

Tema 6: Modelos Lineales Generalizados

Modelos Lineales

Grados en Estadística y Matemáticas

- 1 Familia exponencial de distribuciones
- 2 Modelo Lineal Generalizado
- 3 Estimación
- 4 Contraste de hipótesis

Familia exponencial de distribuciones

Familia exponencial

Sea Θ un intervalo de \mathbb{R} , y sea $\{f_\theta: \theta \in \Theta\}$ una familia de funciones de densidad (o funciones masa de probabilidad) definidas sobre \mathbb{R}^n . Supondremos que el conjunto $\{\mathbf{y} \in \mathbb{R}^n: f_\theta(\mathbf{y}) > 0\}$ es independiente de θ .

Definición

Diremos que la familia $\{f_\theta: \theta \in \Theta\}$ es una familia exponencial uniparamétrica, si existen funciones reales $Q(\theta)$ y $D(\theta)$ sobre Θ y funciones reales medibles $T(\mathbf{y})$ y $S(\mathbf{y})$ sobre \mathbb{R}^n tales que:

$$f_\theta(\mathbf{y}) = \exp\{Q(\theta)T(\mathbf{y}) + D(\theta) + S(\mathbf{y})\}.$$

Para el caso $n = 1$, si $T(y) = y$ diremos que la familia viene expresada en **forma canónica**. Además, si Y es una v.a. cuya distribución pertenece a la familia exponencial en forma canónica se verifica que:

$$E[Y] = \frac{-D'(\theta)}{Q'(\theta)}, \quad \text{Var}[Y] = \frac{Q''(\theta)D'(\theta) - D''(\theta)Q'(\theta)}{Q'(\theta)^3}$$

Distribución Normal

Si $Y \sim N(\mu, \sigma)$, considerando σ fijado, y tomando como parámetro $\theta = \mu$, su función de densidad se puede escribir como

$$f_{\mu}(y) = \exp\left\{\frac{\mu}{\sigma^2}y - \frac{\mu^2}{2\sigma^2} - \frac{y^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma)\right\}, \quad y \in \mathbb{R}$$

Por tanto, pertenece a la familia exponencial, en forma canónica, con

$$Q(\mu) = \frac{\mu}{\sigma^2}, \quad D(\mu) = -\frac{\mu^2}{2\sigma^2}, \quad S(y) = -\frac{y^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma)$$

Distribución Binomial

Si $Y \sim B(n, p)$, considerando n fijado, y tomando como parámetro $\theta = p$, su función de probabilidad se puede escribir como

$$f_p(y) = \exp\left\{\log\left(\frac{p}{1-p}\right)y + n \log(1-p) + \log\binom{n}{y}\right\}, \quad y = 0, 1, \dots, n$$

Por tanto, pertenece a la familia exponencial, en forma canónica, con

$$Q(p) = \log\left(\frac{p}{1-p}\right), \quad D(p) = n \log(1-p), \quad S(y) = \log\binom{n}{y}$$

Distribución de Poisson

Si $Y \sim \text{Poisson}(\mu)$, tomando como parámetro $\theta = \mu$, su función de densidad se puede escribir como

$$f_{\mu}(y) = \exp\{\log(\mu)y - \mu - \log(y!)\}, \quad y = 0, 1, \dots$$

Por tanto, pertenece a la familia exponencial, en forma canónica, con

$$Q(\mu) = \log(\mu), \quad D(\mu) = -\mu, \quad S(y) = -\log(y!)$$

Modelo Lineal Generalizado

Sea $\mathbf{Y} = (Y_1, \dots, Y_n)^t$ un vector aleatorio n -dimensional, \mathbf{X} una matriz $n \times p$ ($p < n$) de constantes conocidas:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1^t \\ \vdots \\ \mathbf{X}_n^t \end{pmatrix},$$

$g: \mathbb{R} \rightarrow \mathbb{R}$ una función monótona y derivable (llamada función de enlace ó link function) y $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^t$ un vector de parámetros desconocidos.

