UNAP
FINESI

IV Semestre

Modelos Discretos

Regresión Logística Simple

Mtr. Alcides Ramos Calcina

Regresión Logística Simple

Mtr. Alcides Ramos Calcina

Introducción

- La regresión logística es una técnica analítica que nos permite relacionar funcionalmente una variable dicotómica con un conjunto de variables independientes.

- El análisis de regresión logística es muy frecuente en muchos campos de investigación, siendo especialmente empleado en investigación socio-sanitaria.

- En el análisis de datos en las ciencias sociales y las ciencias de la salud, su utilidad deriva de la lectura de los coeficientes -Odd Ratio- para interpretar los efectos que tienen las categorías sobre la variable dependiente.

- La regresión logística puede considerarse una extensión de los modelos de regresión lineal, con la particularidad de que la variable de respuesta está acotado al intervalo [0,1] y que el procedimiento de estimación, en lugar de mínimos cuadrados, utiliza el procedimiento de estimación máxima verosimilitud.

1. Definición

- Sea *Y* una variable dependiente binaria que toma dos valores posibles etiquetados como 0 y 1.

- Sean $X_1, X_2, \cdots, X_k$ un conjunto de variables independientes observadas con el de explicar y/o predecir el valor de Y.

- El objetivo es determinar $P[Y = 1 \mid X_1, \cdots, X_k]$

$$P[Y = 0 \mid X_1, \cdots, X_k] = 1 - P[Y = 1 \mid X_1, \cdots, X_k]$$

- Se construye un modelo de la forma: $P[Y = 1 \mid X_1, \cdots, X_k] = p(X_1, \cdots, X_k; \beta)$

Donde $p(X_1, \cdots, X_k; \beta)$ es una función que recibe el nombre de función de enlace (función de probabilidad) cuyo valor depende de un vector de parámetros

$$\beta = (\beta_1, \beta_2, \cdots, \beta_k)$$

2. Transformación Logit

- Consideremos la situación en que la variable Y es dicotómica tal que:

$$Y = \begin{cases} 1 & \textit{Presenta característica de interés} \\ 0 & \textit{En caso contrario} \end{cases}$$

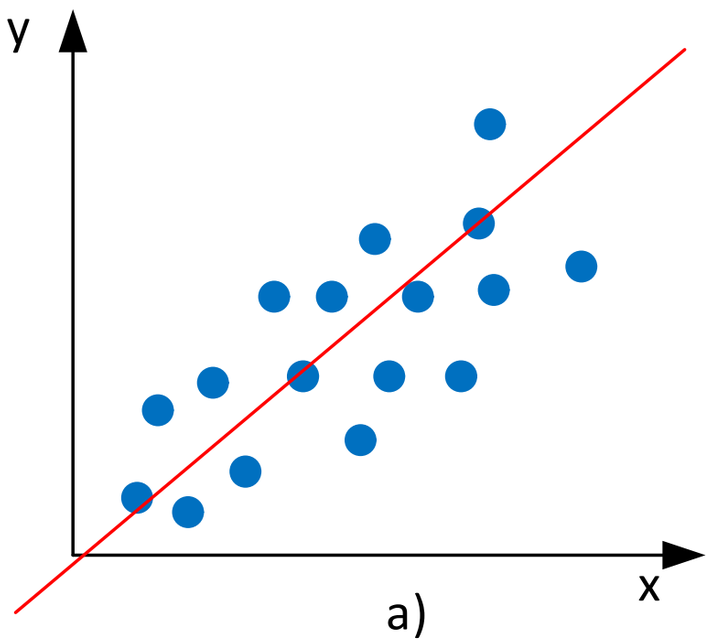
- La variable cualitativa con dos niveles (dicotómica) se codifica como 1 y 0, matemáticamente es posible ajustar un modelo de regresión lineal por mínimos cuadrados ordinarios (MCO)

