Modelo Lineal Generalizado

Introducción

Comenzaremos con un ejemplo que nos servirá para ilustrar el análisis de datos binarios.

Nuestro interés se centra en relacionar una estructura estocástica en los datos que siguen una distribución binomial y una estructura sistemática en términos de alguna transformación de las variables independientes.

Los siguientes datos tomados de Little (1978) corresponden a 1607 mujeres casadas y fértiles entrevistadas por la Encuesta de Fertilidad Fiji de 1975, cla-

sificadas por edad, nivel de educación, deseo de tener más hijos y el uso de anticonceptivos.

Edad	Educación	Más Hijos?	Uso de Anticonceptivos		Total
			No	Si	
< 25	Baja	Si	53	6	59
		No	10	4	14
	Alta	Si	212	52	264
		No	50	10	60
25–29	Baja	Si	60	14	74
		No	19	10	29
	Alta	Si	155	54	209
		No	65	27	92
30–39	Baja	Si	112	33	145
		No	77	80	157
	Alta	Si	118	46	164
		No	68	78	146
40–49	Baja	Si	35	6	41
		No	46	48	94
	Alta	Si	8	8	16
		No	12	31	43
Total			1100	507	1607

En este ejemplo se considera a la variable *Anticoncepción* como dependiente y a las demás como predictoras. En este caso, todas las predictoras son variables categóricas, sin embargo el modelo que presentaremos permite introducir variables independientes continuas y discretas.

El objetivo es decribir cómo el uso de métodos anticonceptivos varía según la *edad, el nivel de educación y el deseo de tener más hijos*.

Por ejemplo, una pregunta que sería interesante responder es si hay asociación entre educación y anticoncepción, por ejemplo, podríamos querer saber si las mujeres con un nivel de educación más elevado prefieren familias más chicas que las mujeres con niveles de educación inferior.

Componente Aleatoria

La componente aleatoria del modelo involucra a las respuestas Y_i.

Definamos

$$Y_i = \begin{cases} 1 & \text{si usa anticonceptivo} \\ 0 & \text{si no} \end{cases}$$

Y_i toma los valores 1 y 0 con probabilidad π_i y $1 - \pi_i$, respectivamente,y por lo tanto

$$\begin{aligned} E(Y_i) &= \pi_i \\ Var(Y_i) &= \pi_i(1 - \pi_i)\,. \end{aligned}$$

Tanto la media como la varianza dependen de i, por lo tanto cualquier factor que afecte la esperanza también afectará la varianza. Esto sugiere que cualquier modelo que, como el lineal, asuma homoscedasticidad de las observaciones no será adecuado para este problema.

En el ejemplo, de acuerdo con el valor de las variables predictoras, las observaciones pueden ser clasificadas en 16 grupos. Si n_i es el número de observaciones del grupo i e Y_i denota al número de *éxitos*, tendremos que

$$Y_i \sim Bi(n_i, \pi_i)\,.$$

En nuestro caso,

Y_i = número de mujeres que usan anticonceptivos en el i–ésimo grupo.

Luego,

$$\begin{aligned} P(Y_i = k) &= \binom{n_i}{k} \pi_i^k (1-\pi_i)^{n_i-k} \\ E(Y_i) &= n_i \pi_i \\ Var(Y_i) &= n_i \pi_i (1-\pi_i)\,, \end{aligned}$$

para $k = 0, \ldots, n_i$.

Componente sistemática

La componente sistemática del modelo involucra a las covariables $\mathbf{x}_i$ que participan.

El modelo más sencillo podría expresar a π_i como una combinación lineal de las variables independientes:

$$\pi_i = \mathbf{x}_i' \boldsymbol{\beta} \,,$$

siendo $\boldsymbol{\beta}$ el vector de parámetros a estimar.

Este modelo recibe el nombre de **modelo de probabilidad lineal** y su estimación puede basarse en mínimos cuadrados ordinarios.

Un problema evidente de este modelo es que las probabilidades π_i son acotadas, mientras que las $\mathbf{x}_i' \boldsymbol{\beta}$ pueden tomar cualquier valor real. Si bien esto podría controlarse imponiendo complicadas restricciones a los coeficientes, esta solución no resulta muy natural.

Una solución sencilla es *transformar* la probabilidad mediante una función que mapee el intervalo $(0, 1)$ sobre la recta real y luego modelar esta transformación como una función lineal de las variables independientes.

Una manera posible es mediante los **odds** (o chances) definidos como

$$\psi = \frac{\pi}{1-\pi},$$

π	ψ
0.1	0.11
0.2	0.25
0.5	1
0.6	4
0.9	9

De manera que odds menores que 1 están asociados a probabilidades menores que 0.5 y odds mayores que 1 están asociados a probabilidades mayores que 0.5.

Sin embargo, esta transformación sólo mapea sobre los reales positivos. Para extenderla a los negativos introduciremos el log:

$$\text{logit}(\pi) = \log\left[\frac{\pi}{1-\pi}\right] = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p = \mathbf{x}'\boldsymbol{\beta} = \eta$$

La función logit es estrictamente creciente y tiene inversa:

$$\pi = \text{logit}^{-1}(\eta) = \frac{e^{\eta}}{1+e^{\eta}}\,.$$

En el ejemplo tenemos: 507 mujeres usan anticonceptivos entre las 1607, por lo que estimamos la probabilidad como $\frac{507}{1607}$=0.316. Luego, los odds se estimarían por

$$\frac{507/1607}{1100/1607} = \frac{507}{1100} = 0.461$$

Entonces, aproximadamente por cada mujer que usa anticonceptivos hay dos que no los usan. El $\text{logit}(0.461) = \text{-}0.775$.

