Modelo Lineal Generalizado

Introducción

Comenzaremos con un ejemplo que nos servir 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.

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

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

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

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

Tanto la media como la varianza dependen de i, por lo tanto cualquier factor que afecte la esperanza también afectará la varianza. Esto nos sigiere 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,

$$\begin{aligned} P(Y_i = k) &= \left(\begin{array}{c} n_i \\ k \end{array} \right) \Pi_i^k (1 - \Pi_i)^{n_i - k} \\ E(Y_i) &= n_i \Pi \\ Var(Y_i) &= n_i \Pi_i (1 - \Pi_i) \, , \end{aligned}$$

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

Compenente sistemática

El próximo paso en la definición del modelo involucra a las covariables $\mathbf{x}_i$ 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 1.

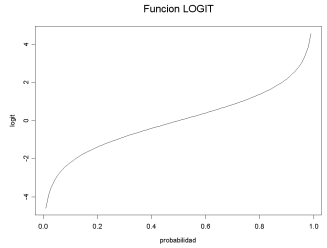
Sin embargo, esta transformación no alcanza, pues sólo mapea sobre los reales positivos. Para extenderla a los negativos introduciremos el **log**:

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

La función logit es estrictamente creciente y tiene inversa:

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

En nuestro ejemplo tenemos: 507 mujeres usan anticonceptivos entre las 1607, por lo que estimamos la probabilidad como $\frac{507}{1607} = 0.316$. Los odds son $\frac{507}{1100} = 0.461$. Entonces, aproximadamente por cada mujer que usa anticoceptivos hay dos que no usan. El $\text{logit}(0.461) = -0.775$.



Modelo de Regresión Logística

Supongamos que $Y_1, \ldots, Y_n$ son 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 probabailidad asociada cuando hay un cambio de una unidad en el j–ésimo predictor y se matienen constantes todas las demás variables.

Como

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

la relación en el miembro derecho es no lineal y por lo tanto no es sencillo 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. Estos temas los iremos desarrollando en el contexto más general de Modelo Lineal Generalizado.

Modelo Lineal Generalizado

El modelo lineal clásico lo podemos definir como:

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

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

1. Componente Aleatoria: $Y \sim N(\mu, \sigma^2)$
2. Componente Sistemática: covariables $x_1, x_2, \ldots, x_p$ que dan origen al predictor lineal $\eta = \Sigma_{j=1}^{p} x_j \beta_j$.
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 $\mathbf{x}_i'\boldsymbol{\beta}$.

Elementos de un GLM

Sea Y la respuesta de una unidad experimental, $\mathbf{x} \in \Re^p$ el vector de covariables asociado de dicha unidad y $\mu = E(Y)$. Definimos el **predictor lineal** como

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

donde $\boldsymbol{\beta} \in \Re^p$ es el parámetro a estimar. Suponemos que

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

para alguna función monótona g, que llamamos función de enlace o link.

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:

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

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,

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

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.

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

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

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

$$E(Y) = \mu$$

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: Sea $Y \sim Bi(n, p)$ y consideremos $\frac{Y}{n}$ = proporción de éxitos.

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

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

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

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

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

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 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: $\eta = \log \frac{\mu}{1-\mu}$ $\qquad (\frac{e^\eta}{1+e^\eta})$
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.

- 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 $Y_1, \ldots, Y_n$ 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

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

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:

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

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

En problemas regulares, el loglikelihood 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 loglikelihoods alejados de una cuadrática NR podría no converger.

Veamos como quedan los distintos elementos de (). Por simplicidad estudiaremos la contribución de cada término Y_i al loglikelihood omitiendo los subíndices superfluos. Salvo constantes:

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

Cuánto vale cada derivada?

$$\begin{aligned} \frac{\partial \ell}{\partial \theta} &= \frac{y - b'(\theta)}{a(\phi)} = \frac{y - \mu}{a(\phi)} \\ \frac{\partial \theta}{\partial \mu} &= \frac{1}{b''(\theta)} = \frac{a(\phi)}{Var(Y)} = \frac{1}{V(\mu)} \\ \frac{\partial \mu}{\partial \eta} &= \text{depende de la función de enlace} \\ \frac{\partial \eta}{\partial \beta_j} &= x_{ij}\,, \end{aligned} \tag{5}$$

luego, resulta

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

De esta manera, las ecuaciones de máxima verosimilitud quedan:

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

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 (8) queda (omitiendo los índices superfluos):

$$\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_j} = \sum_i \frac{\partial}{\partial \beta_k}(Y-\mu)\frac{1}{V}\,\frac{\partial \mu}{\partial \eta}\,x_j + \sum_i (Y-\mu)\frac{\partial}{\partial \beta_k}\left[\frac{1}{V}\,\frac{\partial \mu}{\partial \eta}\,x_j\right]\,. \tag{7}$$

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 $\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_j}$.