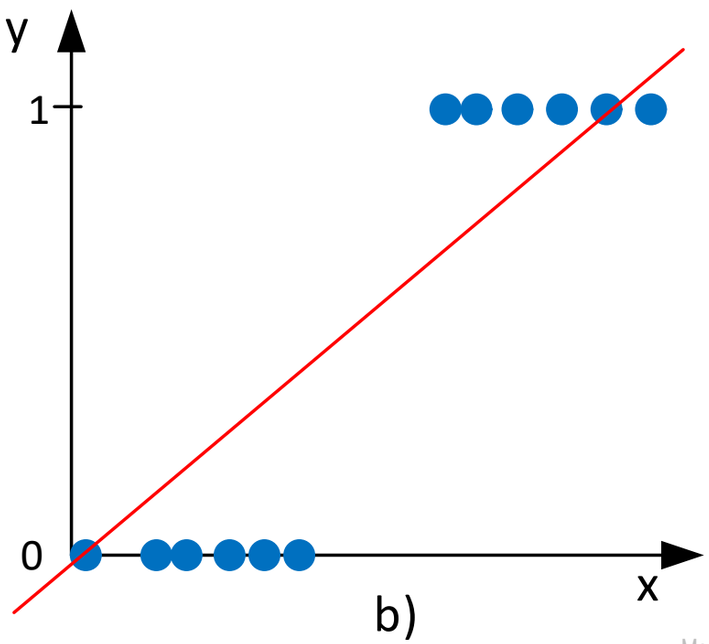
- El problema de esta aproximación es que, al tratarse de una recta, para valores extremos del predictor, se obtienen valores de Y menores que 0 o mayores que 1, lo que entra en contradicción con el hecho de que las probabilidades siempre están dentro del rango [0,1].

2. Transformación Logit

- En gráfico se observa los siguiente:

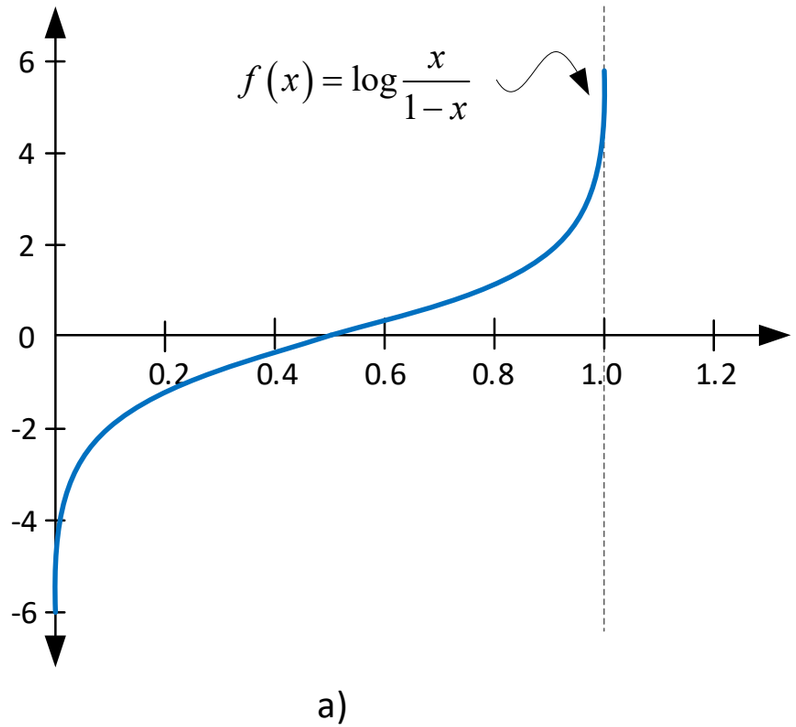
  a) la regresión lineal ayuda a predecir una variable numérica, pero no una variable dicotómica. En cambio, en el gráfico
  b) nos obliga a buscar una alternativa para predecir el evento que sea no lineal.:



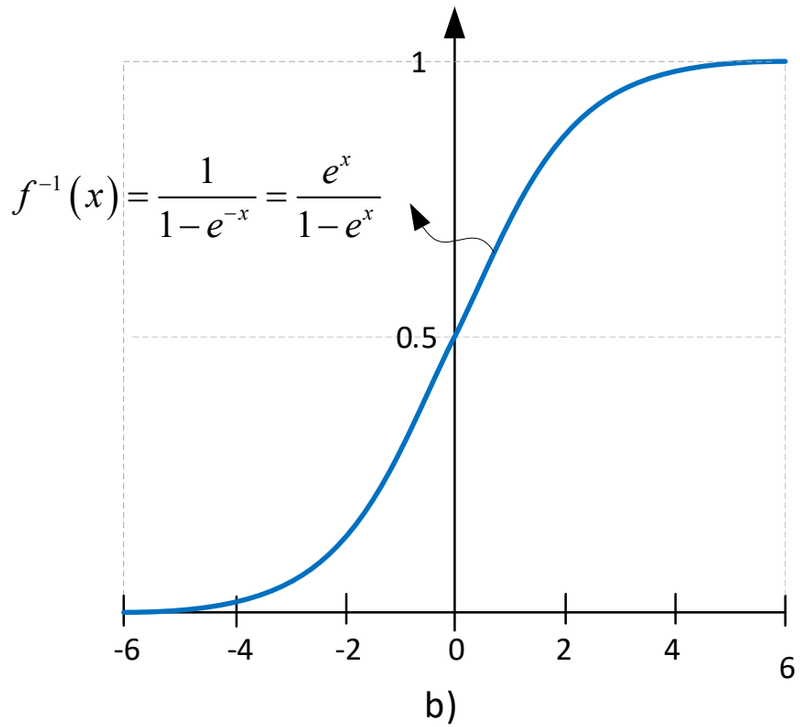


2. Transformación Logit

- Entonces esa alternativa es la función logit, que nos permitirá que los valores se encuentren entre [0,1].