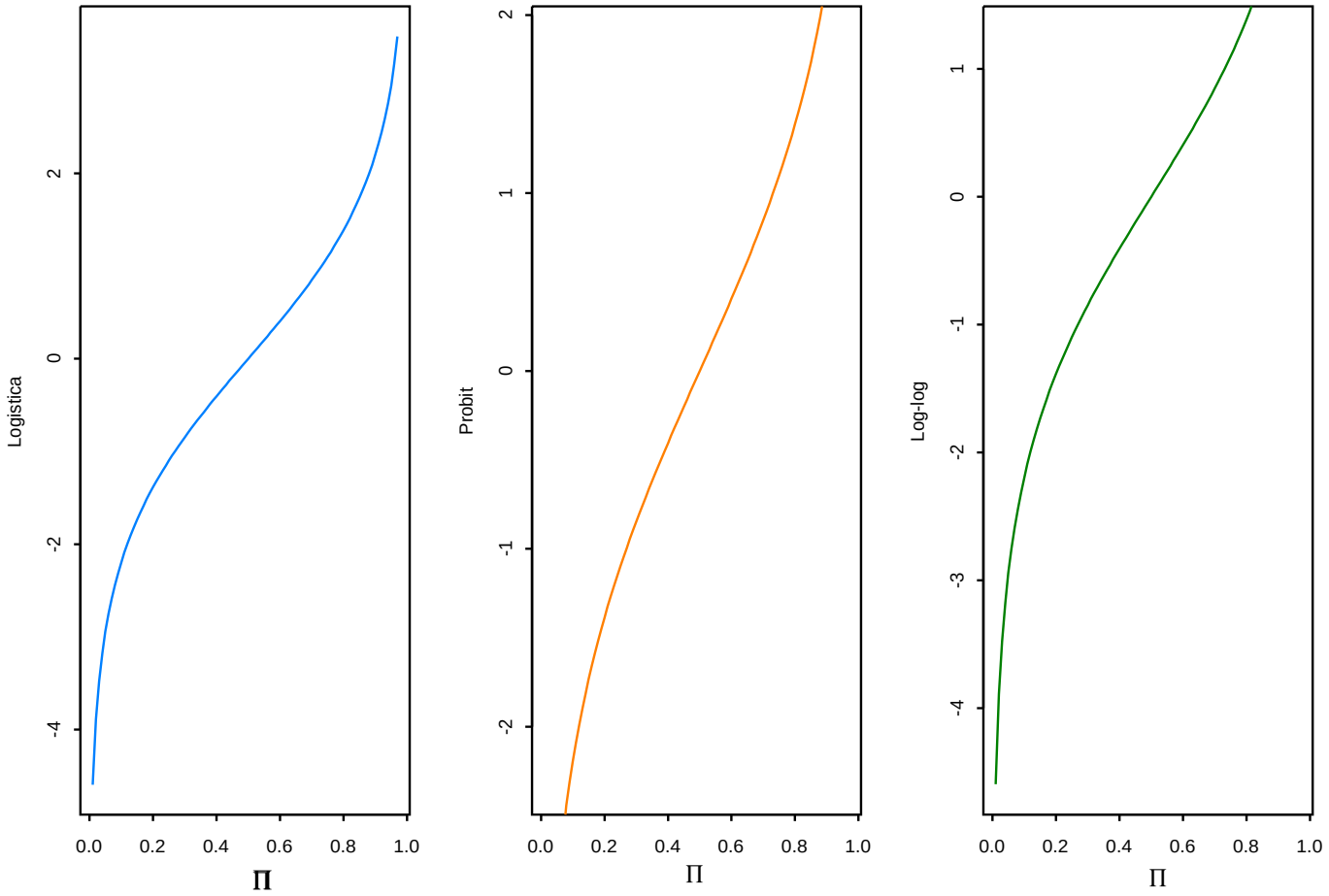

Figura 1: Funciones de enlace o link

Modelo de Regresión Logística

Sean $Y_1, \ldots, Y_n$ v.a. independientes tales que

$$Y_i \sim Bi(n_i, \pi_i)\,. \tag{1}$$

Esto define la componente aleatoria.

Supongamos además que la probabilidad π_i es una función de los predictores:

$$\text{logit}(\pi_i) = \mathbf{x}_i'\boldsymbol{\beta}\,, \tag{2}$$

donde las $\mathbf{x}_i$ son las covariables.

Esto define la componente sistemática del modelo.

El modelo definido por (1) y por (2) es un **modelo lineal generalizado** con respuesta binomial y función de enlace logit.

Los coeficientes $\boldsymbol{\beta}$ tienen una interpretación similar a la que tienen en el modelo lineal, pero debemos tener en cuenta que el miembro de la derecha es un logit y no una media.

