2. REGRESIÓN LOGISTICA BINOMIAL

Regresión Logística

- En estadística, la **regresión logística** es un modelo de regresión para variables dependientes o de respuesta dicotómica con distribución binomial aproximada.

- Es útil para modelar la probabilidad de un evento ocurriendo como función de otros factores.

Regresión Logística

- Objetivo: Modelar la probabilidad de una característica en particular como una función de una o mas variables explicativas o predictivas.

- La distribución de probabilidad Binomial es comúnmente utilizada para datos dicotómicos, que es de la forma:

$$Y_i \sim B(P, n)$$

- Donde se consideran n ensayos Bernoulli conocidos con probabilidad de éxito P desconocida y con valores que se limitan entre 0 y 1

FUNCIÓN LIGA LOGIT

- Variable respuesta - Dicotómica (con dos posibles resultados)
- Variables explicativas - numéricas y/o categóricas
- Es un modelo lineal generalizado (glm) con función liga Logit (transformación logit):

$$g(P) = \log\left(\frac{P}{1-P}\right)$$

- Las probabilidades binomiales desconocidas son representadas por el logaritmo del momio y modeladas como una función lineal de las *k* variables explicativas ($j = 0,\ldots, k$).

$$g(P) = \log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

- Los parámetros desconocidos β_j son usualmente estimados a través del método de máxima verosimilitud.

Regresión Logística con una variable predictiva

- Respuesta- Presencia/Ausencia de cierta característica
- Predictiva - Variable numérica observada en cada caso
- Modelo - $P(x) \equiv$ Probabilidad de presencia al nivel x de la variable predictiva

$$P(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

- $b_1 = 0 \Rightarrow$ P(Presencia) es la misma a cualquier nivel de x
- $b_1 > 0 \Rightarrow$ P(Presencia) se incrementa al incrementar x
- $b_1 < 0 \Rightarrow$ P(Presencia) disminuye al incrementar x

Regresión Logística con una variable predictiva

- b_0 y b_1 son parámetros desconocidos y deben estimarse con apoyo de paquetes estadísticos.
- El interés se centra en la estimación y la prueba de hipótesis sobre b_1

Prueba de Wald (Muestras grandes):

H_0: $b_1 = 0$ H_A: $b_1 \neq 0$

$$X_c^2 = \left(\frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \right)^2$$

$$X_c^2 \geq \chi^2_{\alpha,1}$$

$$P(\chi^2 \geq X_c^2)$$

Regresión Logística con una variable predictiva dicotómica

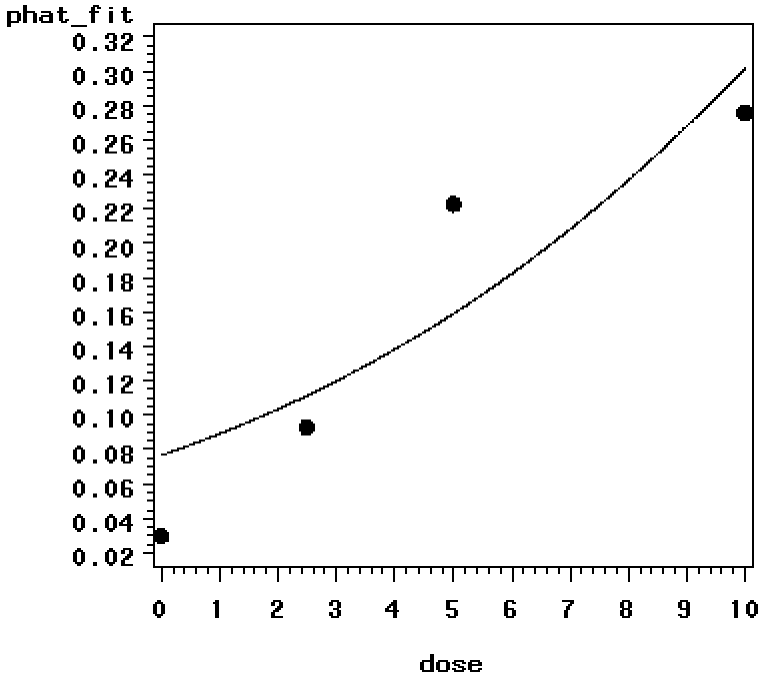
Variable Respuesta	Variable predictiva X	
	x = 1	x = 0
y = 1	$P(1)=\dfrac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$	$P(0)=\dfrac{e^{\beta_0}}{1+e^{\beta_0}}$
y = 0	$1-P(1)=\dfrac{1}{1+e^{\beta_0+\beta_1 x}}$	$1-P(0)=\dfrac{1}{1+e^{\beta_0}}$

