito
0 = no se paga el crédito

Modelo de regresión logística con variable dependiente binaria o dummy

Modelo de probabilidad lineal

La probabilidad de una variable discreta puede calcularse como la suma de todos los valores que la misma adopta multiplicado por su probabilidad de ocurrencia. Siendo que Y = 0 o 1:

$$P(Y = 1 \mid X) \quad P_i = 1$$
$$P(Y = 0 \mid X) \quad 1 - P_i = 0$$
$$E(Y \mid X) = 0 \cdot (1 - P_i) + 1 \cdot P_i$$
$$P_i = P(Y_i = 1 \mid X)$$

Cuando la VD es binaria, la esperanza coincide con la probabilidad del evento analizado. Si es repago es probabilidad de repago, por ejemplo.

Modelo de regresión logística con variable dependiente binaria o dummy

Modelo de probabilidad lineal

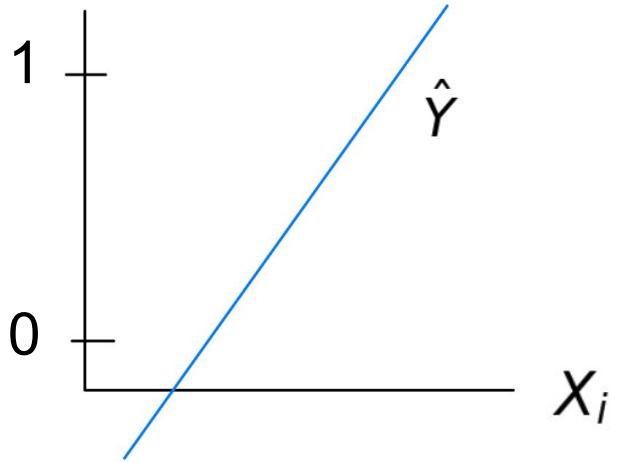
$$P_i = \mathbb{P}(Y_i = 1 \mid X) = \beta_0 + X_i\beta + U_i$$

Se expresa la probabilidad del evento estudiado como función lineal de los coeficientes del modelo.

Modelo de regresión logística con variable dependiente binaria o dummy

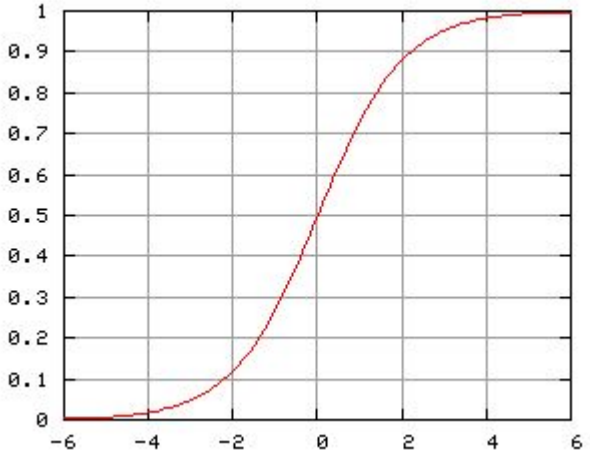
Modelo de probabilidad lineal

Limitación de este modelo: el método de mínimos cuadrados clásicos estima una recta que minimiza la suma de residuos al cuadrado, pero que puede arrojar valores de Y negativos y por encima de 1, es decir fuera del intervalo [0, 1] que es lo que admite la probabilidad.