Los β_j representan entonces el cambio en el logit de la probabilidad asociada cuando hay un cambio de una unidad en el j–ésimo predictor y se matienen constantes todas las demás variables.

Como

$$\pi = \frac{e^{\mathbf{x}'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}'\boldsymbol{\beta}}} = \frac{1}{1 + e^{-\mathbf{x}'\boldsymbol{\beta}}},$$

la relación con π es no lineal, luego no es tan sencillo como en el modelo lineal expresar el cambio en π_i al cambiar un predictor.

Cuando el predictor es continuo, podemos hacer una aproximación tomando derivadas con respecto a la j–ésima coordenada de $\mathbf{x}$, obteniendo

$$\frac{\partial \pi}{\partial x_j} = \beta_j \pi(1-\pi)\,.$$

Luego, el efecto del j–ésimo predictor depende del coeficiente β_j y de la probabilidad π.

Una vez establecido el modelo que queremos ajustar haremos las diferentes etapas de inferencia habituales:

- estimar los parámetros
- hallar intervalos de confianza para los mismos
- evaluar la bondad del ajuste
- realizar algún test de interés que involucre a los parámetros

También tendremos que evaluar la influencia de las observaciones en la determinación de los valores estimados.

Modelo Lineal Generalizado

En el Modelo Lineal Generalizado (GLM) tenemos variables de respuesta asociadas a covariables.

Mientras en el Modelo Lineal combinamos aditividad de los efectos de las covariables con normalidad de las respuestas y homoscedasticidad, en el GLM estas tres cosas no se satisfacen necesariamente.

Los GLM permiten incluir respuestas no normales, como binomial, Poisson o Gamma, y en la teoría clásica la estimación se relaiza mediante el método de máxima verosimilitud.

Supongamos que observamos las variables de respuesta $Y_1, \ldots, Y_n$ que son v.a. independientes relacionadas con las covariables $x_{i1}, x_{i2}, \ldots, x_{ip}$, $1 \leq i \leq n$.

Genéricamente pensemos en una respuesta Y y covariables$x_1, x_2, \ldots, x_p$

Componentes del modelo

Podemos pensar que el GLM posee tres componentes:

1. Componente Aleatoria: la variable de respuesta Y tiene f.d. o f.p.p dada por

$$f(y, \theta) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\},$$

donde θ es el parámetro canónico, ϕ es un parámetro nuisance y las fun-

ciones $a()$, $b()$, y $c()$ son conocidas.

Denotemos $\ell(\theta, y)$ al log–likelihood, es decir $\ell(\theta, y) = \log(f(y, \theta))$. Recordemos que en una familia exponencial se cumple que, si $\ell'(\theta, y) = \frac{\partial \ell(\theta, y)}{\partial \theta}$, entonces

$$E(\ell'(\theta, y)) = 0 \quad \text{y} \quad E(\ell''(\theta, y)) = -E((\ell'(\theta, y))^2)$$

Por lo tanto, se puede verificar que

$$\mu = E(Y) = b'(\theta) \quad \text{y} \quad Var(Y) = a(\phi)b''(\theta)\,.$$

2. Componente Sistemática: el vector de covariables $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$ que da origen al predictor lineal

$$\eta = \sum_{j=1}^{p} x_j \beta_j = \mathbf{x}'\boldsymbol{\beta}\,.$$

siendo $\boldsymbol{\beta}$ el vector a estimar

3. Función de enlace o link: relaciona las dos componentes μ y η

$$g(\mu) = \eta$$

Nota: En algunos casos $a(\phi)$ es de la forma $a(\phi) = \frac{\phi}{w}$, donde w es un peso conocido.

Los modelos lineales generalizados permiten dos extensiones:

I. podemos tratar distribuciones que pertenezcan a una familia exponencial (simétrica o asimétrica, continua o discreta).

II. podemos elegir una función de enlace que sea una función monótona y diferenciable.

El Modelo Lineal Generalizado tuvo mucha difusión a partir del libro de McCullagh y Nelder (1989). En estos modelos la variable de respuesta Y_i sigue una distribución que pertenece a una familia exponencial con media μ_i que es una función, por lo general no lineal, de $\mathbf{x}_i'\boldsymbol{\beta}$.

Ejemplos

1. Normal: $Y \sim N(\mu, \sigma^2)$.

$$f(y, \theta, \phi) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right)$$
$$= \exp\left(\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{1}{2}\left[\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right]\right),$$

por lo tanto $\theta = \mu$, $b(\theta) = \frac{\mu^2}{2}$, $\phi = \sigma^2$, $a(\phi) = \phi$ y $c(y, \phi) = -\frac{1}{2}\left[\frac{y^2}{\phi} + \log(2\pi\phi)\right]$.

$$E(Y) = \mu$$

En el caso heteroscedástico $Y \sim N(\mu, \frac{\sigma^2}{w})$, donde w es un peso *conocido*, tenemos $\phi = \sigma^2$ y $a(\phi) = \frac{\phi}{w}$.