Definición

Diremos que \mathbf{Y} se ajusta a un **modelo lineal generalizado (MLG)** si

- 1 Y_1, \dots, Y_n son v.a. independientes con funciones de densidad (o funciones masa de probabilidad) pertenecientes a una familia exponencial uniparamétrica expresada en forma canónica, i.e.

$$f_{\theta_i}(y) = \exp\{Q(\theta_i)y + D(\theta_i) + S(y)\}, \quad y \in \mathbb{R}; \quad i = 1, \dots, n.$$

- 2 $g(E[Y_i]) = \mathbf{X}_i^t \boldsymbol{\beta} = x_{i1}\beta_1 + \dots + x_{ip}\beta_p, \quad i = 1, \dots, n.$

Ejemplo 1

Se trata de establecer una ecuación que relacione la edad (*age*) con la presencia o ausencia de una enfermedad coronaria (*chd*). Para ello se toman datos de n sujetos seleccionados para participar en un estudio (Datos del archivo *chdage.dat*).

- Variable respuesta: *chd* cuyos datos son $\mathbf{Y} = (Y_1, \dots, Y_n)^t$. Se asumen Y_1, \dots, Y_n independientes e $Y_i \sim B(1, p_i)$.
- Variable predictora: *age* cuyos datos son x_1, \dots, x_n .
- Para describir p_i en términos de x_i necesitamos una función $g : [0, 1] \rightarrow \mathbb{R}$ que sea monótona y derivable.

Funciones de enlace

- **Modelo de regresión logística.** Caracterizado por la función logit:

$$g(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i \quad i = 1, \dots, n$$

- **Modelo de regresión probit.**

$$g(p_i) = F^{-1}(p_i) = \beta_0 + \beta_1 x_i \quad i = 1, \dots, n$$

siendo $F(\cdot)$ la función de distribución de una $N(0, 1)$. El modelo resultante se denomina

- **Modelo cloglog.** Caracterizado por la función log-log complementaria:

$$g(p_i) = \log(-\log(1-p_i)) = \beta_0 + \beta_1 x_i \quad i = 1, \dots, n$$

Ejemplo 2

El número de muertes por Sida en Australia entre 1983 y 1986 evolucionó de una forma vertiginosa. Se propuso estudiar cómo evoluciona el número de muertes con el tiempo, en este caso registrado en periodos de tres meses. (Datos en el archivo `sida.dat`)

- Variable respuesta: `muertes` cuyos datos son Y_1, \dots, Y_n . Se asumen Y_1, \dots, Y_n independientes y con distribución de Poisson, $Y_i \sim \text{Poisson}(\mu_i)$, ya que esta distribución es considerada adecuada para este tipo de recuentos.
- Variable predictora: `periodo` cuyos datos son x_1, \dots, x_n .
- Para describir μ_i en términos de x_i necesitamos una función $g : [0, 1] \rightarrow \mathbb{R}$ que sea monótona y derivable. Tomando $g(x) = \log(x)$ obtenemos el modelo de regresión de Poisson:

$$\log \lambda_i = \beta_0 + \beta_1 x_i \quad i = 1, \dots, n$$

Otra función de enlace adecuada para esta distribución puede ser $g(x) = x$.

Estimación en MLG

Log verosimilitud

La función de log verosimilitud, expresada en términos de los parámetros originales es

$$\ell(\boldsymbol{\theta}; \mathbf{Y}) = \sum_{i=1}^n [Y_i Q(\theta_i) + D(\theta_i) + S(Y_i)] = \sum_{i=1}^n \ell_i \quad \boldsymbol{\theta} = (\theta_1, \dots, \theta_n)^t$$

siendo $\ell_i = Y_i Q(\theta_i) + D(\theta_i) + S(Y_i)$ la parte de log verosimilitud explicada por la i -ésima observación.

Por las hipótesis del MLG tenemos que

- $\mu_i = E[Y_i] = -D'(\theta_i)/Q'(\theta_i) = \mu_i(\theta_i)$ es función de θ_i , bajo ciertas condiciones invertible.
- $g(\mu_i) = \mathbf{X}_i^t \boldsymbol{\beta}$, y como g es invertible, $\mu_i = \mu_i(\theta_i) = g^{-1}(\mathbf{X}_i^t \boldsymbol{\beta})$, $i = 1, \dots, n$.