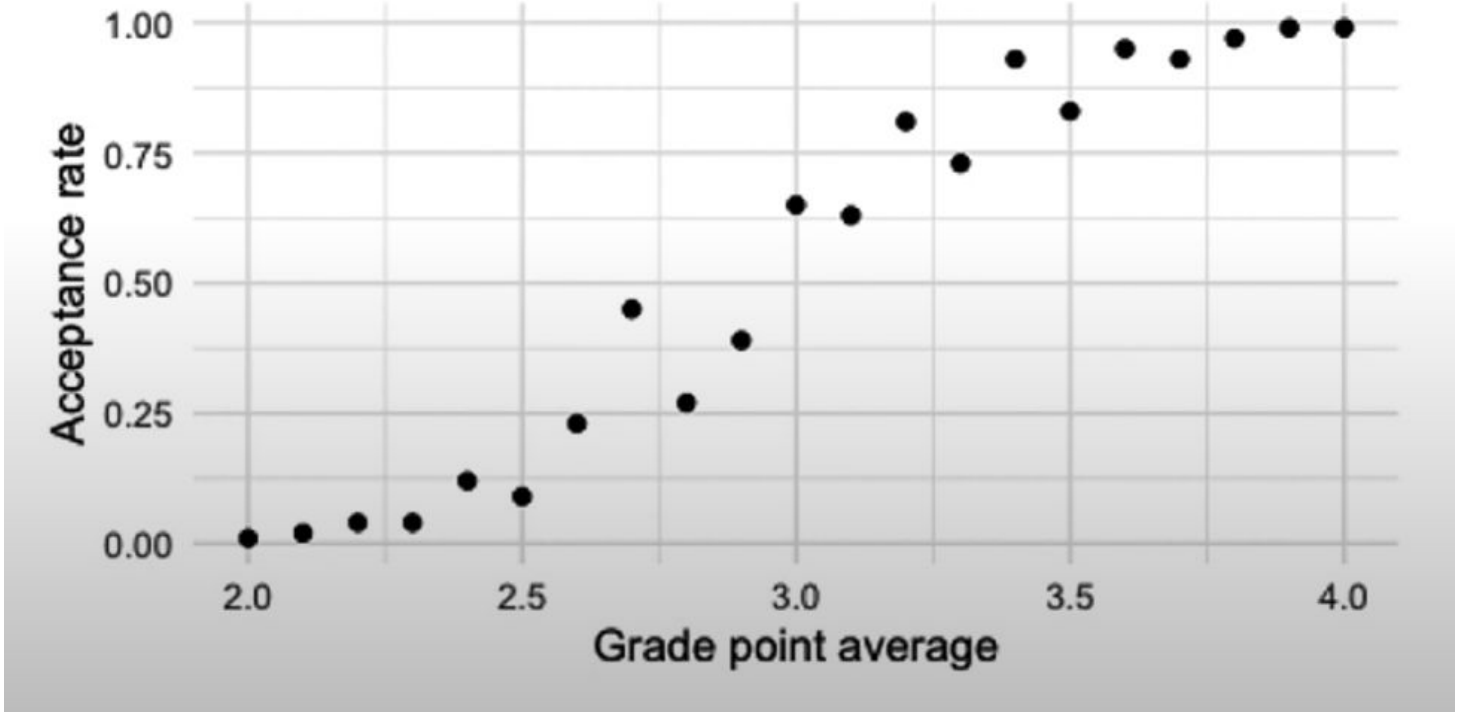
Modelo de regresión logística con variable dependiente binaria o dummy

Para solucionar este problema se utiliza la función logística que transforma las probabilidades originales para que estén entre 0 y 1



$$f(z) = \frac{1}{1 + e^{-z}}$$

Modelo de regresión logística con variable dependiente binaria o dummy



Modelo de regresión logística con variable dependiente binaria o dummy

Definamos la probabilidad de repago de un crédito en función de utilidades:

- i repaga el crédito si

$$U_i^R > U_i^{NR}$$

$$U_i^R = \beta_0^R + \beta^R X_i + U_i^R$$

$$U_i^{NR} = \beta_0^{NR} + \beta^{NR} X_i + U_i^{NR}$$

Modelo de regresión logística con variable dependiente binaria o dummy

- i repaga el crédito si

$$P(Y_i = 1 \mid X) = P(U_i^R > U_i^{NR} \mid X)$$
$$P(\beta_0^R + \beta^R X_i + U_i^R > \beta_0^{NR} + \beta^{NR} X_i + U_i^{NR} \mid X)$$
$$P(U_i^{NR} - U_i^R < \beta_0^R - \beta_0^{NR} + (\beta^R - \beta^{NR}) X_i \mid X)$$

$$P(Y_i = 1 \mid X) = P(U_i < \beta_0 + \beta X_i \mid X)$$

Con distribución logística,

$$P(U_i < \beta_0 + \beta X_i \mid X)$$

representa la función de distribución acumulada hasta $\beta_0 + \beta X_i$

Modelo de regresión logística con variable dependiente binaria o dummy

$f(\beta_0 + \beta X_i)$ ⟵ función de distribución acumulada de la logística hasta $\beta_0 + \beta X_i$

$$P(Y_i = 1 \mid X) = f(\beta_0 + \beta X_i) = \frac{1}{1 + e^{-(\beta_0 + \beta X_i)}} = \frac{e^{\beta_0 + \beta X_i}}{1 + e^{\beta_0 + \beta X_i}}$$