Podemos hallar esta esperanza recordando que:

$$-E\left(\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_j}\right) \;=\; E\left(\frac{\partial \ell}{\partial \beta_k}\frac{\partial \ell}{\partial \beta_j}\right)$$

$$= E\left[\left(\frac{Y-\mu}{Var(Y)}\right)^2\left(\frac{\partial\mu}{\partial\eta}\right)^2 x_{ij}x_{ik}\right]$$
$$= \frac{1}{Var(Y)}\left(\frac{\partial\mu}{\partial\eta}\right)^2 x_{ij}x_{ik}\ .$$

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

$$-\sum_i V^{-1}\left(\frac{\partial\mu}{\partial\eta}\right)^2 x_j x_k$$

que en forma matricial podemos escribir como:

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

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

También notemos que cuando usamos el link natural $V^{-1}\dfrac{\partial\mu}{\partial\eta} = 1$, por lo tanto (7) 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 $\mathbf{V} = diag(V_i^{-1})$, entonces

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

y si volvemos a () queda

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

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:

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

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

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

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:

$$\frac{\partial \ell}{\partial \beta_j} \;=\; \sum_{i=1}^{n} \frac{Y_i - \mu_i}{V_i} \quad \frac{\partial \mu_i}{\partial \eta_i} \quad x_{ij} = 0\,. \tag{8}$$

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

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

Por lo tanto

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

Finalmente, si:

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

resulta

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

donde

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

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:

$$\begin{aligned} Likelihood &= \prod_{i=1}^{n} \frac{n_i!}{y_i!\,(n_i-y_i)!}\,\Pi_i^{y_i}\,(1-\Pi_i)^{n_i-y_i} \\ Likelihood &\propto \prod_{i=1}^{n}\left(\frac{\Pi_i}{1-\Pi_i}\right)^{y_i}\,(1-\Pi_i)^{n_i} \\ Likelihood &\propto \prod_{i=1}^{n} e^{\mathbf{x}_i'\boldsymbol{\beta}y_i}\,(1+e^{\mathbf{x}_i'\boldsymbol{\beta}y_i})^{-n_i} \\ \ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\mathbf{x}_i'\boldsymbol{\beta}\;y_i - \sum_{i=1}^{n} n_i\;\log(1+e^{\mathbf{x}_i'\boldsymbol{\beta}}) \\ \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} &= \sum_{i=1}^{n} y_i\mathbf{x}_{ij} - \sum_{i=1}^{n} n_i\,\frac{1}{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}}\,e^{\mathbf{x}_i'\boldsymbol{\beta}}x_{ij} \end{aligned}$$

$$= \sum_{i=1}^{n} (y_i - \mu_i)\, x_{ij}\,,$$

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:

$$\begin{aligned}
\mathit{Likelihood} &= \prod_{i=1}^{n} \frac{n_i!}{y_i!\, (n_i - y_i)!}\, \Pi_i^{y_i}\, (1 - \Pi_i)^{n_i - y_i} \\
\ell'(\boldsymbol{\beta}) &= \mathbf{X}'(y - \mu)\,, \\
\ell''(\boldsymbol{\beta}) &= -\mathbf{X}\mathbf{W}\mathbf{X}\,,
\end{aligned}$$

donde

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

Newton–Raphson resulta:

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

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 $Var(Y_i) = \dfrac{\Pi_i (1 - \Pi_i)}{n_i}$. 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

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

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 usada de la inferencia estadística son los intervalos de confianza y los tests de hipótesis.

Los tests de hipótesis son realizados 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 herurística 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.