Función Logit



Inversa de la función logit.

2. Transformación Logit

- Necesitamos que valores se encuentren entre [0, 1] en el eje Y, para ello es necesario ocupar la función inversa de logit, denominada función sigmoidal.

$$\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^{x}}{1+e^{x}}$$

- Para valores de x muy grandes positivos, el valor de e^{-x} es aproximadamente 0 entonces la función sigmoide es 1.

- Para valores de x muy grandes negativos, el valor e^{-x} tiende a infinito por lo que el valor de la función sigmoide es 0.

2. Transformación Logit

- Sustituyendo la x de la función sigmoide por la función lineal $\beta_0 + \beta_1 X$ se obtiene que:

$$P[Y = k \mid X = x] = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}} = \frac{1}{\frac{e^{\beta_0 + \beta_1 X}}{e^{\beta_0 + \beta_1 X}} + \frac{1}{e^{\beta_0 + \beta_1 X}}} = \frac{1}{\frac{1 + e^{\beta_0 + \beta_1 X}}{e^{\beta_0 + \beta_1 X}}}$$

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

Donde $P[Y = k \mid X = x]$ puede interpretarse como: la probabilidad de que la variable cualitativa Y adquiera el valor k (el nivel de referencia, codificado como 1), dado que el predictor X tiene el valor x.

$$P[Y = 1 \mid X = x] = p = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

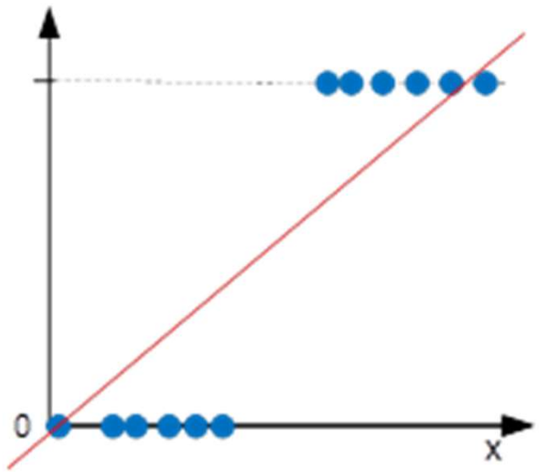
2. Transformación Logit

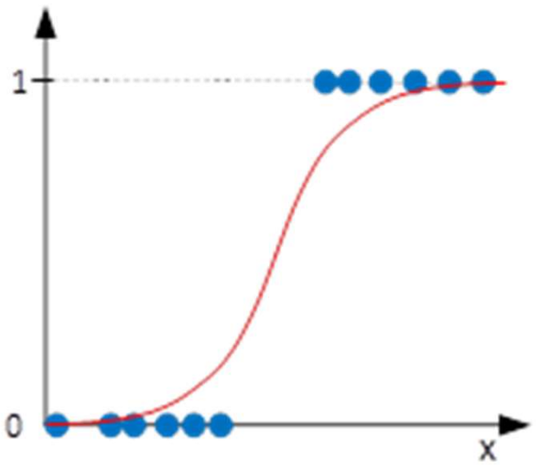
- Esta función, puede ajustarse de forma sencilla con métodos de regresión lineal si se emplea su versión logarítmica, obteniendo lo que se conoce como Log of Odds.

$$\log\left(\frac{P[Y=k \mid X=x]}{1-P[Y=k \mid X=x]}\right)=\beta_0+\beta_1 X \qquad \text{también} \qquad \log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1 X$$

- El siguiente gráfico nos muestra el ajuste de un modelo de regresión lineal y el ajuste del modelo logístico para Y,

a) Ajuste con regresión lineal





b) ajuste con regresión logística

2. Transformación Logit

- De esta manera el modelo que nos puede permitir, en principio, resolver problemas que tenemos planteado puede representarse de la forma:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

2.1. Odd y Odd Ratio

Para formular un modelo de regresión logística se hace una transformación de probabilidades a 'razones de probabilidades', o proporción de casos favorables a desfavorables. De forma habitual esta razón o ratio suele denominarse con el término Odd.

$$Odd = \frac{p}{(1-p)} = \frac{p}{q}$$

2. Transformación Logit

- El término Odd en inglés se refiere a la razón que se establece entre la ocurrencia -o su probabilidad- de un suceso respecto a su no ocurrencia.

