

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, clasificadas 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
25–29	Alta	Si	212	52	264
		No	50	10	60
	Baja	Si	60	14	74
		No	19	10	29
30–39	Alta	Si	155	54	209
		No	65	27	92
	Baja	Si	112	33	145
		No	77	80	157
40–49	Alta	Si	118	46	164
		No	68	78	146
	Baja	Si	35	6	41
		No	46	48	94
Total		Si	8	8	16
		No	12	31	43
Total			1100	507	1607

En este ejemplo se considera a *Anticoncepción* como variable 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 describir 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 la asociación entre educación y anticoncepción es afectada por el hecho de que mujeres con un nivel de educación más elevado prefieren familias más chicas que las mujeres con niveles de educación inferior.

Compenente Aleatoria

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

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

Tanto la media como la varianza dependen de i , por lo tanto cualquier fac-

tor que afecte la esperanza también afectará la varianza. Esto nos sugiere que cualquier modelo, que como el lineal, asuma homoscedasticidad de las observaciones no será adecuado para este problema.

En nuestro ejemplo, de acuerdo con el valor de las variables predictoras, las observaciones pueden ser clasificadas en 16 grupos. Si llamamos n_i al 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 $k = 0, \dots, n_i$.

Compenente sistemática

El próximo paso en la definición del modelo involucra a las covariables \mathbf{x}_i que participan en lo que llamaremos **componente sistemática**.

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 de hacer esto es mediante los **odds** definidos como

$$\Psi = \frac{\Pi}{1 - \Pi},$$

es decir la razón entre los casos favorables y los no favorables. Veamos unos ejemplos:

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 no alcanza, pues sólo mapea sobre los reales positivos. Para extenderla a los negativos introduciremos el **log**:

$$\text{logit}(\Pi) = \log \left[\frac{\Pi}{1 - \Pi} \right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p = \mathbf{x}'\boldsymbol{\beta} = \eta$$

La función logit es estrictamente creciente y tiene inversa:

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

En nuestro 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 calculan como

$$\frac{\frac{507}{1607}}{\frac{1100}{1607}} = \frac{507}{1100} = 0,461.$$

Entonces, aproximadamente por cada mujer que usa anticonceptivos hay dos que no usan. El $\text{logit}(0,461) = -0,775$.

Modelo de Regresión Logística

Supongamos que Y_1, \dots, Y_n son v.a.independientes tales que

$$Y_i \sim Bi(n_i, \Pi_i). \quad (1)$$

Esto define la *componente aleatoria*.

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

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

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 mantienen constantes todas las demás variables.

Como

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

la relación con π_i es no lineal, luego no es tan sencillo como en el modelo lineal expresar el cambio en Π_i al cambiar un predictor. Sin embargo, cuando el predictor es continuo, podemos hacer una aproximación tomando derivadas con respecto a la j -ésima coordenada de \mathbf{x}_i , obteniendo

$$\frac{\partial \Pi_i}{\partial x_{ij}} = \beta_j \Pi_i (1 - \Pi_i).$$

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

Una vez establecido el modelo que queremos ajustar deberemos estimar los parámetros, hallar intervalos de confianza para los mismos, evaluar la bondad del ajuste y es probable que nos interese realizar algún test que involucre a los parámetros. También tendremos que evaluar la influencia de las observaciones en la determinación de los valores estimados.

Modelo Lineal Generalizado

El modelo lineal clásico lo podemos definir como:

$$\begin{aligned}\mathbf{Y} &= (Y_1, \dots, Y_n)' \sim N(E(\mathbf{Y}), \Sigma_{\mathbf{Y}}) \quad \text{donde} \\ E(\mathbf{Y}) &= \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} \\ \Sigma_{\mathbf{Y}} &= \sigma^2 I\end{aligned} \tag{3}$$

Podemos pensar el modelo (3) como un modelo con tres componentes:

1. Componente Aleatoria: $Y \sim N(\mu, \sigma^2)$
2. Componente Sistemática: covariables x_1, x_2, \dots, x_p que dan origen al predictor lineal $\eta = \sum_{j=1}^p x_j \beta_j$.
3. Función de enlace: enlace entre las dos componentes $\mu = \eta$.

Si escribimos $\eta = g(\mu)$, g es la llamada **función de enlace o link**.

Los modelos lineales generalizados permiten dos extensiones:

- I. podemos tratar distribuciones que pertenezcan a una familia exponencial.

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 $\mathbf{x}'_i\boldsymbol{\beta}$.

Nota

Recordemos que en la expresión clásica del modelo lineal tenemos un error aleatorio aditivo

$$Y = \mathbf{x}'\boldsymbol{\beta} + \epsilon.$$

Los modelos GLM no tienen esta estructura. Por ejemplo, en el caso del logit no podemos escribir

$$\log\left(\frac{\Pi}{1 - \Pi}\right) = \mathbf{x}'\boldsymbol{\beta} + \epsilon.$$

Para este modelo, el error aleatorio ya está incluido en $Y \sim Bi(n, \Pi)$ y $g(\mu) = \eta$ es una relación funcional.

