R-DEMO

LOGIT

PONENTE: DRA. ANA ESCOTO

DÍA 9

CONTENIDO

- Regresión logística simple
  - La función logística
  - Interpretación de coeficientes.
  - Supuestos
  - Regresión logística múltiple
- Ejemplos

¿POR QUÉ USAR LA REGRESIÓN LOGÍSTICA

- Hay muchos temas de investigación importantes para los cuales la variable dependiente es "limitada".
- Por ejemplo: el voto, la morbilidad o la mortalidad, y los datos de participación no son continuos o distribuidos normalmente.
- La regresión logística binaria es un tipo de análisis de regresión donde la variable dependiente es una variable ficticia: codificado 0 (no votó) o 1 (votó)

SI CORRIÉRAMOS UN MODELO LINEAL MCO

En la regression por OLS tendríamos:

- $Y = \gamma + \varphi X + e$; donde $Y = (0, 1)$

Los términos de error son heteroscedásticos.

e no se distribuye normalmente e porque Y toma solo dos valores

Las probabilidades predichas pueden ser mayores que 1 o menores que 0

MODELOS LINEALES GENERALIZADOS

- En Estadística el modelo lineal generalizado (GLM) es una generalización flexible de la regresión lineal ordinaria que permite variables de respuesta que tienen modelos de distribución de errores distintos a una distribución normal.
- El GLM generaliza la regresión lineal al permitir que el modelo lineal se relacione con la variable de respuesta a través **de una función de enlace** y al permitir que la magnitud de la varianza de cada medición sea una función de su valor predicho

MODELOS LINEALES GENERALIZADOS

- Familia de modelos de regresión.
  - Tipo de modelo de respuesta

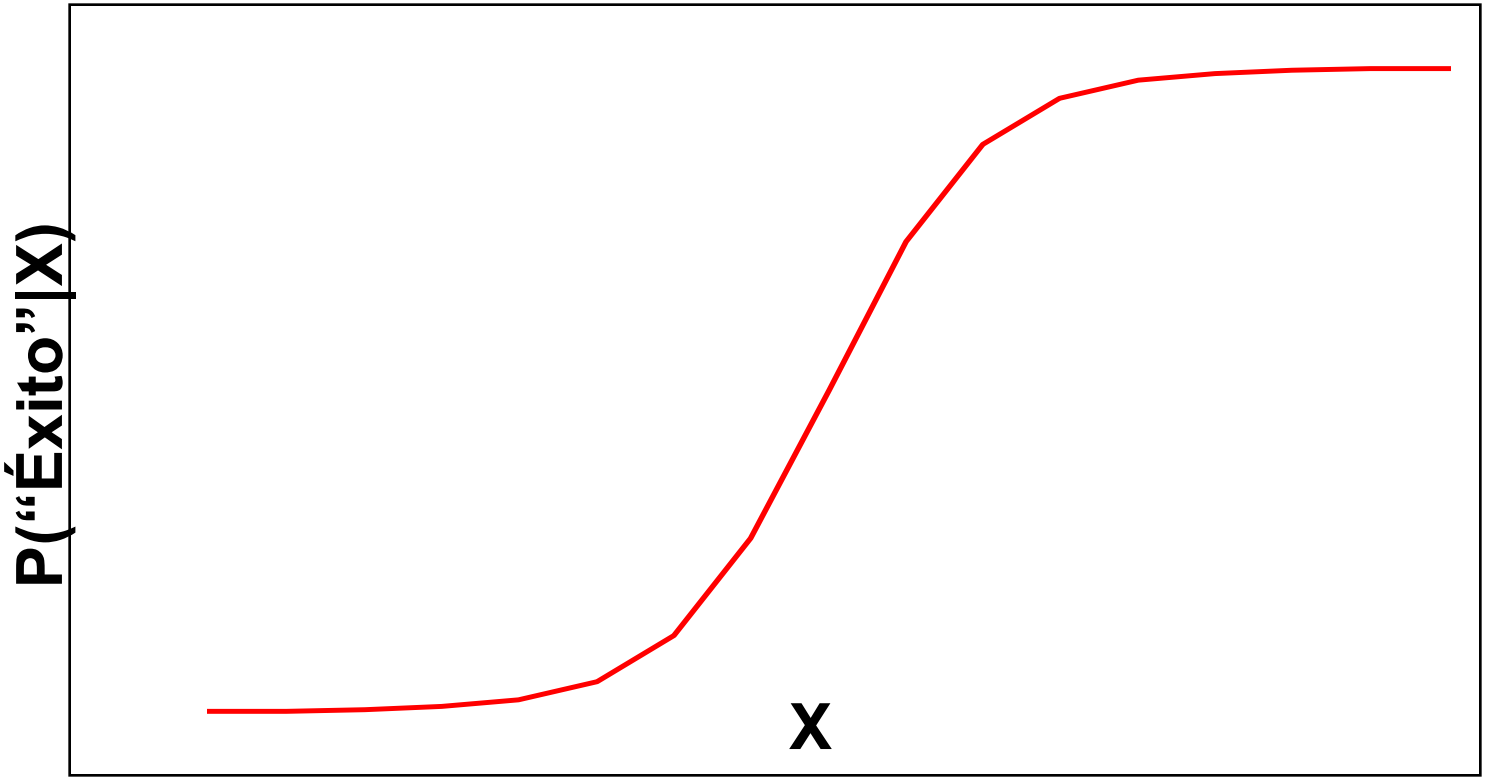
Respuesta	Tipo de modelo
Continuo	Regresión lineal
Cuenta	Regresión de poisson
Tiempos de supervivencia	Modelo de cox
Binomial	Regresión logística

- Usos
- Control para factores de confusión potenciales.
- Construcción de modelos, predicción de riesgos.