2. Caso Binomial: $Y \sim Bi(n, p)$

Consideremos $\frac{Y}{n}$ = proporción de éxitos.

$$P(\frac{Y}{n} = y) = P(Y = ny) = \left(\begin{array}{c} n \\ ny \end{array} \right) p^{ny}(1-p)^{n-ny}$$

$$= \exp\left(\frac{y \log\left(\frac{p}{1-p}\right) + \log(1-p)}{1/n} + \log\left(\begin{array}{c} n \\ ny \end{array} \right) \right)$$

por lo tanto $\theta = \log \frac{p}{1-p}$, $b(\theta) = \log(1 + e^{\theta})$, $\phi = n$, $a(\phi) = \frac{1}{\phi}$ y $c(y, \phi) = \left(\begin{array}{c} \phi \\ \phi y \end{array} \right)$.

$$E\left(\frac{Y}{n}\right) = p = \frac{e^{\theta}}{1 + e^{\theta}} = \frac{1}{1 + e^{-\theta}}$$

3. Caso Poisson: $Y \sim P(\lambda)$.

$$\begin{aligned} P(Y = y) &= e^{-\lambda} \frac{\lambda^y}{y!} \\ &= \exp\left(y \log \lambda - \lambda - \log y!\right) \end{aligned}$$

por lo tanto $\theta = \log \lambda$, $b(\theta) = e^{\theta}$, $\phi = 1$, $a(\phi) = 1$ y $c(y, \phi) = -\log y!$

$$E(Y) = \lambda = e^{\theta}$$

Función de enlace o link

Esta función relaciona el predictor lineal η con la esperanza μ de la respuesta Y. A diferencia del modelo lineal clásico, aquí introducimos una función uno–a–uno continua y diferenciable, $g(\mu)$, tal que

$$\eta = g(\mu)\,.$$

Ejemplos de $g(t)$ son la función identidad, el log, la función logística y la probit. Como la función g es biyectiva podremos invertirla, obteniendo:

$$\mu = g^{-1}(\eta) = g^{-1}(\mathbf{x}'\boldsymbol{\beta}) = h(\mathbf{x}'\boldsymbol{\beta})\,.$$

En el caso Binomial, por ejemplo, tenemos que $\mu \in (0,1)$ y el link tiene que mapear sobre la recta real. Suelen usarse 3 links:

1. Logit: $\eta = \log \frac{\mu}{1-\mu}$ $\qquad$ $(\frac{e^\eta}{1+e^\eta})$
2. Probit: $\eta = \Phi^{-1}(\mu)$
3. Complemento log–log: $\eta = \log(-\log(1-\mu))$

Links Canónicos:

En el caso normal mostramos que si $Y \sim N(\mu, \sigma^2)$ el parámetro canónico es $\theta = \mu$.

En el caso binomial $Y \sim Bi(n, p)$ en el que consideremos $\frac{Y}{n}$ vimos que el canónico es $\theta = \text{logit}(\pi)$. Estos son los links más usados en cada caso.

Cuando usamos $\eta = \theta$ el modelo tiene el *link canónico* o *natural*. Es conveniente usar el link natural, ya que algunas cosas se simplifican, pero la posibilidad de usarlo dependerá de los datos con los que estemos trabajando.

- Normal: $\eta\mu$
- Poisson: $\eta = \log \mu$
- Binomial: $\eta = \log \frac{\mu}{1-\mu}$
- Gamma: $\eta = \mu^{-1}$

Función de Verosimilitud para el GLM

Sea Y una v.a. con función de densidad o probabilidad perteneciente a una familia exponencial dada por:

$$f_Y(y, \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\},$$

para algunas funciones conocidas $a(\phi)$, $b(\theta)$ y $c(y, \phi)$. Si ϕ es un parámetro conocido, ésta es una familia exponencial con **parámetro canónico o natural** θ.

Si ϕ no es conocido, ésta puede ser una familia exponencial en (θ, ϕ) o no. ϕ es un parámetro de dispersión o de forma.

La media $E(Y)$ es sólo función de θ y es por lo tanto el parámetro de interés; ϕ en general es tratado como un **parámetro nuisance o de ruido**. En la mayoría de los casos ϕ no será tratado tal como es tratado θ. Estimaremos y haremos inferencia bajo un valor asumido de ϕ y si ϕ necesita ser estimado, lo estimaremos y luego será tomado como un valor fijo y conocido.

Esta familia incluye distribuciones simétricas, asimétricas, discretas y continuas, tales como la distribución Normal, Binomial, Poisson o Gamma.

Momentos de una familia exponencial

Deduciremos el primer y segundo momento de una familia exponencial a partir del logaritmo de su verosimilitud.

$$\ell(\theta, y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\,.$$

Su primera derivada o *score* es:

$$\ell'(\theta, y) = \frac{\partial \ell(\theta, y)}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)}\,,$$

mientras que su derivada segunda es:

$$\ell''(\theta, y) = \frac{\partial \ell^2(\theta, y)}{\partial^2 \theta} = \frac{-b''(\theta)}{a(\phi)}\,.$$

Como $E\left(\dfrac{\partial \ell(\theta, y)}{\partial \theta}\right) = 0$, entonces

$$0 = E\left(\ell'(\theta, y)\right) = E\left[\frac{y - b'(\theta)}{a(\phi)}\right]$$

y por lo tanto

$$\mu = E(Y) = b'(\theta)\,.$$

Además, sabemos que

$$E(\ell''(\theta, y)) = -E\left[(\ell'(\theta, y))^2\right],$$

entonces

$$Var(\ell'(\theta, y)) = E\left[(\ell'(\theta, y))^2\right] = -E(\ell''(\theta, y)) = \frac{b''(\theta)}{a(\phi)}\,.$$

Por otro lado,

$$Var(\ell'(\theta, y)) = Var\left(\frac{y - b'(\theta)}{a(\phi)}\right) = \frac{1}{a^2(\phi)} Var(Y)$$

y en consecuencia

$$Var(Y) = a(\phi)b''(\theta)\,.$$

La varianza es el producto de dos funciones: una que depende del parámetro natural, θ, y otra que depende sólo del parámetro nuisance ϕ. $V(\theta) = b''(\theta)$ es llamada la función de varianza del modelo.