- El Odd Ratio es una razón de Odds, -abreviadamente OR- y puede interpretarse como razón de probabilidades. Cuando el Odd Ratio alcanza el valor 1 quiere decir que no hay diferencias.

$$Odd\ Ratio = OR = \frac{Odd_A}{Odd_B} = \frac{\frac{p_A}{(1-p_A)}}{\frac{p_B}{(1-p_B)}}$$

2. Transformación Logit

2.2. La relación entre Odd y proporción. El Logit

A partir del Odd podemos definir el logit, simplemente como el logaritmo del Odd. Así:

$$Logit = \ln\left(Odd\right) = \ln\left(\frac{p}{1-p}\right)$$

Por ejemplo, podemos definir el Logit para un Odd Ratio. El Odd Ratio entre dos categorías de hombres y mujeres lo definimos como:

$$Odd\ Ratio = \frac{\frac{p_h}{1-p_h}}{\frac{p_m}{1-p_m}}$$

3. Características

Que hacemos con nuestro modelo de regresión logística entonces:

- **MODELA** la probabilidad de que ocurra un evento partiendo de un conjunto de variables.
- **ESTIMA** la probabilidad de que un evento ocurra para una observación al azar contra la probabilidad de que el evento no ocurra (ODDS).
- **PREDICE** el efecto de una serie de variables en una variable categórica binaria.
- **CLASIFICA** observaciones a través de la estimación de probabilidad de que se encuentre en una categoría determinada.

4. Prueba de significancia de los coeficientes

4.1. Estadístico de Wald

Una vez que se ha finalizado el proceso de máxima verosimilitud, se han encontrado los coeficientes betas, es importante saber si esos βi son o no distintos de cero.

$$H_0: \beta_k = 0 \qquad \text{vs} \qquad H_1: \beta_k \neq 0$$

Teniendo en cuenta el estimador sugerido se puede elaborar un test basado en el valor del coeficiente partido por la desviación típica elevado al cuadrado, que se ajusta a un χ_1^2

$$W = \left(\frac{\beta_i}{S_{\beta_i}} \right)^2 \approx \chi_1^2$$

Si $p < 0.05$ rechazaremos la H_0 y por tanto, ese valor será distinto de cero.

4. Prueba de significancia de los coeficientes

4.2. Devianza

Otro indicador importante para estudiar el ajuste del modelo logístico es la devianza que se define como el doble logaritmo del estadístico de verosimilitud, es decir,

$$devianza = -2\times\log -likelihood \text{ y se representa como } -2LL$$

Simplemente tomamos la devianza del nuevo modelo y le restamos la devianza del modelo referencia.

$$\chi^2 = 2LL(nulo) - 2LL(residual) \sim \chi^2_{(k-1)g.l.}$$

Esta diferencia se lo conoce como $ratio-likelihood$.

5. El modelo Logístico Binario Simple

Consideremos por ahora la situación en que sólo se dispone de una variable predictora, dicho en términos epidemiológicos, un solo *factor de riesgo*.

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

donde p representa la probabilidad de que un individuo presente la característica de interés y x es la única predictora.

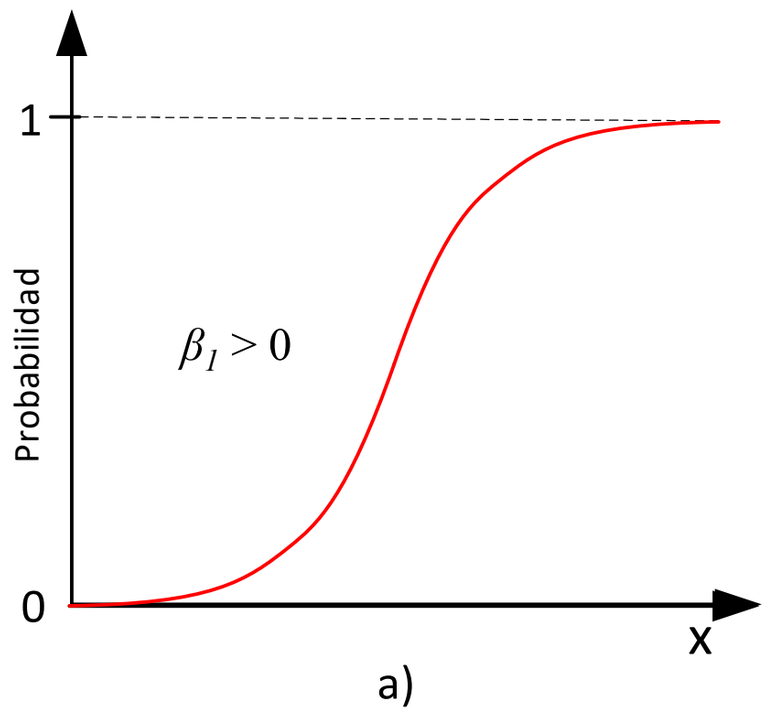
Expresión equivalente: $\dfrac{p}{1-p} = e^{\beta_0 + \beta_1 X}$