LOGISTIC REGRESSION

- Relación de modelos entre conjunto de variables Xi.
  - dicotómico (sí / no, fumador / no fumador, ...)
  - categóricos (clase social, raza, ...)
  - continuas (edad, peso, ingreso)
- y
  - Respuesta categórica dicotómica variable Y
  - p.ej. Éxito / Fracaso,

FUNCIÓN LOGÍSTICA



$$P("Success"|\ X) = \frac{e^{\beta_o + \beta_1 X}}{1 + e^{\beta_o + \beta_1 X}}$$

EL LOGITO

- Nuestra probabilidad de éxito se modela así

$$P(Y \mid X) = \frac{e^{\beta_o + \beta_1 X}}{1 + e^{\beta_o + \beta_1 X}}$$

- Que es equivalente a

$$\underbrace{\ln\left(\frac{P(Y \mid X)}{1 - P(Y \mid X)}\right)} = \beta_o + \beta_1 X$$

Esto es el logito: el logaritmo de los momios

UN PREDICTOR DICOTÓMICO: LA RAZÓN DE MOMIOS

	Pobreza (X)	
Voto (Y)	Pobreza (X = 1)	No Pobreza (X = 0)
Yes (Y = 1)	$P(Y=1\|X=1)$	$P(Y=1\|X=0)$
No (Y = 0)	$1-P(Y=1\|X=1)$	$1-P(Y=1\|X=0)$

- Con una variable explicativa dicotómica(X) que representa (1 =Pobreza)

$$\frac{P}{1-P} = e^{\beta_o+\beta_1 X}$$

En caso de pobreza $\dfrac{P(Y=1\,|\,X=1)}{1-P(Y=1\,|\,X=1)} = e^{\beta_o+\beta_1}$

En caso de no pobreza $\dfrac{P(Y=1\,|\,X=0)}{1-P(Y=1\,|\,X=0)} = e^{\beta_o}$

Por tanto la razón de momios sería

$$= \frac{e^{\beta_o+\beta_1}}{e^{\beta_o}} = e^{\beta_1}$$

1.1 aumentó .10

0.9 disminyó .10

ENTONCES… TENEMOS

- Por lo tanto, para el odds ratio asociado a la presencia de riesgo tenemos

$$OR = e^{\beta_1}$$

- Tomando el logaritmo natural que tenemos.

$$\ln(OR) = \beta_1$$

- por lo tanto, el coeficiente de regresión estimado asociado con un predictor dicotómico codificado 0-1 es el registro natural de la OR asociado con la presencia de riesgo.

El modelo logístico puede ser escrito.

$$\ln\left(\frac{P(Y \mid X)}{1-P(Y \mid X)}\right) = \ln\left(\frac{P}{1-P}\right) = \beta_o + \beta_1 X$$

Esto implica que las probabilidades de éxito pueden expresarse como

$$\frac{P}{1-P} = e^{\beta_o + \beta_1 X}$$

Esta relación es la clave para interpretar los coeficientes en un modelo de regresión logística.

SUPUESTOS

- La única limitación "real" en la regresión logística es que el resultado debe ser discreto.
- Relación de casos a variables: el uso de variables discretas requiere que haya suficientes respuestas en cada categoría dada
- Si hay demasiadas celdas sin respuesta, las estimaciones de los parámetros y los errores estándar probablemente explotarán
- También puede hacer que los grupos sean perfectamente separables (por ejemplo, multicolineales) lo que hará imposible la estimación de máxima probabilidad.

SUPUESTOS

- Linealidad en el logit: la ecuación de regresión debe tener una relación lineal con la forma logit de la VD. No se supone que los predictores estén relacionados linealmente entre sí..
- Ausencia de multicolinealidad.
- No hay valores atípicos
- Independencia de errores
  - Si se asume un diseño entre sujetos.
  - Hay otras formas si el diseño está dentro de los sujetos.

EN POCAS PALABRAS

- En regresión logística, la respuesta (Y) es una variable categórica dicotómica.
- Las estimaciones de los parámetros dan la razón de probabilidades asociadas a las variables en el modelo.
- Estos odds ratios se ajustan para las otras variables en el modelo.
- También se puede calcular P (Y | X) si eso es de interés, por ejemplo. Dado la condición de pobreza y otro resto de vars, ¿cuál es la probabilidad estimada de tener que se vote por cierto candidato?

PARA PROFUNDIZAR

- https://stats.stackexchange.com/questions/86351/interpretation-of-rs-output-for-binomial-regression
- https://stats.stackexchange.com/questions/20523/difference-between-logit-and-probit-models/30909#30909
- https://statisticalhorizons.com/hosmer-lemeshow
- Agresti, A. (2002), *Categorical data analysis*, New York, Wiley-Interscience.
- Hosmer, D. W., Jr., S. A. Lemeshow, and R. X. Sturdivant. 2013. Applied Logistic Regression. 3rd ed. Hoboken, NJ: Wiley.