Modelo de regresión logística con variable dependiente binaria o dummy

Se puede modelar la probabilidad de repago de un crédito de la siguiente manera:

$$\Pr(Y_i = 1) = f(X_i\beta) = \frac{1}{1 + e^{-X_i\beta}}$$

*donde $X\beta$ admite K variables explicativas

En este caso, el modelo es no lineal en los parámetros, por lo que MCC no dará buenos estimadores

Modelo de regresión logística con variable dependiente binaria o dummy

Se utiliza el método de maximización de la función de maximización de verosimilitud, y puntualmente la maximización de su logaritmo natural. La función de verosimilitud es una función de probabilidad conjunta

Ejemplo: muestra de 3 clientes de un banco:

i	Y_i
1	1
2	0
3	1

Modelo de regresión logística con variable dependiente binaria o dummy

Ejemplo: muestra de 3 clientes de un banco:

i	Y_i
1	1
2	0
3	1

Si suponemos que se trata de una muestra aleatoria, representativa, en la que las variables son independientes, ¿cómo se estimaría la probabilidad conjunta observada?: $\Pr(Y_1 = 1 \text{ y } Y_2 = 0 \text{ y } Y_3 = 1 \mid X)$

Modelo de regresión logística con variable dependiente binaria o dummy

Ejemplo: muestra de 3 clientes de un banco:

i	Y_i
1	1
2	0
3	1

Si suponemos que se trata de una muestra aleatoria, representativa, en la que las variables son independientes, ¿cómo se estimaría la probabilidad conjunta observada?: $\Pr(Y_1 = 1 \text{ y } Y_2 = 0 \text{ y } Y_3 = 1 \mid X)$

En este caso se cumple que la probabilidad conjunta es igual a la multiplicación de las probabilidades individuales

Modelo de regresión logística con variable dependiente binaria o dummy

$$\Pr(Y_1 = 1 \text{ y } Y_2 = 0 \text{ y } Y_3 = 1 \mid X)$$

$$\frac{e^{x_1\beta}}{1+e^{x_1\beta}} \times \left(1 - \frac{e^{x_2\beta}}{1+e^{x_2\beta}}\right) \times \frac{e^{x_3\beta}}{1+e^{x_3\beta}}$$

Probabilidad de pagar

Probabilidad de no pagar

Modelo de regresión logística con variable dependiente binaria o dummy

$$\Pr(Y_1 = 1 \text{ y } Y_2 = 0 \text{ y } Y_3 = 1 \mid X)$$

$$\frac{e^{x_1\beta}}{1+e^{x_1\beta}} \times \left(1 - \frac{e^{x_2\beta}}{1+e^{x_2\beta}}\right) \times \frac{e^{x_3\beta}}{1+e^{x_3\beta}}$$

Probabilidad de pagar

Probabilidad de no pagar

Esta es la función de probabilidad conjunta cuando el argumento son las X.
Cuando en cambio el argumento son los β, se trata de una función de verosimilitud.

Modelo de regresión logística con variable dependiente binaria o dummy
Función de verosimilitud

En términos generales:

$$L = \prod_{i=1}^{n} P(Y_i = 1|x)^{Y_i} \cdot (1 - P(Y_i = 1|x))^{1 - Y_i}$$

Aplicado a la muestra de n=3

$$L = P(Y_1 = 1|x)^{Y_1} \cdot (1 - P(Y_1 = 1|x))^{1 - Y_1} \cdot P(Y_2 = 1|x)^{Y_2} \cdot$$
$$(1 - P(Y_2 = 1|x))^{1 - Y_2} \cdot P(Y_3 = 1|x)^{Y_3} \cdot (1 - P(Y_3 = 1|x))^{1 - Y_3}$$

Modelo de regresión logística con variable dependiente binaria o dummy
Función de verosimilitud