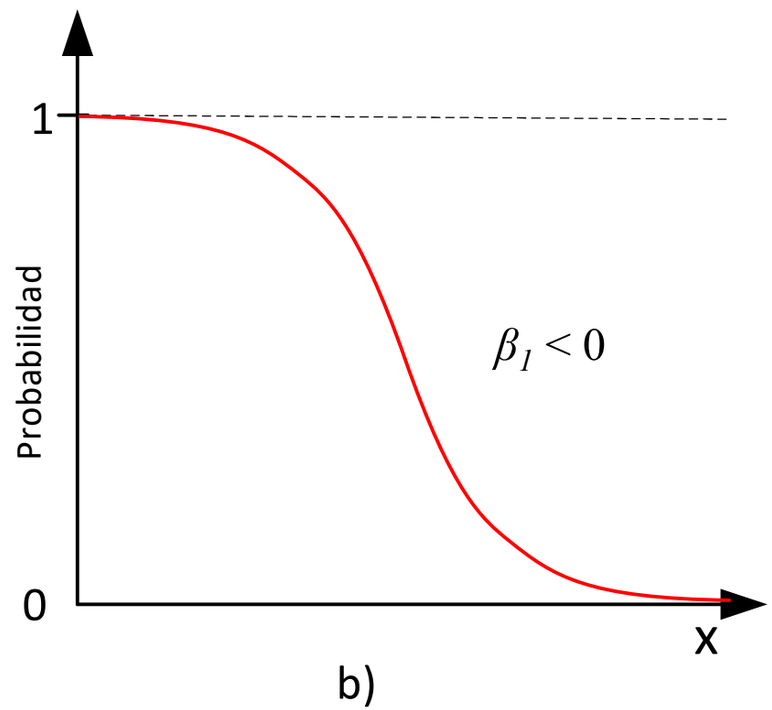
y despejando p obtenemos otra forma de escribir el modelo logístico

$$E(Y) = p = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

5. El modelo Logístico Binario Simple

Considerando la variable x continua, la representación gráfica del modelo es como se muestra en figura, dependiendo de que el parámetro β_1 sea positivo o negativo.





5. El modelo Logístico Binario Simple

Interpretación de los coeficientes

En las estimaciones de la regresión logística, los coeficientes miden el cambio en el logaritmo de la razón de probabilidad de éxito frente al fracaso (conocida como *odds*) cuando X incrementa en una unidad.

Los coeficientes del modelo $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$ no se interpretan,

se interpreta su exponencial $\frac{p}{1-p} = e^{\beta_0 + \beta_1 X} = e^{\beta_0}.e^{\beta_1 X}$

y puede adoptar la siguiente forma: $OR = e^{\beta}$

5. El modelo Logístico Binario Simple

Por lo que, el exponencial de β no es más que el *odds ratio* entre dos individuos que se diferencian en una unidad de la variable independiente. Entonces:

- $\beta = 0$ equivale a que $OR = 1$, es decir, que la variable independiente en cuestión no esta asociada a la probabilidad de enfermar.

- $OR > 1$ significa que por cada incremento en una unidad de la variable X, podemos apostar OR a 1 que la variable $P(Y = 1)$ le ocurriría dicho evento.

- $OR < 1$, se recomienda aplicar la inversa, para efectos de interpretación, es decir $1/OR$

6. Predicciones

Una vez estimados los coeficientes del modelo logístico, es posible conocer la probabilidad de que la variable dependiente pertenezca al nivel de referencia, dado un determinado valor del predictor. Para ello se emplea la ecuación del modelo:

$$\hat{p}\left[Y = 1 \mid X\right] = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}}$$

FINESI

Modelos Discretos

IV Semestre

https://aulavirtual2.unap.edu.pe/

GRACIAS