Ejemplo - Rizatriptan para la Migraña

- Respuesta – Alivio de Dolor Efectivo en 2 horas (Si/No)
- Predictiva - Dosis (*mg*): Placebo (0),2.5,5,10

Dosis	# de Pacientes	# de Aliviados	% de Aliviados
0	67	2	3.0%
2.5	75	7	9.3%
5	130	29	22.3%
10	145	40	27.6%

Ejemplo - Rizatriptan para la Migraña



$$\hat{P}(x) = \frac{e^{-2.490+0.165x}}{1+e^{-2.490+0.165x}}$$

$$H_0 : \beta = 0 \quad H_A : \beta \neq 0$$

$$X^2_{obs} = \left(\frac{0.165}{0.037}\right)^2 = 19.819$$

$$X^2_{obs} \geq \chi^2_{.05,1} = 3.84$$

$$P = 0.000$$

Razón de Momios (Odds Ratio)

- Interpretación del Coeficiente de Regresión (β):
  - Se puede demostrar que:

$$\frac{odds(x+1)}{odds(x)} = e^{\beta} \qquad \left(odds(x) = \frac{P(x)}{1-P(x)} \right)$$

• Por lo tanto e^b representa el cambio en el momio (odds) de la respuesta (multiplicativamente) al incrementar x en 1 unidad

• Si $b = 0$, el momio y la probabilidad son lo mismo a todos los niveles de x (e^b=1)

• Si $b > 0$, el momio y la probabilidad se incrementan al incrementar x (e^b>1)

• Si $b < 0$, el momio y la probabilidad disminuyen al aumentar x (e^b<1)

Intervalo de Confianza al 95% para la Razón de Momios

- Paso1: Construir un IC al 95% para *b* :

$$\hat{\beta} \pm 1.96\hat{\sigma}_{\hat{\beta}} \equiv \left(\hat{\beta} - 1.96\hat{\sigma}_{\hat{\beta}} \, , \, \hat{\beta} + 1.96\hat{\sigma}_{\hat{\beta}} \right)$$

- Paso 2: Elevar la base *e* = 2.718 a la potencia de los límites inferior y superior del IC:

$$\left(e^{\hat{\beta} - 1.96\hat{\sigma}_{\hat{\beta}}} \, , \, e^{\hat{\beta} + 1.96\hat{\sigma}_{\hat{\beta}}} \right)$$

- Si el intervalo esta sobre 1, se concluye que existe asociación positiva
- Si el intervalo está 1, se concluye que existe asociación negativa
- Si el intervalo contiene al 1, no se puede concluir que existe asociación

Ejemplo - Rizatriptan para la Migraña

- IC al 95% para *b* :

$$\hat{\beta} = 0.165 \qquad \hat{\sigma}_{\hat{\beta}} = 0.037$$

$$0.165 \pm 1.96(0.037) \quad \equiv \quad (0.0925, 0.2375)$$

- IC al 95% para la razón de momios poblacional:

$$\left(e^{0.0925}, e^{0.2375}\right) \equiv (1.10, 1.27)$$

- Se concluye que existe asociación positiva entre la dosis y la probabilidad de alivio efectivo en 2 hrs.