En consecuencia podemos asumir que la log-verosimilitud es función de $\boldsymbol{\beta}$.

$$\ell = \ell(\boldsymbol{\beta}; \mathbf{Y})$$

Scores

Se definen los scores del modelo como:

$$U_j = \frac{\partial \ell}{\partial \beta_j}, \quad j = 1, \dots, p.$$

Asumimos que se satisfacen suficientes condiciones de regularidad para asegurar que el máximo global de la función de log-verosimilitud, $\hat{\beta}$, se obtiene de manera única como solución de las ecuaciones

$$U_j(\hat{\beta}) = 0, \quad j = 1, \dots, p$$

Cálculo de los scores

Para cada $i = 1, \dots, n$, utilizando la regla de la cadena, se obtiene:

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{d\ell_i}{d\theta_i} \frac{d\theta_i}{d\mu_i} \frac{\partial \mu_i}{\partial \beta_j}, \quad j = 1, \dots, p.$$

- $\frac{d\ell}{d\theta_i} = Y_i Q'(\theta_i) + D'(\theta_i)$
- $\frac{d\theta_i}{d\mu_i} = \frac{1}{Q'(\theta_i) \text{Var}[Y_i]}$
- Tomando $\mathbf{X}_i^t = (x_{i1}, \dots, x_{ip})$ y $\eta_i = \mathbf{X}_i^t \boldsymbol{\beta}$, se verifica que

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial g^{-1}(\eta_i)}{\partial \beta_j} = (g^{-1})'(\eta_i) \frac{\partial \mathbf{X}_i^t \boldsymbol{\beta}}{\partial \beta_j} = (g^{-1})'(\mathbf{X}_i^t \boldsymbol{\beta}) x_{ij}$$

Cálculo de los scores

$$U_j = U_j(\boldsymbol{\beta}) = \sum_{i=1}^n \frac{\partial \ell_i}{\partial \beta_j} = \sum_{i=1}^n \frac{(Y_i - \mu_i)x_{ij}}{\text{Var}[Y_i]} (g^{-1})'(\mathbf{X}_i^t \boldsymbol{\beta}), \quad j = 1, \dots, p$$

siendo x_{ij} es el j -ésimo elemento de \mathbf{X}_i^t .

Propiedades de los scores

- $E[U_j] = 0$, $j = 1, \dots, p$.
- Si $\mathcal{I} = \mathcal{I}(\boldsymbol{\beta}) = \text{Cov}[\mathbf{U}] = E[\mathbf{U}\mathbf{U}^t]$ es la matriz de covarianzas de \mathbf{U} (denominada *matriz de información de Fisher*), entonces

$$\mathcal{I}_{jk} = \mathcal{I}_{jk}(\boldsymbol{\beta}) = \sum_{i=1}^n \frac{x_{ij}x_{ik}}{\text{Var}[Y_i]} ((g^{-1})'(\mathbf{X}_i^t \boldsymbol{\beta}))^2, \quad j, k = 1, \dots, p$$

Asumiremos que esta matriz es definida positiva, y por tanto no singular.

Cálculo de $\hat{\beta}$. Método de Newton-Raphson

Como ya dijimos, $\hat{\beta}$, se obtiene como solución de las ecuaciones

$$U_j(\hat{\beta}) = 0, \quad j = 1, \dots, p$$

En la mayoría de los casos, no es posible encontrar una solución explícita de estas ecuaciones. Se utilizan métodos numéricos como el método de Newton-Raphson multidimensional.

Si denotamos $\mathbf{b}^{(m)}$ a la aproximación de $\hat{\beta}$ obtenida en la m -ésima iteración, se obtiene la siguiente ecuación:

$$\mathbf{b}^{(m)} = \mathbf{b}^{(m-1)} - \left(\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} \left(\mathbf{b}^{(m-1)} \right) \right)_{j,k=1,\dots,p}^{-1} \mathbf{U}^{(m-1)}$$

donde $\mathbf{U}^{(m-1)} = \mathbf{U}(\mathbf{b}^{(m-1)})$. La convergencia del Método de Newton-Raphson nos garantiza que

$$\mathbf{b}^{(m)} \xrightarrow{m \rightarrow \infty} \hat{\beta}$$

Cálculo de $\hat{\beta}$. Método de Scoring.