Resumiendo:

$$\begin{aligned} E(Y) &= b'(\theta) \\ Var(Y) &= a(\phi)b''(\theta) \end{aligned}$$

Estimación de los parámetros: Método de Newton–Raphson y Fisher–scoring

Supongamos que $Y_1, \ldots, Y_n$ son variables aleatorias que satisfacen los supuestos de un GLM y que queremos maximizar el loglikelihood $\ell(\boldsymbol{\beta}, \mathbf{y})$ respecto a $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$. Queremos resolver

$$\frac{\partial}{\partial \beta_j} \ell(\boldsymbol{\beta}, \mathbf{y}) = 0 \quad j = 1, \ldots p$$

En general éste es un sistema *no lineal*. Abusando de la notación, a fin de simplificarla, notaremos:

$$\ell'(\boldsymbol{\beta}) = \ell'(\boldsymbol{\beta}, \mathbf{y}) = 0\,.$$

Aproximaremos la ecuación linealmente en la vecindad de un punto $\boldsymbol{\beta}^{(t)}$ mediante el algoritmo de Newton–Raphson.

Método de de Newton–Raphson

Supongamos que queremos resolver

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(x_1, \ldots, x_p) = \begin{pmatrix} f_1(x_1, \ldots, x_p) \\ \vdots \\ f_p(x_1, \ldots, x_p) \end{pmatrix} = 0 \, .$$

Supongamos además que ξ es solución y que $\mathbf{x}_0$ es un punto próximo a ξ. Usando una expansión de Taylor de primer orden alrededor de $\mathbf{x}_0$ tenemos que

$$0 = \mathbf{f}(\xi) \approx \mathbf{f}(\mathbf{x}_0) + \nabla \mathbf{f}(\mathbf{x}_0)(\xi - \mathbf{x}_0)$$

donde

$$\nabla \mathbf{f}(\mathbf{x}_0) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_p} \\ \cdots & \ddots & \cdots \\ \frac{\partial f_p}{\partial x_1} & \cdots & \frac{\partial f_p}{\partial x_p} \end{pmatrix}_{\mathbf{x}=\mathbf{x}_0} \, .$$

Luego,

$$\xi = \mathbf{x}_0 - [\nabla \mathbf{f}(\mathbf{x}_0)]^{-1} \mathbf{f}(\mathbf{x}_0)$$

El método de Newton Raphson es un método iterativo con un punto inicial $\mathbf{x}_0$ y tal que

$$\mathbf{x}_{i+1} = \mathbf{x}_i - [\nabla \mathbf{f}(\mathbf{x}_i)]^{-1} \mathbf{f}(\mathbf{x}_i)$$

Para el caso que nos interesa resolver resultaría

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left[\ell''(\boldsymbol{\beta}^{(t)})\right]^{-1} \ell'(\boldsymbol{\beta}^{(t)}) \tag{3}$$

Si $\ell(\boldsymbol{\beta})$ fuera cuadrática, entonces $\ell'(\boldsymbol{\beta})$ sería lineal y el algoritmo iterativo convergería en un sólo paso a partir de un punto inicial.

En problemas regulares, el log–likelihood se hace aproximadamente cuadráti-

co a medida que n crece. En estas situaciones el método de NR funcionará bien, mientras que en muestras pequeñas y con log–likelihoods alejados de una cuadrática NR podría no converger.

Veamos como quedan los distintos elementos de (3). Por simplicidad estudiaremos la contribución de cada término Y_i al log–likelihood omitiendo los subíndices superfluos. Tenemos que:

$$\begin{aligned} \ell(\theta, y) &= \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \\ \frac{\partial \ell}{\partial \beta_j} &= \frac{\partial \ell}{\partial \theta}\frac{\partial \theta}{\partial \mu}\frac{\partial \mu}{\partial \eta}\frac{\partial \eta}{\partial \beta_j} \end{aligned}$$

Recordemos que

$$\begin{aligned} f(\theta, y) &= \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\} \\ \mu &= E(Y) = b'(\theta) \quad \text{y} \quad Var(Y) = a(\phi)b''(\theta) \\ \eta &= \mathbf{x}'\boldsymbol{\beta} \end{aligned}$$

$$g(\mu) = \eta$$

¿Cuánto vale cada derivada?