Función de Verosimilitud para el GLM

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

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

para algunas funciones $a(\phi)$, $b(\theta)$ 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. 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:

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

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

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

y por lo tanto

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

Además, sabemos que

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

entonces

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

Por otro lado,

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

y en consecuencia

$$\text{Var}(Y) = a(\phi)b''(\theta).$$

La varianza es el producto de dos funciones: una que depende del parámetro natural, θ y otra que depende sólo del parámetro nuisance ϕ .

Supuestos del modelo

- la variable de respuesta Y tiene distribución

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

donde θ es el parámetro canónico, para el cual

$$\mu = E(Y) = b'(\theta) \quad y \quad Var(Y) = a(\phi)b''(\theta)$$

- el predictor lineal

$$\eta = \mathbf{x}'\boldsymbol{\beta}$$

siendo \mathbf{x} el vector de covariables y $\boldsymbol{\beta}$ el vector a estimar

- la función de enlace que relaciona a η y μ

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

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

Ejemplos

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

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

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

$$E(Y) = \mu$$

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.

$$\begin{aligned} P\left(\frac{Y}{n} = \frac{y}{n}\right) &= \binom{n}{y} p^y (1-p)^{n-y} \\ &= \exp\left(\frac{y}{n} n \log\left(\frac{p}{1-p}\right) + n \log(1-p) + \log\left(\binom{n}{\frac{y}{n}n}\right)\right) \end{aligned}$$

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

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

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

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

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 identidad, el log, la logística y la probit. Como la función g es biyectiva podremos invertirla, obteniendo:

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

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: $\eta = \log \frac{\mu}{1-\mu} \quad \left(\frac{e^\eta}{1+e^\eta} \right)$

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: $\eta = \mu^{-1}$

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

Supongamos que Y_1, \dots, 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 $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$. Queremos resolver

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

En general éste es un sistema *no lineal*.

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

Usando una expansión de Taylor de primer orden, tenemos que:

$$\begin{aligned}\ell'(\boldsymbol{\beta}) &\cong \ell'(\boldsymbol{\beta}^{(t)}) + (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}) \ell''(\boldsymbol{\beta}^{(t)}) \\ \boldsymbol{\beta} &= \boldsymbol{\beta}^{(t)} - [\ell''(\boldsymbol{\beta}^{(t)})]^{-1} \ell'(\boldsymbol{\beta}^{(t)})\end{aligned}\tag{4}$$

Si $\ell(\boldsymbol{\beta})$ es cuadrática, entonces $\ell'(\boldsymbol{\beta})$ es lineal y el algoritmo iterativo con-



vergerá en un solo paso a partir de un punto inicial.

En problemas regulares, el log-likelihood se hace aproximadamente cuadrático 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 (4). Por simplicidad estudiaremos la contribución de cada término Y_i al log-likelihood omitiendo los subíndices superfluos. Salvo constantes tenemos que:

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

Cuánto vale cada derivada?

Recordemos que

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

$$\mu = E(Y) = b'(\theta) \quad y \quad Var(Y) = a(\phi)b''(\theta)$$

$$\eta = \mathbf{x}'\boldsymbol{\beta}$$

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

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

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

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

$$\frac{\partial \eta}{\partial \beta_j} = x_{ij},$$

luego, resulta

$$\frac{\partial \ell}{\partial \beta_j} = \frac{Y - \mu}{Var(Y)} \frac{\partial \mu}{\partial \eta} x_{ij}.$$

De esta manera, las ecuaciones de máxima verosimilitud quedan:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^n \frac{Y_i - \mu_i}{V_i} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} = 0 \quad (5)$$

Por ejemplo, si usamos el link natural tenemos que

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

y además

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

por lo tanto

$$V^{-1} \frac{\partial \mu}{\partial \eta} = 1.$$

Si consideramos la derivada segunda a partir de (5) queda:

$$\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_j} = \sum_i \frac{\partial}{\partial \beta_k} (Y_i - \mu_i) \frac{1}{V_i} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} + \sum_i (Y_i - \mu_i) \frac{\partial}{\partial \beta_k} \left[\frac{1}{V_i} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} \right]. \quad (6)$$

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:

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

Cuando usamos el link natural queda

$$E \left(\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_j} \right) = - \frac{b''(\theta)}{a(\phi)} x_{ij} x_{ik} .$$

Si volvemos a la muestra tendremos

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

que en forma matricial podemos escribir como:

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

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

También notemos que cuando usamos el link natural $V^{-1} \frac{\partial \mu}{\partial \eta} = 1$, por lo tanto (6) queda

$$\sum_i \frac{\partial}{\partial \beta_k} (Y - \mu) \frac{1}{V} \frac{\partial \mu}{\partial \eta} x_j,$$

por lo tanto, en este caso, Newton–Raphson coincide con Fisher scoring.