Regresión Logística Múltiple

- Extensión a más de una variable predictiva (variables numéricas y/0 dummy). Con *k* predictivas, el modelo se escribe como:

$$P(\underline{x}) = \frac{e^{\alpha+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\alpha+\beta_1 x_1+\cdots+\beta_k x_k}}$$

- Razón de Momios Ajustada para el incremento de x_i en 1 unidad, manteniendo constantes todas las variables predictivas adicionales:

$$OR_i = e^{\beta_i}$$

- Muchos modelos tienen variables predictivas de tipo nominal y ordinal, por lo que comúnmente se utilizan variables dummy

Prueba de los Coeficientes de Regresión

- Prueba del modelo completo:

Hipótesis estadísticas:

$$H_0 : \beta_1 = \cdots = \beta_k = 0$$

$$H_A : \text{No todas las } \beta_i = 0$$

Estadístico de prueba:

$$X^2_{obs} = (-2\log(L_0)) - (-2\log(L_1))$$

Criterio de decisión:

$$X^2_{obs} \geq \chi^2_{\alpha,k}$$

Significancia:

$$P = P(\chi^2 \geq X^2_{obs})$$

- L_0, L_1 son valores que maximizan la función de verosimilitud, calculados iterativamente por los paquetes estadísticos. Esta lógica se usa para comparar el modelo completo y modelos reducidos con base en subconjuntos de variables predictivas.

Disfunción Eréctil en Hombres Mayores

- Respuesta: Presencia/Ausencia de DE (n=1688)
- Predictivas:
  - Grupo de Edad (50-54*, 55-59, 60-64, 65-69, 70-78)
  - Fuma (No Fumador*, Fumador)
  - Grupo de IMC (<25*, 25-30, >30)
  - Síntomas del Tracto Urinario STU (No*, Leve, Moderado, Severo)
  - Tratamiento de síntomas cardiacos (No*, Si)
  - Tratamiento de Enfermedad pulmonar obstructiva crónica EPOC (No*, Si)

* Grupo control para variables Dummy.

Disfunción Eréctil en Hombres Mayores

Predictiva	b	s_b	OR Ajustado (IC 95%)
Edad 55-59 (vs 50-54)	0.83	0.42	2.3 (1.0 – 5.2)
Edad 60-64 (vs 50-54)	1.53	0.40	4.6 (2.1 – 10.1)
Edad 65-69 (vs 50-54)	2.19	0.40	8.9 (4.1 – 19.5)
Edad 70-78 (vs 50-54)	2.66	0.41	14.3 (6.4 – 32.1)
Fumador (vs no fumador)	0.47	0.19	1.6 (1.1 – 2.3)
IMC 25-30 (vs <25)	0.41	0.21	1.5 (1.0 – 2.3)
IMC >30 (vs <25)	1.10	0.29	3.0 (1.7 – 5.4)
STU Leve (vs No)	0.59	0.41	1.8 (0.8 – 4.3)
STU Moderado (Si vs No)	1.22	0.45	3.4 (1.4 – 8.4)
STU Severo (Si vs No)	2.01	0.56	7.5 (2.5 – 22.5)
T. de S. cardiacos (Si vs No)	0.92	0.26	2.5 (1.5 – 4.3)
EPOC (Si vs No)	0.64	0.28	1.9 (1.1 – 3.6)

Interpretación: El riesgo de Disfunción Eréctil :

- Incrementa con la edad, el IMC, y los STU
- Es mayor en fumadores
- Es mayor en hombres bajo tratamiento cardiaco o EOPC

FUNCIÓN PROBIT

- Se le llama Función Probit a la inversa de la función de distribución normal o función cuantil asociada con la Distribución Normal Estándar.

- la distribución normal estándar (a menudo denotada por N(0,1)) la función de distribución se denota comúnmente por Φ.

- Φ es una función sigmoide continua y creciente, cuyos dominio y recorrido son la recta real en el intervalo (0, 1).