$$\frac{\partial \ell}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)} = \frac{y - \mu}{a(\phi)}$$

$$\frac{\partial \theta}{\partial \mu} = \frac{1}{b''(\theta)} = \frac{a(\phi)}{Var(Y)}$$

$$\frac{\partial \mu}{\partial \eta} = \text{depende de la función de enlace}$$

$$\frac{\partial \eta}{\partial \beta_j} = x_j \,,$$

Luego, resulta

$$\frac{\partial \ell}{\partial \beta_j} = \frac{Y - \mu}{Var(Y)} \quad \frac{\partial \mu}{\partial \eta} \quad x_j \,.$$

De esta manera, las ecuaciones de máxima verosimilitud quedan:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{Y_i - \mu_i}{V_i} \quad \frac{\partial \mu_i}{\partial \eta_i} \quad x_{ij} = 0 \tag{4}$$

Por ejemplo, si usamos el link natural tenemos que

$$V = a(\phi) b''(\theta) = a(\phi) b''(\eta)$$

y además

$$\begin{aligned} \mu &= b'(\theta) = b'(\eta) \\ \frac{\partial \mu}{\partial \eta} &= b''(\eta)\,, \end{aligned}$$

por lo tanto el peso queda constante

$$\frac{1}{V} \frac{\partial \mu}{\partial \eta} = \frac{1}{a(\phi)}\,.$$

Si consideramos la derivada segunda a partir de (4) queda:

$$\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_j} = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_k}[Y_i - \mu_i] \frac{1}{V_i} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} + \sum_{i=1}^{n} (Y_i - \mu_i) \frac{\partial}{\partial \beta_k} \left[\frac{1}{V_i} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} \right]. \quad (5)$$

En el método de **Fisher–scoring** se propone utilizar $E\left(\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_j}\right)$ en lugar de $\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_j}$ con el fin de obtener resultados más estables.

Podemos hallar esta esperanza recordando que:

$$\begin{aligned} -E\left(\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_j}\right) &= E\left(\frac{\partial \ell}{\partial \beta_k} \frac{\partial \ell}{\partial \beta_j}\right) \\ &= E\left[\left(\frac{Y - \mu}{Var(Y)}\right)^2 \left(\frac{\partial \mu}{\partial \eta}\right)^2 x_{ij} x_{ik}\right] \\ &= \frac{1}{Var(Y)} \left(\frac{\partial \mu}{\partial \eta}\right)^2 x_{ij} x_{ik} \ . \end{aligned}$$

Si volvemos a la muestra tendremos

$$- \sum_{i=1}^{n} V_i^{-1} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 x_{ij} x_{ik}$$

que en forma matricial podemos escribir como:

$$- \mathbf{X'WX}$$

siendo $\mathbf{W} = diag\left(V_i^{-1} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right)$.

Cuando usamos el link natural queda $V^{-1} \dfrac{\partial \mu}{\partial \eta} = a^{-1}(\phi)$, que es constante por lo tanto, en este caso, Newton–Raphson coincide con Fisher–scoring.
Finalmente, si $\mathbf{V}^{-1} = diag(V_i^{-1})$ y $\frac{\partial \mu}{\partial \eta} = diag(\frac{\partial \mu_i}{\partial \eta_i})$, entonces

$$\frac{\partial \ell}{\partial \beta} = \mathbf{X'V}^{-1} \frac{\partial \mu}{\partial \eta} (\mathbf{Y} - \mu) \ ,$$

y si volvemos a (3) usando Fisher–scoring queda

$$
\begin{aligned}
\boldsymbol{\beta}^{(t+1)} &= \boldsymbol{\beta}^{(t)} + (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{V}^{-1}\frac{\partial \mu}{\partial \eta}(Y-\mu) \\
\boldsymbol{\beta}^{(t+1)} &= (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\left[\mathbf{X}'\mathbf{W}\mathbf{X}\;\boldsymbol{\beta}^{(t)} + \mathbf{X}'\mathbf{V}^{-1}\frac{\partial \mu}{\partial \eta}(Y-\mu)\right] \\
\boldsymbol{\beta}^{(t+1)} &= (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\left[\mathbf{X}'\mathbf{W}\mathbf{X}\;\boldsymbol{\beta}^{(t)} + \mathbf{X}'\mathbf{V}^{-1}\left[\frac{\partial \mu}{\partial \eta}\right]^2\frac{\partial \eta}{\partial \mu}(Y-\mu)\right] \\
\boldsymbol{\beta}^{(t+1)} &= (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{W}\mathbf{z}\,,
\end{aligned}
$$

donde definimos la psuedo–observación $\mathbf{z}$ como

$$\mathbf{z} = \eta + \frac{\partial \eta}{\partial \mu}(Y-\mu)$$

De esta manera vemos al método de Fisher–scoring como mínimos cuadrados pesados iterados (IRWLS) usando pesudo–observaciones $\mathbf{z}$ y los pesos $\mathbf{W}$ que se actualizan en cada paso para actualizar el valor de $\boldsymbol{\beta}$.