Si denotamos ℓ alloglikelihood, en el GLM el score con respecto a β_j es

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

Como ya recordamos, si $\mathbf{U} = (U_1, \ldots, U_p)'$,

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

siendo $\mathcal{I}$ la matriz de información de Fisher. Por el TCL la distribución asintótica de $\mathbf{U}$ es $N_p(\mathbf{O}, \mathcal{I})$

Supongamos que ℓ tiene un único máximo en $\mathbf{b}$ y que $\mathbf{b}$ yace en un entorno del valor verdadero $\boldsymbol{\beta}$. Haciendo un desarrollo de Taylor de primer orden de $\mathbf{U}(\boldsymbol{\beta})$ alrededor de $\mathbf{b}$ obtenemos

$$\mathbf{U}(\boldsymbol{\beta}) \simeq \mathbf{U}(\mathbf{b}) + \mathbf{H}(\mathbf{b})(\boldsymbol{\beta} - \mathbf{b})\,,$$

donde

$$\mathbf{H}(\mathbf{b}) = \frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k}|_{\mathbf{b}}\,.$$

Asintóticamente $\mathbf{H}$ es equivalente a $E(\mathbf{H})$, que está relacionada con la matriz de información

$$\mathcal{I} = E(\mathbf{U}\mathbf{U}') = E(-\mathbf{H})\,.$$

Por lo tanto, si n es suficientemente grande tenemos

$$\mathbf{U}(\boldsymbol{\beta}) \simeq \mathbf{U}(\mathbf{b}) - \mathcal{I}(\boldsymbol{\beta} - \mathbf{b})\,.$$

Como $\mathbf{b}$ es el máximo $\mathbf{U}(\boldsymbol{\beta}) = 0$, si $\mathcal{I}$ es no singular

$$(\mathbf{b} - \boldsymbol{\beta}) \overset{(a)}{\sim} \mathcal{I}^{-1}\mathbf{U}\,.$$

Luego, como $\mathbf{U}$ es asintóticamente normal, viendo a $\mathcal{I}$ como fija, $\mathbf{b} - \boldsymbol{\beta}$ también lo es y por lo tanto para n suficientemente grande

$$(\mathbf{b} - \boldsymbol{\beta}) \overset{(a)}{\sim} N(\mathbf{O}, \mathcal{I}^{-1})$$

y el **estadítico de Wald**

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

Cuando $\mathcal{I}$ depende de $\boldsymbol{\beta}$ en las aplicaciones la estimaremos por $\mathcal{I}(\widehat{\boldsymbol{\beta}})$.

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

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

Para n sufcientemente grande, una aproxímación razonable esperamos que sea

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

siendo

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

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

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

siendo

$$\widehat{\sigma}(\widehat{\beta}_j) = \widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}})_{jj} \,.$$

Inferencia acerca de una función de los coeficientes

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

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

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

$$g(\widehat{\boldsymbol{\beta}}) \overset{(a)}{\sim} N(g(\boldsymbol{\beta}), g^{(1)}(\widehat{\boldsymbol{\beta}})'\widehat{\mathbf{V}}(\widehat{\boldsymbol{\beta}})g^{(1)}(\widehat{\boldsymbol{\beta}})) \,,$$

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

Ejemplo

Supongamos un problema de dosis–respuesta en el que un grupo de animales son espuestos 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 $M_{50} = -\dfrac{\beta_0}{\beta_1}$. Por lo tanto,

$$\begin{aligned} \frac{\partial M_{50}}{\partial \beta_0} &= \frac{-1}{\beta_1} \\ \frac{\partial M_{50}}{\partial \beta_1} &= \frac{\beta_0}{{\beta_1}^2} \end{aligned}$$

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

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

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

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. 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 parsimonía 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* $G1 - G_o$ (eventualmente G_1/G_o).
4. A partir de la distribución de $G1 - G_o$ testeamos $G_1 = G_o$ vs. la alternativa $G_1 \neq G_o$.
5. Si la hipótesis $G_1 = G_o$ no es rechazada, tampoco lo es H_o y preferimos el modelo M_o. Si rechazamos $G_1 = G_o$ elegiremos H_1.

Estadístico de Cociente de Verosimilitud

Una forma de determinar si el modelo es adecuado es compararlo con un modelo más general. 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.

Si n es el tamaño de la muestra, el modelo saturado puede especificarse con n parámetros. 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 parámertos que puede especificarse.