Finalmente, si $\mathbf{V}^{-1} = \text{diag}(V_i^{-1})$, entonces

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

y si volvemos a (4) queda

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

donde

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

De esta manera vemos al método de Fisher–scoring como mínimos cuadrados pesados iterados (IRWLS)

- 1) En cada ciclo usamos el valor actual de $\boldsymbol{\beta}$ para construir la variable de trabajo \mathbf{z} y nuevos pesos \mathbf{W} .
- 2) Hacemos la regresión de \mathbf{z} sobre \mathbf{x} usando los pesos \mathbf{W} para actualizar el valor de $\boldsymbol{\beta}$.

Recordemos el algoritmo de cálculo del estimador:

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

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

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

Las ecuaciones de máxima verosimilitud quedan:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^n \frac{Y_i - \mu_i}{V_i} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} = 0. \quad (7)$$

La derivada segunda es:

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

Método de **Fisher–scoring**: usamos

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

Por lo tanto

$$\begin{aligned} E \left(\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_j} \right) &= - \sum_i V_i^{-1} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 x_{ij} x_{ik}. \\ &= - \sum_i \frac{\partial \mu_i}{\partial \eta_i} V_i^{-1} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} x_{ik}. \end{aligned}$$

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:

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

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 $Y_i \sim Bi(n_i, \Pi_i)$. 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:

$$Likelihood = \prod_{i=1}^n \frac{n_i!}{y_i! (n_i - y_i)!} \Pi_i^{y_i} (1 - \Pi_i)^{n_i - y_i}$$

$$Likelihood \propto \prod_{i=1}^n \left(\frac{\Pi_i}{1 - \Pi_i} \right)^{y_i} (1 - \Pi_i)^{n_i}$$

$$Likelihood \propto \prod_{i=1}^n e^{\mathbf{x}'_i \boldsymbol{\beta} y_i} (1 + e^{\mathbf{x}'_i \boldsymbol{\beta}})^{-n_i}$$

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^n \mathbf{x}'_i \boldsymbol{\beta} y_i - \sum_{i=1}^n n_i \log(1 + e^{\mathbf{x}'_i \boldsymbol{\beta}})$$

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

donde $\mu_i = E(Y_i) = n_i \Pi_i$.

Derivadas segundas:

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

Usemos la notación matricial:

$$Likelihood = \prod_{i=1}^n \frac{n_i!}{y_i! (n_i - y_i)!} \Pi_i^{y_i} (1 - \Pi_i)^{n_i - y_i}$$

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

donde

$$\mathbf{W} = \text{diag}(n_i \Pi_i (1 - \Pi_i)).$$

Newton–Raphson resulta:

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

Si como antes, pensamos a Y como la proporción de éxitos en los n_i ensayos, tendríamos

$n_i Y_i \sim Bi(n_i, \Pi_i)$. Tenemos que $Var(Y_i) = \frac{\Pi_i(1 - \Pi_i)}{n_i}$. La función de varianza resulta:

$$V(\Pi_i) = \Pi_i(1 - \Pi_i).$$

Bajo el modelo logístico

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

por lo tanto

$$\mathbf{W} = \text{diag} (n_i \Pi_i (1 - \Pi_i)) .$$

Por último 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 usadas 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

Haremos una deducción heurística de la 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 condiciones de regularidad allí establecidas.

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

$$D_{ij} = \frac{\partial \mu_i}{\partial \beta_j}$$

$$V = \text{Diag}(V(\mu_i))$$

evaluadas en $\boldsymbol{\beta}_0$

Fahrmeir y Kaufmann (1985) probaron que si

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

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

para todo $\boldsymbol{\beta} \in N_n(\delta)$ si $n \geq n_1(\delta)$, donde $N_n(\delta)$ es un entorno de $\boldsymbol{\beta}_0$ y c es independiente de δ .

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

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

donde

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

es una matriz de información normalizada.