En términos generales:

$$L = \prod_{i=1}^{n} P(Y_i = 1|x)^{Y_i} \cdot (1 - P(Y_i = 1|x))^{1 - Y_i}$$

Aplicado a la muestra de n=3

$$L = P(Y_1 = 1|x)^{\overset{1}{Y_1}} \cdot (1 - P(Y_1 = 1|x))^{\overset{0}{1 - Y_1}} \cdot P(Y_2 = 1|x)^{\overset{0}{Y_2}} \cdot$$
$$(1 - P(Y_2 = 1|x))^{\overset{1}{1 - Y_2}} \cdot P(Y_3 = 1|x)^{\overset{1}{Y_3}} \cdot (1 - P(Y_3 = 1|x))^{\overset{0}{1 - Y_3}}$$

$$L = P(Y_1 = 1|X) \cdot (1 - P(Y_2 = 1|X)) \cdot P(Y_3 = 1|X)$$

Modelo de regresión logística con variable dependiente binaria o dummy
Función de verosimilitud

La función de verosimilitud obtenida es igual a la probabilidad conjunta

$$P(Y_1 = 1) \cdot P(Y_2 = 0) \cdot P(Y_3 = 1)$$

Pero la fórmula previa tiene la ventaja de poder ser aplicada para muchos casos, “encendiendo” o “apagando” las probabilidades indicadas

$$\prod_{i=1}^{n} \left(\frac{e^{x_i \beta}}{1 + e^{x_i \beta}} \right)^{Y_i} \cdot \left(\frac{1}{1 + e^{x_i \beta}} \right)^{1 - Y_i}$$

Modelo de regresión logística con variable dependiente binaria o dummy
Función de verosimilitud

$$L = \prod_{i=1}^{n} \left(\frac{e^{x_i \beta}}{1 + e^{x_i \beta}} \right)^{Y_i} \cdot \left(\frac{1}{1 + e^{x_i \beta}} \right)^{1 - Y_i}$$

El procedimiento busca maximizar la función de verosimilitud eligiendo los β

$$\max_{\{\beta\}} L$$

En la práctica, suele maximizarse el logartimo natural de la función de verosimilitud (porque es una transformación monótona en la que pasamos de estimar un producto a una sumatoria, lo cual facilita la maximización)

$$l = \ln L$$

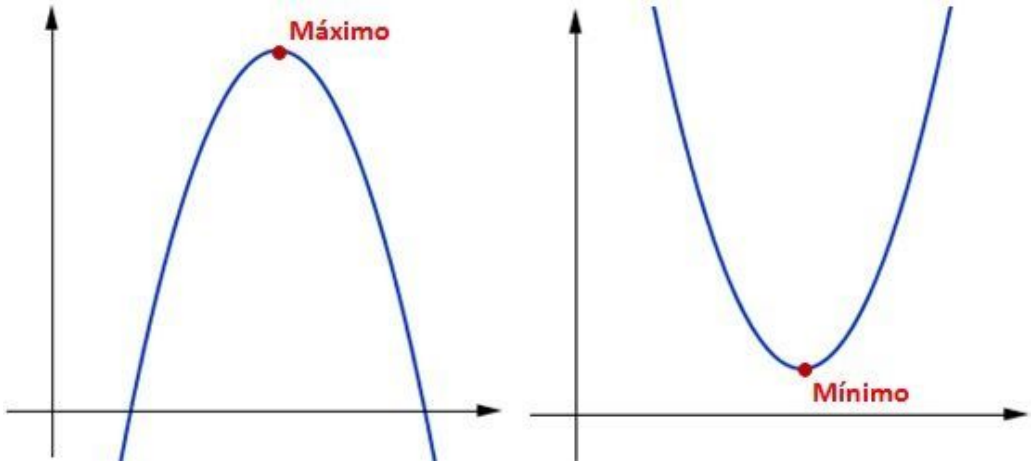
Modelo de regresión logística con variable dependiente binaria o dummy
Función de verosimilitud

Para encontrar el máximo de la función eligiendo los parámetros (β) la condición de primer orden es igualar la derivada de l con respecto a β a 0. En este punto la función no está aumentando ni disminuyendo, por lo que se trata de un punto crítico

$$CPO = \frac{dl}{d\beta} = 0$$

Modelo de regresión logística con variable dependiente binaria o dummy
Función de verosimilitud

$$CPO = \frac{dl}{d\beta} = 0$$



Este punto crítico puede ser un máximo o un mínimo. Como en este caso buscamos un máximo, es decir el punto en el que la función alcanza el valor más alto posible en un intervalo, necesitamos la condición de segundo orden.
La función sube hasta el punto máximo y luego comienza a bajar, así que en ese máximo la pendiente cambia de positiva a negativa, pasando por un punto en el que la pendiente es = 0.