Si se reemplaza en la ecuación anterior la matriz Hessiana de ℓ por su esperanza y se tiene en cuenta que

$$E \left[\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} \right] = -E[U_j U_k] = -\mathcal{I}_{jk}$$

se obtiene que:

$$\mathbf{b}^{(m)} = \mathbf{b}^{(m-1)} + \mathcal{I}(\mathbf{b}^{(m-1)})^{-1} \mathbf{U}^{(m-1)}$$

Distribución de los estimadores

- $\mathcal{I}^{-1/2}\mathbf{U} \stackrel{aprox}{\sim} N_p(\mathbf{0}, \mathbf{I}_p)$ (equivalentemente $\mathbf{U} \stackrel{aprox}{\sim} N_p(0, \mathcal{I})$).
- $\hat{\boldsymbol{\beta}} \stackrel{aprox}{\sim} N_p(\boldsymbol{\beta}, \mathcal{I}^{-1})$
- $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^t \mathcal{I} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \stackrel{aprox}{\sim} \chi^2(p)$ (estadístico de Wald)

Intervalos y regiones de confianza

- Un intervalo de confianza aproximado al nivel $1 - \alpha$ ($0 < \alpha < 1$) para el parámetro β_i , viene dado por la expresión:

$$[\hat{\beta}_i - z_{\alpha/2} \sqrt{v_{ii}}; \hat{\beta}_i + z_{\alpha/2} \sqrt{v_{ii}}]$$

- Una región de confianza p -dimensional al nivel $1 - \alpha$ ($0 < \alpha < 1$) para el parámetro $\boldsymbol{\beta}$:

$$\mathcal{R} = \{\mathbf{z} \in \mathbb{R}^p : (\mathbf{z} - \hat{\boldsymbol{\beta}})^t \mathcal{I} (\mathbf{z} - \hat{\boldsymbol{\beta}}) \leq \chi_{p, \alpha}\}$$

Contraste de Hipótesis en MLG

Modelo saturado

Se denomina **modelo saturado** asociado a un MLG al modelo con la misma distribución y función de enlace que dicho MLG, en el cual el número de parámetros es igual al de observaciones.

El estimador de máxima verosimilitud para este modelo se denota $\hat{\beta}_{max}$ y la log-verosimilitud máxima es $\ell(\hat{\beta}_{max}; \mathbf{Y})$.

Deviance

Se define la **deviance** de un MLG com

$$D = 2(\ell(\hat{\beta}_{max}; \mathbf{Y}) - \ell(\hat{\beta}; \mathbf{Y}))$$

Se verifica que

$$D \overset{approx}{\sim} \chi^2(n - p)$$

La utilizamos para los contrastes de hipótesis, ya que nos permite comparar modelos.

Contraste de bondad de ajuste

La deviance compara la verosimilitud de un modelo con la del modelo saturado y es por tanto una medida de la bondad de ajuste de nuestro modelo. A mayor deviance peor ajuste del MLG. Podemos utilizarla para el siguiente contraste de hipótesis:

H_0 : el modelo SÍ es adecuado para describir el conjunto de datos

H_1 : el modelo NO es adecuado para describir el conjunto de datos

Se rechaza H_0 al nivel α si $D \geq \chi_{n-p, \alpha}^2$

Contraste de Hipótesis sobre β .

$$H_0 : \quad \beta_{q+1} = \dots = \beta_p = 0$$

$$H_1 : \quad \beta_i \neq 0 \text{ para algún } i = q + 1, \dots, p$$

- D_1 : deviance asociada al modelo original
- D_0 : deviance asociada al modelo reducido por H_0
- $\Delta D = D_0 - D_1 \stackrel{\text{aprox}}{\sim} \chi^2(p - q)$.

Se rechaza H_0 al nivel α si $\Delta D \geq \chi_{p-q, \alpha}^2$.