Existencia y Consistencia

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

$$\widehat{\beta}_n \xrightarrow{p} \beta_0$$

Distribución Asintótica

Entonces, bajo (D) y (N) existe

$$(\mathcal{I}_n)^{-1/2} \mathbf{U} \xrightarrow{\mathcal{D}} N(0, I)$$

y

$$(\mathcal{I}_n)^{1/2} (\widehat{\beta}_n - \beta_0) \xrightarrow{\mathcal{D}} N(0, I)$$

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

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

$$(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)' \mathcal{I}_n(\widehat{\boldsymbol{\beta}}_n) (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \stackrel{(a)}{\approx} \chi_p^2.$$

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

$$(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \stackrel{(a)}{\approx} N(\mathbf{0}, (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}).$$

Para n suficientemente grande, una aproximación razonable esperamos que sea

$$(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \stackrel{(a)}{\approx} N(\mathbf{0}, \widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}}_n)),$$

siendo

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

Si queremos computar un intervalo de confianza de nivel asintótico $1 - \alpha$ para

β_j , éste será:

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

siendo

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

Inferencia acerca de una función de los coeficientes

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

$$(\mathbf{a}'\widehat{\boldsymbol{\beta}}_n - \mathbf{a}'\boldsymbol{\beta}_0) \stackrel{(a)}{\approx} N(\mathbf{0}, \mathbf{a}'\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}}_n)\mathbf{a}).$$

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

$$g(\widehat{\boldsymbol{\beta}}_n) \stackrel{(a)}{\approx} N(g(\boldsymbol{\beta}_0), g^{(1)}(\widehat{\boldsymbol{\beta}}_n)' \widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}}_n) g^{(1)}(\widehat{\boldsymbol{\beta}}_n)),$$

donde hemos notado

$g^{(1)}$ al Jacobiano de g .

Ejemplo

Supongamos un problema de dosis–respuesta en el que un grupo de animales son expuestos a una sustancia peligrosa en distintas concentraciones. Sea n_i el número de animales que recibe la dosis i , Y_i el número de animales que muere y por lo tanto $p_i = Y_i/n_i$ la proporción de muertos en el i –ésimo grupo.

Llamemos Π_i a la probabilidad de muerte y modelemos a Π_i en términos de $z_i = \log_{10}(\text{concentración})$.

Proponemos el modelo:

$$\text{logit}(\Pi_i) = \beta_0 + \beta_1 z_i .$$

Un parámetro de interés en estos problemas suele ser el valor de z para el cual se obtiene el 50 % de muertes. Llamemos a dicho valor M_{50} .

Como $\text{logit}(1/2) = 0$, tenemos que $M_{50} = -\frac{\beta_0}{\beta_1}$. Por lo tanto,

$$\frac{\partial M_{50}}{\partial \beta_0} = \frac{-1}{\beta_1}$$

$$\frac{\partial M_{50}}{\partial \beta_1} = \frac{\beta_0}{\beta_1^2}$$

La varianza estimada de $-\frac{\widehat{\beta}_0}{\widehat{\beta}_1}$ es

$$\begin{bmatrix} -1 \\ \widehat{\beta}_1 \end{bmatrix}, \frac{\widehat{\beta}_0}{\widehat{\beta}_1^2} (\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X})^{-1} \begin{bmatrix} -1 \\ \widehat{\beta}_1 \end{bmatrix},$$

donde $\widehat{\mathbf{W}} = \text{diag}(n_i \widehat{\Pi}_i (1 - \widehat{\Pi}_i))$.

Tests de Hipótesis

En el contexto de GLM abordaremos el problema de comparar dos modelos cuando tienen la misma distribución subyacente y la misma función link.

Consideraremos la comparación de dos modelos anidados, es decir la diferencia entre los dos modelos será que la componente lineal de un modelo tendrá más

parámetros que el otro. El modelo más simple, que corresponderá a H_o , será un caso especial de un modelo más general. Si el modelo más simple ajusta a los datos tan bien como el más general, entonces, en virtud del principio de parsimonia no rechazaremos H_o . Si el modelo más general ajusta significativamente mejor, rechazaremos H_o en favor de H_1 , que corresponde al modelo más general. Para realizar estas comparaciones deberemos usar medidas de *bondad de ajuste*.

Las medidas de bondad de ajuste pueden basarse en en el máximo valor de la función de verosimilitud, en el máximo valor del log de la función de verosimilitud, en el mínimo valor de la suma de cuadrados o en un estadístico combinado basado en los residuos.

El proceso de comparación será como siempre:

1. Especificamos un modelo M_o correspondiente a H_o y un modelo más general, M_1 , que corresponde a H_1 .
2. Ajustamos M_o y calculamos el estadístico de bondad de ajuste G_o . Idem con M_1 y su correspondiente G_1 .
3. Computamos la *mejoría* $G_1 - G_o$ (eventualmente G_1/G_o).

4. A partir de la distribución de $G1 - G_o$ testeamos H_o vs. la alternativa H_1 , es decir M_o vs. M_1 .
5. Si la hipótesis H_o no es rechazada, preferimos el modelo M_o . Si rechazamos H_o elegiremos M_1 .

Estadístico de Cociente de Verosimilitud

El modelo con el máximo número de parámetros que pueden ser estimados se conoce como **modelo saturado**. Es un GLM con la misma distribución subyacente y la misma función de enlace que el modelo de interés, que podría tener tantos parámetros como observaciones. Si hay observaciones que tienen las mismas covariables (replicaciones), el modelo saturado podría determinarse con menos de n parámetros. Llamemos m al máximo número de parámetros que puede especificarse.