Recordemos el algoritmo de cálculo del estimador:

$$\boldsymbol{\beta} = \boldsymbol{\beta}^{(t)} - \left[\ell''(\boldsymbol{\beta}^{(t)})\right]^{-1} \ell'(\boldsymbol{\beta}^{(t)})$$

La contribución de cada término Y_i al loglikelihood es, salvo constantes:

$$\ell_i(\theta_i, Y_i) = \frac{Y_i\theta_i - b(\theta_i)}{a(\phi)} + c(Y_i, \phi)$$

Su derivada respecto de β_j

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{Y_i - \mu_i}{Var(Y_i)} \quad \frac{\partial \mu_i}{\partial \eta_i} \quad x_{ij} \, .$$

Las ecuaciones de máxima verosimilitud quedan:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{Y_i - \mu_i}{V_i} \quad \frac{\partial \mu_i}{\partial \eta_i} \quad x_{ij} = 0\,. \tag{6}$$

La derivada segunda es:

$$\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_j} = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_k}(Y_i - \mu_i)\frac{1}{V_i}\frac{\partial \mu_i}{\partial \eta_i} x_{ij} + \sum_{i=1}^{n}(Y_i - \mu_i)\frac{\partial}{\partial \beta_k}\left[\frac{1}{V_i}\frac{\partial \mu_i}{\partial \eta_i} x_{ij}\right]\,.$$

Método de **Fisher–scoring**: usamos

$$E\left(\frac{\partial^2 \ell_i}{\partial \beta_k \partial \beta_j}\right) = -\frac{1}{Var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 x_{ij} x_{ik}\,.$$

Por lo tanto

$$E\left(\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_j}\right) = -\sum_{i=1}^{n} V_i^{-1} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 x_{ij} x_{ik} \,.$$
$$= -\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \eta_i} V_i^{-1} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} x_{ik} \,.$$

entonces, en forma matricial

$$E\left(\frac{\partial^2 \ell}{\partial \beta \partial \beta}\right) = -\sum_{i} \left(\frac{\partial \mu_i}{\partial \beta}\right)' V_i^{-1} \frac{\partial \mu_i}{\partial \beta}.$$

Finalmente, si:

$$\mathbf{W}^{(t)} = diag\left(V_i^{-1} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right)$$
$$(\mathbf{V}^{(t)})^{-1} = diag(V_i^{-1})$$

resulta

$$\begin{aligned}\boldsymbol{\beta}^{(t+1)} &= \boldsymbol{\beta}^{(t)} + \left(\mathbf{X}'\mathbf{W}^{(t)}\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{V}^{(t)})^{-1}\frac{\partial\mu}{\partial\eta}(Y-\mu)\\ \boldsymbol{\beta}^{(t+1)} &= \left(\mathbf{X}'\mathbf{W}^{(t)}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}^{(t)}\mathbf{z}^{(t)},\end{aligned}$$

donde $\mu = \mu^{(t)}$ y $\eta = \eta^{(t)}$ y

$$\mathbf{z}^{(t)} = \eta + \frac{\partial\eta}{\partial\mu}(Y-\mu)$$

Casos Particulares

Distribución Binomial: regresión logística

Sean $Y_i \sim Bi(1, \pi_i)$. Supongamos que $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i'\boldsymbol{\beta}$, con lo cual

$$\pi_i = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}} = \frac{1}{1 + e^{-\mathbf{x}_i'\boldsymbol{\beta}}}$$

Tenemos las siguientes igualdades:

$$\begin{aligned}
Likelihood &= \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i} \\
Likelihood &= \prod_{i=1}^{n} \left(\frac{\pi_i}{1-\pi_i}\right)^{y_i} (1-\pi_i) \\
Likelihood &= \prod_{i=1}^{n} e^{\mathbf{x}_i'\boldsymbol{\beta} y_i} (1 + e^{\mathbf{x}_i'\boldsymbol{\beta}})^{-1} \\
\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \mathbf{x}_i'\boldsymbol{\beta}\, y_i - \sum_{i=1}^{n} \log(1 + e^{\mathbf{x}_i'\boldsymbol{\beta}})
\end{aligned}$$

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} &= \sum_{i=1}^{n} y_i \mathbf{x}_{ij} - \sum_{i=1}^{n} \frac{e^{\mathbf{x}_i' \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i' \boldsymbol{\beta}}} x_{ij} \\ &= \sum_{i=1}^{n} (y_i - \mu_i)\, x_{ij}\,,\end{aligned}$$

donde $\mu_i = E(Y_i) = \pi_i$.

Derivadas segundas:

$$\begin{aligned}\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} &= -\sum_{i=1}^{n} x_{ij} \frac{\partial}{\partial \beta_k} \left(\frac{e^{\mathbf{x}_i' \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i' \boldsymbol{\beta}}} \right) \\ &= -\sum_{i=1}^{n} \pi_i\, (1 - \pi_i) x_{ij} x_{ik}\end{aligned}$$

Usemos la notación matricial:

$$Likelihood = \prod_{i=1}^{n} \pi_i^{y_i}\, (1 - \pi_i)^{1 - y_i}$$

$$\begin{aligned} \ell'(\boldsymbol{\beta}) &= \mathbf{X}'(\mathbf{Y} - \mu)\,, \\ \ell''(\boldsymbol{\beta}) &= -\mathbf{XWX}\,, \end{aligned}$$

donde

$$\mathbf{W} = diag(\pi_i(1 - \pi_i))\,.$$

Newton–Raphson resulta:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \left(\mathbf{X}'\mathbf{W}^{(t)}\mathbf{X}\right)^{-1}\mathbf{X}'\left(y - \mu^{(t)}\right)\,.$$