En el modelo saturado los μ's derivados ajustan 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 es usaulmente usado ya que no resume la información presente en los datos, sin embargo 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 para todas las observaciones.

Denotemos $\widehat{\boldsymbol{\beta}}_s$ al estimador de $\boldsymbol{\beta}_s$, el verdadero parámetro bajo el modelo saturado. El $L(\widehat{\boldsymbol{\beta}}_s, \mathbf{y})$, 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 $L(\widehat{\boldsymbol{\beta}}, \mathbf{y})$ el máximo valor del likelihood para el modelo de interés. El cociente de verosimilitud será

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

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

En la práctica se usa el logaritmo de este cociente

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

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

$$D = 2\left[\ell(\widehat{\boldsymbol{\beta}}_s, \mathbf{y}) - \ell(\widehat{\boldsymbol{\beta}}, \mathbf{y})\right]\,.$$

NOTA: Muchas 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 $\ell(\widehat{\mu}, \mathbf{y})$ al likelihood maximizado sobre $\boldsymbol{\beta}$, mientras que el máximo alcanzado en el modelo saturado sería $\ell(\mathbf{y}, \mathbf{y})$. Si denotamos $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\widehat{\mu})$ y $\widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta}(\mathbf{y})$ tendremos

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

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

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

Para realizar los tests de bondad de ajuuste debemos conocer la distribución de D.

Deduciremos en forma heurística la distribución de D. Mediante un desarrollo de Taylor de primer orden y usando la notación anterior tenemos que:

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

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

$$\ell(\boldsymbol{\beta}) - \ell(\mathbf{b}) \overset{(a)}{\sim} \chi^2_p \,.$$

de este resultado , obtenemos

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

Luego,

$$D \overset{(a)}{\sim} \chi^2_{m-p,\nu}\,,$$

siendo

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

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

Aplicaciones a Test de Hipótesis

Consideremos la hipótesis nula:

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

y una hipótesis más general

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

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

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

Si ambos modelos fueran razonables tendríamos que $D_0 \overset{(a)}{\sim} \chi^2_{m-q}$ y $D_1 \overset{(a)}{\sim} \chi^2_{m-p}$, por lo tanto $D \overset{(a)}{\sim} \chi^2_{p-q}$.

Si el valor observado de D fuera mayor que el percentil $\chi^2_{p-q,\alpha}$ 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.	$x_i = \log_{10} concentrac.$	y_i	$n_i - y_i$	y_i/n_i	$\widehat{\Pi}_i$
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)

Al hacer un summary obtenemos:

summary(salida)

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

Deviance Residuals:

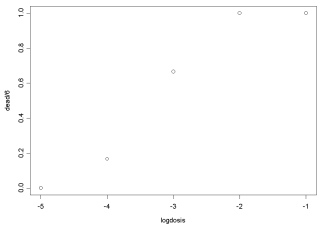
1	2	3	4	5
-0.3122076	0.282141	-0.291303	0.5080521	0.1210355

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

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

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

$$\chi^2 = \sum \frac{(y_i - \widehat{\mu}_i^2)}{V(\widehat{\mu}_i)},$$

donde $V(\widehat{\mu}_i)$ 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.

Análisis de la deviance

El análisis de la deviance es una generalización del análisis de la varianza para los GLM obtenido para una secuencia de modelos anidados (cada uno incluyendo más términos que los anteriores).

Dada una secuencia de modelos anidados usamos la deviance como una medida de discrepancia y podemos formar una tabla de diferencias de deviances.

Sean $M_{p_1}, M_{p_2}, \ldots, M_{p_r}$ una sucesión de modelos anidados de dimensión $p_1 < p_2 < \ldots < p_r$ y matrices de diseño $\mathbf{X}_{p_1}, \mathbf{X}_{p_2}, \ldots, \mathbf{X}_{p_r}$ y deviances $D_{p_1} > D_{p_2} > \ldots > D_{p_r}$. Suponemos en todos ellos la misma distribución y la misma función link. LAS DESIGUALDADES ENTRE LAS DEVIANCES NO SE VERIFICAN NECESARIAMENTE ENTRE LOS ESTADÍSTICOS χ^2 DE PEARSON.