En el modelo saturado los μ justan exactamente a los datos. Por lo tanto, en el modelo saturado se asigna toda la variación a la componente sistemática y ninguna a la componente aleatoria. Este modelo no se usa ya que no resume la información presente en los datos, pero provee una base para medir la discrepancia para un modelo intermedio entre el modelo saturado y el **modelo nulo**, en el que hay un único parámetro para todas las observaciones.

Si llamamos $\widehat{\boldsymbol{\theta}}_s$ al al valor estimado bajo el modelo saturado, el $L(\widehat{\boldsymbol{\theta}}_s, \mathbf{y})$, likelihood evaluado en dicho estimador, tomará el valor más grande posible para estas observaciones, asumiendo la misma distribución subyacente y la misma función de enlace.

Sea $L(\widehat{\boldsymbol{\theta}}, \mathbf{y})$ el máximo valor del likelihood para el modelo de interés. El cociente de verosimilitud será

$$\lambda = \frac{L(\widehat{\boldsymbol{\theta}}_s, \mathbf{y})}{L(\widehat{\boldsymbol{\theta}}, \mathbf{y})},$$

que nos da una idea de cuán bueno es el ajuste del modelo.

En la práctica se usa el logaritmo de este cociente

$$\log(\lambda) = \ell(\widehat{\boldsymbol{\theta}}_s, \mathbf{y}) - \ell(\widehat{\boldsymbol{\theta}}, \mathbf{y}).$$

Grandes valores de $\log(\lambda)$ sugieren un pobre ajuste del modelo respecto al modelo saturado.

Un estadístico cercano y muy usado en el contexto de GLM es la **deviance**, introducida por Nelder y Wedderburn (1972).

La **deviance** se define como

$$D = 2 [\ell(\widehat{\boldsymbol{\theta}}_s, \mathbf{y}) - \ell(\widehat{\boldsymbol{\theta}}, \mathbf{y})]$$

$$D = 2 \sum_{i=1}^n a_i^{-1}(\phi) \{y_i(\widehat{\boldsymbol{\theta}}_{si} - \widehat{\boldsymbol{\theta}}_i) - b(\widehat{\boldsymbol{\theta}}_{si}) + b(\widehat{\boldsymbol{\theta}}_i)\}.$$

NOTA: A veces es conveniente expresar el log likelihood en términos de las medias μ 's más que de $\boldsymbol{\beta}$ o $\boldsymbol{\theta}$. En ese caso llamaríamos $\ell(\widehat{\boldsymbol{\mu}}, \mathbf{y})$ al likelihood maximizado sobre $\boldsymbol{\beta}$, mientras que el máximo alcanzado en el modelo saturado sería $\ell(\mathbf{y}, \mathbf{y})$.

Ejemplos

Caso Normal

Recordemos que $\theta = \mu$, $b(\theta) = \frac{\theta^2}{2}$, $\Phi = \sigma^2$ ($w_i = 1$).

Entonces

$$D = 2 \sum_{i=1}^n (y_i(y_i - \mu_i) - \frac{1}{2}y_i^2 + \frac{1}{2}\mu_i^2) = \sum_{i=1}^n (y_i - \mu_i)^2.$$

Caso Binomial

Recordemos que $\theta = \log\left(\frac{\Pi}{1-\Pi}\right)$, es decir $\Pi = \frac{e^\theta}{1+e^\theta}$,

$b(\theta) = -\log(1 - \Pi) = \log(1 + e^\theta)$, entonces

$$\begin{aligned} D &= 2 \sum_{i=1}^n n_i \left\{ \frac{y_i}{n_i} (\widehat{\theta}_{si} - \widehat{\theta}_i) - b(\widehat{\theta}_{si}) + b(\widehat{\theta}_i) \right\} \\ &= 2 \sum_{i=1}^n n_i \left[\frac{y_i}{n_i} \left(\log \frac{y_i/n_i}{1 - y_i/n_i} - \log \frac{\widehat{\Pi}_i}{1 - \widehat{\Pi}_i} \right) + \right. \\ &\quad \left. \log\left(1 - \frac{y_i}{n_i}\right) - \log(1 - \widehat{\Pi}_i) \right] \\ &= 2 \sum_{i=1}^n \left[y_i \log \frac{y_i/n_i}{\widehat{\Pi}_i} + y_i \log \frac{1 - \widehat{\Pi}_i}{1 - y_i/n_i} + \log \frac{1 - y_i/n_i}{1 - \widehat{\Pi}_i} \right] \\ &= 2 \sum_{i=1}^n \left[y_i \log \frac{y_i/n_i}{\widehat{\Pi}_i} + (1 - y_i) \log \frac{1 - y_i/n_i}{1 - \widehat{\Pi}_i} \right] \\ &= 2 \sum_{i=1}^n \left[y_i \log \frac{y_i}{\widehat{\mu}_i} + (n_i - y_i) \log \frac{n_i - y_i}{n_i - \widehat{\mu}_i} \right] \end{aligned}$$

Para realizar los tests de bondad de ajuste debemos conocer la distribución

de D .