FUNCIÓN PROBIT

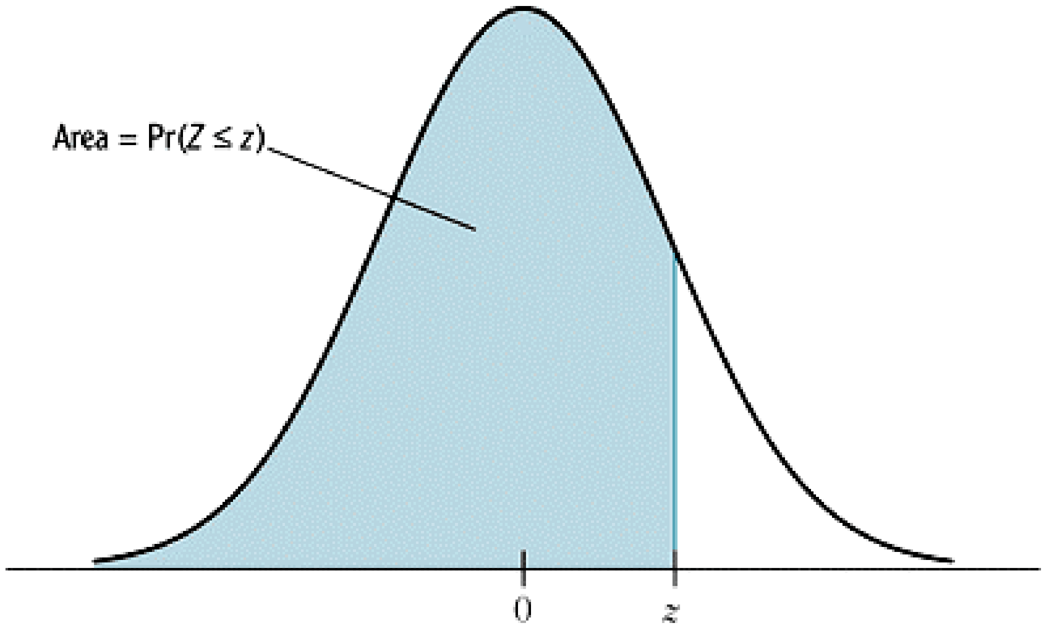
- Por ejemplo, la distribución N(0, 1) tiene un 95% de probabilidad entre -1,96 y 1,96 y es simétrica en un entorno de cero. De ahí se deduce que:

$$\Phi(-1{,}96) = 0{,}025 = 1 - \Phi(1{,}96)$$

- La función probit es el cálculo inverso, generando un valor de una variable aleatoria N(0, 1) asociado a una probabilidad acumulada bajo su curva.

- Formalmente, la función probit es la inversa de $\Phi(z)$, denotada $\Phi^{-1}(p)$.

TABLE 1 The Cumulative Standard Normal Distribution Function, $\Phi(z)$ = Pr(Z ″ z)



z	Second Decimal Value of z									
	0	1	2	3	4	5	6	7	8	9
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019

REGRESIÓN PROBIT

- La regresión probit simple, representa la probabilidad de que Y = 1 usando la función de distribución normal estándar evaluada en $z = \beta_0 + \beta_1 X$:

$$\Pr(Y = 1|X) = \Phi(\beta_0 + \beta_1 X)$$

- $z = \beta_0 + \beta_1 X$ es el predictor lineal en el modelo probit

REGRESIÓN PROBIT

La curva "en forma de S" nos da lo que queremos:

- $0 \leq Pr(Y = 1|X) \leq 1$ para todo X
- $Pr(Y = 1|X)$ creciente en X (para $\beta_1 > 0$)
- Tiene una interpretación relativamente directa:
- El valor $z = b_0 + b_1X$ es el valor z estimado, dado X
- β_1 es el cambio en el valor z para un cambio unitario en X

REGRESIÓN PROBIT

Regresión probit con varios regresores

$$\Pr(Y = 1|X_1, X_2) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$$

- Φ es la función de distribución normal acumulada.
- $z = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ es el predictor lineal del modelo probit
- β_1 es el efecto en el "valor z" de un cambio unitario en X_1, manteniendo constante X_2