Dos términos A y B son ortogonales si la reducción que A (o B) causa en el desvío de M_{p_i} es independiente de si B (o A) está o no incluido en M_{p_i}. En general ocurre la no–ortogonalidad de los términos de un GLM, lo que cambia la interpretación de la tabla de deviances.

La diferencia $D_{p_i} - D_{p_j}$, $p_j > p_i$, es interpretada como una medida de la variación de los datos explicada por los términos que están en M_{p_j} y no están en M_{p_i}, incluidos los efectos de los términos de que están en M_{p_i} e ignorando los efectos cualquier término que no está en M_{p_j}.

De esta manera, si $D_{p_i} - D_{p_j} > \chi^2_{p_j - p_i, \alpha}$ los efectos de los términos que están en M_{p_j} y no están en M_{p_i} son significativos.

Cada secuencia de modelos corresponde a una tabla de análisis de la varianza diferente. La secuencia de los modelos estará determinada por el interés del investigador.

Residuos

Para el GLM necesitamos extender la noción de residuo a todas las distribuciones que pueden reemplazar a la Normal.

Residuos de Pearson

Los residuos de Pearson se definen como

$$r_i^P = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{Var}(y_i)}},$$

siendo $\widehat{Var}(y_i) = a(\Phi)V(\widehat{\mu}_i)$.

Qué nos queda en el caso Poisson ?

Recordemos que si $Y \sim P(\mu)$, entonces $E(Y) = \mu = Var(Y)$.

$$\frac{y - \widehat{\mu}_i}{\sqrt{\widehat{\mu}_i}}$$

Residuos Deviance

La deviance D puede escribirse como una suma

$$D = \sum_{i=1}^{n} d_i \ ,$$

donde las d_i son los que se conocen como componentes de la *deviance*. Esta medida, tal como vimos suele usarse como una medida de discrepanciaen un GLM y en ese sentido d_i es la contribución de cada dato.

Podemos definir los *residuos deviance* como

$$r_i = sg(y_i - \widehat{\mu}_i)\sqrt{d_i}\,.$$

Para el caso Poisson, recordemos que

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

luego, $\theta = \log\mu$, $b(\theta) = e^{\theta}$, $\phi = 1$, $a(\phi) = 1$ y $c(y, \phi) = -\log y!$

Cada componente de la deviance resulta

$$d_i = sg(y_i - \widehat{\mu}_i)\{2(y_i\log(y_i/\widehat{\mu}_i) - y_i + \widehat{\mu}_i)\}^{1/2}$$

Residuos de Anscombe

Una desventaja de ri^P es que en general su distribución para datos no Normales es asimétrica y por lo tanto no es de esperar que posean propiedades similares a las que poseen bajo normalidad.

Anscombe definió unos residuos baasados en una función $A(y)$ en lugar de y, de manera que la distribución de $A(Y)$ sea tan Normal como sea posible.

Wedderburn demostró que para funciones de verosimilitud en el GLM, la función $A(.)$ es

$$A(.) = \int \frac{d\mu}{V^{1/3}(\mu)} .$$

Por ejemplo, en el caso Poisson queda

$$A(.) = \int \frac{d\mu}{\mu^{1/3}} = \frac{3}{2}\mu^{2/3} .$$

entonces basaría mos los residuos en $y^{2/3} - \mu^{2/3}$.

La transformación que "normaliza" no es la misma necesariamente que la que estabiliza varianzas. Por lo tanto, debemos dividir por el desvío de $A(Y)$. Una aproximación de primer orden a esta varianza es $A'(\mu)\sqrt{V(\mu)}$.

En el caso Poisson, resulta

$$r_i^A = \frac{\frac{3}{2}(y^{2/3} - \mu^{2/3})}{\mu^{1/6}} .$$

Si bien, los residuos de Anscombe y de la deviance parecen muy diferentes, los valores que toman para y y μ dados son muy similares, tal como se muestra en la siguiente tabla.

Caso Binomial

En el caso de la distribución binomial quedaría

$$r_i^P = \frac{y_i - n_i\widehat{\Pi}_i}{\sqrt{n_i\widehat{\Pi}_i(1-\widehat{\Pi}_i)}},$$

$$r_i = 2sg(y_i - \widehat{\Pi}_i)\left[y_i \log\left(\frac{y_i}{n_i\widehat{\Pi}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i\widehat{\Pi}_i}\right)\right]$$