Heurísticamente podríamos deducir la la distribución de D . Si hacemos un desarrollo de Taylor de segundo orden tenemos que:

$$\ell(\boldsymbol{\beta}) \simeq \ell(\mathbf{b}) + (\boldsymbol{\beta} - \mathbf{b})' \mathbf{U}(\mathbf{b}) - \frac{1}{2}(\boldsymbol{\beta} - \mathbf{b})' \mathcal{I}(\mathbf{b}) \boldsymbol{\beta} - \mathbf{b}.$$

donde $\mathbf{U} = (U_1, \dots, U_p)'$

$$\begin{aligned} U_j &= \frac{\partial \ell(\boldsymbol{\beta}, \mathbf{y})}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial \ell_i(\boldsymbol{\beta}, y_i)}{\partial \beta_j} \\ &= \sum_{i=1}^n \frac{(Y_i - \mu_i)}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_j} x_{ij} \quad j = 1, \dots, p. \end{aligned}$$

$$E(\mathbf{U}) = 0 \quad E(\mathbf{U}\mathbf{U}') = \mathcal{I},$$

siendo \mathcal{I} la matriz de información de Fisher.

Si \mathbf{b} es el punto donde ℓ alcanza su máximo, entonces

$$\ell(\boldsymbol{\beta}) - \ell(\mathbf{b}) \simeq -\frac{1}{2}(\boldsymbol{\beta} - \mathbf{b})' \mathcal{I}(\mathbf{b}) \boldsymbol{\beta} - \mathbf{b}.$$

Por lo tanto

$$2(\ell(\mathbf{b}) - \ell(\boldsymbol{\beta})) \simeq (\boldsymbol{\beta} - \mathbf{b})' \mathcal{I}(\mathbf{b}) (\boldsymbol{\beta} - \mathbf{b}).$$

y en consecuencia, para n suficientemente grande

$$\ell(\boldsymbol{\beta}) - \ell(\mathbf{b}) \stackrel{(a)}{\simeq} \chi_p^2.$$

de este resultado, obtenemos

$$\begin{aligned} D &= 2 [\ell(\widehat{\boldsymbol{\beta}}_s, \mathbf{y}) - \ell(\widehat{\boldsymbol{\beta}}, \mathbf{y})] \\ &= 2 [\ell(\widehat{\boldsymbol{\beta}}_s, \mathbf{y}) - \ell(\boldsymbol{\beta}_s, \mathbf{y})] \\ &\quad - 2 [\ell(\widehat{\boldsymbol{\beta}}, \mathbf{y}) - \ell(\boldsymbol{\beta}, \mathbf{y})] + 2 [\ell(\boldsymbol{\beta}_s, \mathbf{y}) - \ell(\boldsymbol{\beta}, \mathbf{y})] \end{aligned}$$

Luego,

$$D \stackrel{(a)}{\simeq} \chi_{m-p, \nu}^2,$$

siendo

$$\nu = 2 [\ell(\boldsymbol{\beta}_s, \mathbf{y}) - \ell(\boldsymbol{\beta}, \mathbf{y})] ,$$

donde ν es una constante positiva cercana a 0 si el modelo propuesto ajusta a los datos *tan bien* como el modelo saturado.

En el caso particular de la distribución binomial demostraremos este límite usando las aproximaciones que ya vimos.

Teorema. Bajo las condiciones de regularidad supuestas para los resultados de convergencia de los tests χ^2 y G^2 y suponiendo que $\boldsymbol{\Pi}_0 = \boldsymbol{\Pi}(\boldsymbol{\theta}, 0)$, entonces

$$G^2(M_0) - G^2(M_1) \xrightarrow{\mathcal{D}} \chi_{\nu_1 - \nu_0}^2$$

donde $\nu_1 = \dim(\boldsymbol{\theta}, \boldsymbol{\psi})$ $\nu_0 = \dim(\boldsymbol{\theta}, 0)$

Usaremos en la demostración el resultado dado por el siguiente

Teorema: Sea Y un vector con distribución $N(\boldsymbol{\nu}, \boldsymbol{\Sigma})$. Una condición necesaria y suficiente para que $(Y - \boldsymbol{\nu})' \mathbf{C} (Y - \boldsymbol{\nu})$ tenga distribución χ^2 es que $\boldsymbol{\Sigma} \mathbf{C} \boldsymbol{\Sigma} \mathbf{C} \boldsymbol{\Sigma} = \boldsymbol{\Sigma} \mathbf{C} \boldsymbol{\Sigma}$, donde los grados de libertad serán el rango de $\mathbf{C} \boldsymbol{\Sigma}$ (si $\boldsymbol{\Sigma}$ es no singular la condición se simplifica a $\mathbf{C} \boldsymbol{\Sigma} \mathbf{C} = \mathbf{C}$). (Rao, 1965, p. 150)