Modelo de regresión logística con variable dependiente binaria o dummy
Función de verosimilitud

Condición de segundo orden:

Derivada segunda de l en ese punto negativa

$$\frac{d^2 l}{d\beta} < 0$$

Modelo de regresión logística con variable dependiente binaria o dummy
Función de verosimilitud

Los modelos que hacen esta maximización construyen las CPO asignando valores iniciales a los β para comprobar que las condiciones se cumplan con igualdad 0.
Se trata de un algoritmo de optimización que ajusta los valores iniciales hasta alcanzar el máximo. Es decir, hasta que todas las CPO se cumplan con igualdad 0.
Frecuentemente los softwares utilizan valores iniciales 0 para todos los coeficientes que acompañan las X y en caso de que con estos no se cumplan las CPO, prueban con 1, y así sucesivamente. Al modelo le lleva entre 3 y 5 iteraciones alcanzar el óptimo, en general.

Modelo de regresión logística con variable dependiente binaria o dummy
Interpretación de los coeficientes

En un modelo lineal:

$$\beta = \frac{dy}{dX} \longrightarrow$$ Captura el efecto marginal de X sobre Y

En un modelo no lineal:

$$Pr(Y_i = 1 \mid X) = \frac{e^{X_i\beta}}{1 + e^{X_i\beta}} = f(\beta_0 + \beta X_i)$$

$$\frac{d}{dX_i} f(\beta_0 + \beta X_i)$$

β no representa cuánto cambia la probabilidad de ocurrencia del evento cuando cambia la variable explicativa

Modelo de regresión logística con variable dependiente binaria o dummy
Interpretación de los coeficientes

$$\frac{d}{dX_i} f(\beta_0 + \beta X_i)$$

Cómo cambia la función de densidad varía según en qué punto de la curva nos encontremos. La altura de la función de densidad puede ser más alta o más baja, por lo que el efecto marginal no es constante a lo largo de toda la función.

La única interpretación posible respecto de los β es su signo: si es positivo el efecto de las variables explicativas sobre la probabilidad estimada es directo, y si es negativo, el efecto de las variables explicativas en la probabilidad es inverso. Pero no nos dice nada sobre la magnitud del impacto de X en Y.

Modelo de regresión logística con variable dependiente binaria o dummy
Interpretación de los coeficientes

En la práctica, se suele el efecto marginal en un punto. Frecuentemente se utiliza la media → efecto marginal promedio
Es decir, cuánto cambia la probabilidad de repago de un crédito cuando varía el ingreso del individuo promedio, por ejemplo.

Modelo de regresión logística con variable dependiente binaria o dummy
Interpretación de los coeficientes

Otra forma de interpretar los coeficientes del modelo logit es a través de los **Odds ratio**

$$CP = \frac{Pr(Y_i = 1|X_i = 1)}{Pr(Y_i = 0|X_i = 1)} / \frac{Pr(Y_i = 1|X_i = 0)}{Pr(Y_i = 0|X_i = 0)}$$

Nos indica cuánto más probable es que un individuo con Xi=1 (por ejemplo: una buena historia crediticia) repague el crédito con respecto a uno en el que Xi=0

Modelo de regresión logística con variable dependiente binaria o dummy
Medida de bondad de ajuste

En el modelo lineal utilizamos el R^2 que nos indica qué porcentaje de la variación de Y se explica por el modelo. Es decir:

$$\mathbf{R^2 = SCE / SCT}$$

Pero en un modelo no lineal, no estamos minimizando la suma de los residuos a cuadrado.
Se utilizan medidas de bondad de ajuste alternativas → Pseudo R^2
Uno de los más utilizados es el R^2 de McFadden:

$$R^2_{MF} = 1 - \frac{l(\hat{\beta_0}, \hat{\beta})}{l(\hat{\beta_0})}$$

Si β es 0, es decir que cambios en X no alteran la probabilidad de ocurrencia, esta medida dará 0. El R^2 MF va de 0 a 1 y tiene la misma interpretación que el $\mathbf{R^2}$