Tenemos que $Var(Y_i) = \pi_i(1 - \pi_i)$. La función de varianza resulta:

$$V(\pi) = \pi(1 - \pi)\,.$$

Bajo el modelo logístico

$$\frac{\partial \eta_i}{\partial \pi_i} = \frac{1}{\pi_i(1-\pi_i)},$$

por lo tanto

$$\mathbf{W} = diag\left(\pi_i(1-\pi_i)\right) .$$

y la variable dependiente ajustada es:

$$z_i = \eta_i + \frac{y_i - \pi_i}{\pi_i(1-\pi_i)} = \mathbf{x}_i'\boldsymbol{\beta} + \frac{y_i - \pi_i}{\pi_i(1-\pi_i)} .$$

Intervalos de Confianza y Tests de Hipótesis

Dos de las herramientas más usada de la inferencia estadística son los intervalos de confianza y los tests de hipótesis.

Por ejemplo, los tests de hipótesis son necesarios para comparar el ajuste de dos modelos ajustados a los datos.

Tanto para realizar tests como intervalos de confianza necesitamos las distribuciones muestrales de los estadísticos involucrados.

Distribución Asintótica

Fahrmeir y Kaufmann (1985, Annals of Statistics, 13, 342–368) deducen la consistencia y la distribución asintótica de los estimadores de máxima verosimilitud en el GLM bajo ciertas condiciones de regularidad.

Sea $\mathcal{I}_n = \mathcal{I}_n(\boldsymbol{\beta}_0) = D'V^{-1}D$ donde

$$\begin{aligned} D_{ij} &= \frac{\partial \mu_i}{\partial \beta_j} \\ V &= Diag(V(\mu_i)) \end{aligned}$$

evaluadas en $\boldsymbol{\beta}_0$

Fahrmeir y Kaufmann (1985) probaron que si

- (D) (Diveregencia) $\lambda_{min}(\mathcal{I}_n) \to \infty$
- (C) (Cota inferior) Para todo $\delta > 0$

$$\mathcal{I}_n(\beta) - c\mathcal{I}_n \text{ es semidefinida positiva}$$

para todo $\beta \in N_n(\delta)$ si $n \geq n_1(\delta)$, donde $N_n(\delta)$ es un entorno de β_0 y c es independiente de δ.

- (N) (Convergencia y Continuidad) Para todo $\delta > 0$

$$\max_{\beta \in N_n(\delta)} \|V_n(\beta) - I\| \to 0$$

donde

$$V_n(\boldsymbol{\beta}) = \mathcal{I}_n^{-1/2} \mathcal{I}_n(\boldsymbol{\beta}) \mathcal{I}_n^{-1/2}$$

es una matriz de información normalizada.

Existencia y Consistencia

Entonces, bajo (C) y (D) exite el EMV $\widehat{\boldsymbol{\beta}}$ y además

$$\widehat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}_0$$

Distribución Asintótica

Entonces, bajo (D) y (N)

$$(\mathcal{I}_n)^{1/2} (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N_p(0, \mathbf{I})$$

En la práctica, usaremos un estimador de matriz de covarianza asintótica: $\mathcal{I}_n(\widehat{\boldsymbol{\beta}}_n)$

Esto nos servirá para deducir intervalos de confianza para los parámetros y para deducir tests tipo Wald en tanto

$$(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)' \mathcal{I}_n(\widehat{\boldsymbol{\beta}}_n) (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \overset{(a)}{\sim} \chi^2_p \,.$$

Por lo que ya vimos, entonces para n es suficientemente grande

$$(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \overset{(a)}{\sim} N(\mathbf{O}, (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}) \,.$$

Para n sufcientemente grande, una aproximación razonable esperamos que sea

$$(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \overset{(a)}{\sim} N(\mathbf{O}, \widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}}_n)) \,,$$

siendo

$$\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}}_n) = (\mathbf{X}'\mathbf{W}(\widehat{\beta}_n)\mathbf{X})^{-1} \,.$$

Si queremos computar un intervalo de confianza de nivel asintótico $1-\alpha$ para β_j, éste será:

$$\widehat{\beta}_{nj} \pm z_{\alpha/2}\, \widehat{\sigma}(\widehat{\beta}_{nj})\,,$$

con

$$\widehat{\sigma}(\widehat{\beta}_j) = \left[\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}})_{jj}\right]^{1/2}\,.$$

Inferencia acerca de una función de los coeficientes

Para una función lineal de los prámetros $\Psi = \mathbf{a}'\boldsymbol{\beta}_0$, una aproximación razonable para n suficientemente grande es

$$(\mathbf{a}'\widehat{\boldsymbol{\beta}}_n - \mathbf{a}'\boldsymbol{\beta}_0) \overset{(a)}{\sim} N(\mathbf{O}, \mathbf{a}'\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}}_n)\mathbf{a})\,.$$

Para una función no lineal $\Psi = r(\boldsymbol{\beta}_0)$, para n grande tendremos

$$r(\widehat{\boldsymbol{\beta}}_n) \overset{(a)}{\sim} N(r(\boldsymbol{\beta}_0), \nabla r(\widehat{\boldsymbol{\beta}}_n)'\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}}_n)\nabla r(\widehat{\boldsymbol{\beta}}_n))\,.$$