Aplicaciones a Test de Hipótesis

Consideremos la hipótesis nula:

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 = (\beta_{01}, \dots, \beta_{0q})'$$

y una hipótesis más general

$$H_1 : \boldsymbol{\beta} = \boldsymbol{\beta}_1 = (\beta_{01}, \dots, \beta_{0p})', \quad \text{con } q < p < n .$$

Si testeamos H_0 vs. H_1 usando la diferencia de los estadísticos de cociente del logaritmo de la verosimilitud tenemos

$$\begin{aligned} D &= D_0 - D_1 \\ &= 2 [\ell(\widehat{\boldsymbol{\beta}}_s, \mathbf{y}) - \ell(\widehat{\boldsymbol{\beta}}_0, \mathbf{y})] - 2 [\ell(\widehat{\boldsymbol{\beta}}_s, \mathbf{y}) - \ell(\widehat{\boldsymbol{\beta}}_1, \mathbf{y})] \\ &= 2 [\ell(\widehat{\boldsymbol{\beta}}_1, \mathbf{y}) - \ell(\widehat{\boldsymbol{\beta}}_0, \mathbf{y})] . \end{aligned}$$

Compararíamos a D con una χ_{p-q}^2 ya que bajo H_0 tendríamos que $D \stackrel{(a)}{\sim} \chi_{p-q}^2$.

Si el valor observado de D fuera mayor que el percentil $\chi_{p-q,\alpha}^2$ rechazaríamos a H_0 en favor de H_1 , bajo el supuesto de que H_1 da una mejor descripción de los datos (aún cuando H_1 no provea un muy buen ajuste).

Ejemplo: los siguientes datos corresponden a un experimento de dosis–respuesta en el que 5 grupos de 6 animales fueron expuestos a una sustancia peligrosa (Schafer, 2000). Y_i denota al número de animales que murieron al ser expuestos a la i –ésima dosis.

obs.	$x_i = \log_{10} \text{concentrac.}$	y_i	$n_i - y_i$	y_i/n_i	$\hat{\Pi}_i$
1	-5	0	6	0.000	0.0080899
2	-4	1	5	0.1667	0.1267669
3	-3	4	2	0.667	0.7209767
4	-2	6	0	1.000	0.9787199
5	-1	6	0	1.000	0.9987799

El comando S–plus que usamos es:

```
salida <- glm(SF ~ logdosis, family=binomial)
```

```
yy<- c(0,1,4,6,6)
sf<- cbind(yy,6-yy)
logdosis<- -c(5:1)
```

```
salida<- glm(sf~logdosis,family=binomial)
```

```
summary(salida)
```

```
Call: glm(formula = sf ~ logdosis, family = binomial)
```

```
Deviance Residuals:
```

	1	2	3	4	5
	-0.3122076	0.282141	-0.291303	0.5080521	0.1210355

```
Coefficients:
```

	Value	Std. Error	t value
(Intercept)	9.586802	3.703679	2.588454
logdosis	2.879164	1.101315	2.614296

```
(Dispersion Parameter for Binomial family taken to be 1 )
```

```
Null Deviance: 28.009 on 4 degrees of freedom
```

Residual Deviance: 0.5347011 on 3 degrees of freedom

Number of Fisher Scoring Iterations: 5

Correlation of Coefficients:

(Intercept)

logdosis 0.9820848

salida\$deviance

[1] 0.5347011

pchisq(salida\$deviance,3)

[1] 0.0887958

1-pchisq(salida\$deviance,3)

[1] 0.9112042

Resumiendo

Call: glm(formula = SF ~ logdosis, family = binomial)

Deviance Residuals:

1	2	3	4	5
-0.3122076	0.282141	-0.291303	0.5080521	0.1210355

Coefficients:

	Value	Std. Error	t value
(Intercept)	9.586802	3.703679	2.588454
logdosis	2.879164	1.101315	2.614296

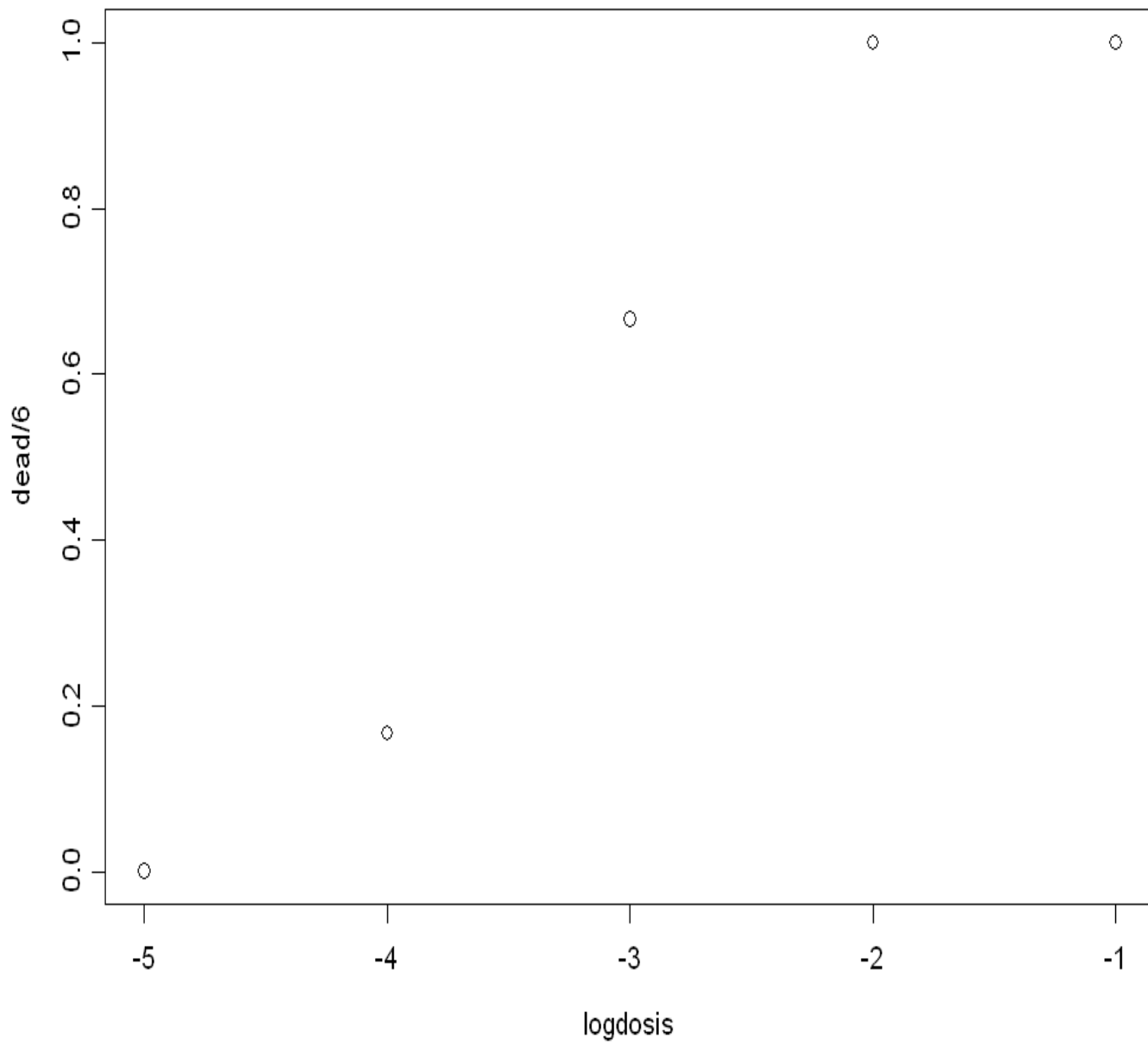
Null Deviance: 28.009 on 4 degrees of freedom

Residual Deviance: 0.5347011 on 3 degrees of freedom

Number of Fisher Scoring Iterations: 5

Correlation of Coefficients: (Intercept)

logdosis 0.9820848



En el caso de la binomial el test G^2 es:

$$G^2 = 2 \sum_{i=1}^N \left\{ y_i \log \left(\frac{y_i}{\hat{\mu}_i} \right) + (n_i - y_i) \log \left(\frac{n_i - y_i}{n_i - \hat{\mu}_i} \right) \right\}$$

Otra medida importante de discrepancia es el estadístico generalizado de *Pearson* χ^2 , que es de la forma

$$\chi^2 = \sum \frac{(y_i - \hat{\mu}_i)^2}{\widehat{V}(Y_i)} = \sum_{i=1}^N \frac{(y_i - n_i \hat{\pi}_i)^2}{n_i \hat{\pi}_i (1 - \hat{\pi}_i)},$$

donde $V(\hat{\mu}_i)$ es la función de varianza estimada para la distribución subyacente.

Para la distribución Normal, χ^2 es la suma de cuadrados habitual.

Tanto la deviance como el estadístico χ^2 tienen distribución exacta χ^2 bajo normalidad y sólo obtenemos resultados asintóticos bajo otras distribuciones.

La ventaja de la deviance como medida de discrepancia es que es aditiva para modelos anidados si se usan estimadores de máxima verosimilitud, mientras que en general χ^2 no es aditiva.