

Índice general

1. Modelos lineales generalizados	3
1.1. Introducción	3
1.2. Modelos lineales	4
1.3. Modelos lineales generalizados	6
1.3.1. Familia de distribuciones exponenciales	7
1.3.2. Ajuste de los modelos lineales generalizados	9
1.3.3. Funciones de enlace canónicas	11
1.3.4. Residuos	11
Bibliografía	12

Capítulo 1

Modelos lineales generalizados

1.1. Introducción

Como bien introdujimos antes, los modelos estadísticos pretenden explicar la relación entre dos o mas variables, en particular, tratan de describir el comportamiento de una variable respuesta (o dependiente), que se suele denotar por Y , mediante la información que otorgan las variables predictoras (o independientes), que se suelen denotar como X_1, \dots, X_p .

La forma más general de expresar matemáticamente la intención de los modelos estadísticos es la siguiente:

$$Y = f(X_1, \dots, X_p) + \epsilon$$

Donde f es una función desconocida cuyo propósito es el de representar de la “mejor”¹ manera posible la relación entre las variables X_1, \dots, X_p e Y , y ϵ es un error aleatorio independiente de las variables predictoras que deberá cumplir ciertas condiciones según el tipo de modelo que estemos tratando.

Este trabajo tiene como finalidad el definir un conjunto de estrategias y técnicas que proporcionan un ajuste óptimo de la función f , es decir, que asemeje lo mejor posible la relación de las variables al fenómeno que se esté estudiando. Además, se incluye una posterior aplicación práctica de dichos resultados en el ámbito del cambio climático.

En lo que a este capítulo respecta, partiremos definiendo los modelos lineales de una forma breve y más general, ya que se ve de manera más extensa en varias asignaturas durante el grado. Tras ello daremos varios resultados básicos para el entendimiento y desarrollo de los Modelos Lineales Generalizados.

¹Escribimos mejor entre comillas pues hay diversas formas de evaluar los modelos.

1.2. Modelos lineales

El modelo lineal ocupa un lugar clave en el manual de herramientas de todo estadístico aplicado. Esto se debe a su simple estructura, a la fácil interpretación de sus resultados y al sencillo desarrollo de la teoría de mínimos cuadrados. Sin embargo, a la hora de dar su definición se deben tener en cuenta ciertas restricciones que deben cumplir las variables y los errores del modelo. Estas condiciones hacen que el modelo no sea capaz de adaptarse bien a todos los fenómenos que uno se propone describir pero, a cambio, otorga esa sencillez de visualización antes mencionada.

Definición 1.2.1 (Modelo Lineal). Sean X_1, \dots, X_p un conjunto de p vectores aleatorios de n componentes (con $p \leq n$) e $Y = (Y_1, \dots, Y_n)^T$ un vector aleatorio de n componentes tal que $E[Y] = \mu$. Entonces, se entiende por modelo lineal (multivariante) aquel que determina la relación entre los vectores aleatorios mediante una combinación lineal de parámetros de la siguiente forma:

$$\begin{aligned}\mu &= X\beta \\ Y &= \mu + \epsilon\end{aligned}$$

O vectorialmente como:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_{(1,1)} & \cdots & x_{(1,j)} & \cdots & x_{(1,p)} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{(i,1)} & \cdots & x_{(i,j)} & \cdots & x_{(i,p)} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{(n,1)} & \cdots & x_{(n,j)} & \cdots & x_{(n,p)} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_i \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Donde:

- β es un vector de β_i parámetros a determinar que reflejan la magnitud del efecto lineal (constante) de los incrementos unitarios en las variables explicativas X_i sobre la variable explicada Y .
- ϵ es un vector aleatorio tal que los ϵ_i son variables aleatorias independientes e idénticamente distribuidas por una distribución normal de esperanza nula y varianza σ^2 ($\epsilon_i \sim N(0, \sigma^2)$). Representa el término de error del modelo y corresponde con $\epsilon = Y - E[Y]$.

Observación 1.2.1. Gracias a lo notado en el párrafo anterior podemos ver que: $Y \sim N(\mu, \sigma^2 I_n)$ ya que:

$$\begin{aligned}E[Y] &= E[X\beta + \epsilon] = X\beta + E[\epsilon] = X\beta = \mu \\ Cov(Y) &= Cov(X\beta + \epsilon) = Cov(\epsilon) = \sigma^2 I_n\end{aligned}$$

Veamos ahora cómo se pueden obtener los valores de los parámetros β .

Definición 1.2.2 (Estimador por mínimos cuadrados). Elegiremos los valores de β que minimicen la suma de cuadrados:

$$S = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (Y_i - \mu_i)^2 = \sum_{i=1}^n (Y_i - (X\beta)_i)^2 \quad (1.1)$$

Es decir:

$$\hat{\beta} = \arg_{\beta \in \mathbb{R}^p} \min ||Y - X\beta||^2$$

Donde $|| \cdot ||$ denota la norma euclídea.

Para encontrar tal mínimo se razona derivando respecto de cada β_i y luego igualando a 0. De este modo los parámetros β_i vienen dados por la solución del siguiente sistema:

$$\begin{cases} \frac{\partial S}{\partial \beta_1} = 0 \\ \vdots \\ \frac{\partial S}{\partial \beta_p} = 0 \end{cases}$$

Desarrollando este sistema de ecuaciones llegamos a que es equivalente a $X^T X \hat{\beta} = X^T Y$, por lo que el estimador por mínimos cuadrados del vector de parámetros β viene dado por:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (1.2)$$

Observación 1.2.2. Notemos que esta expresión tiene sentido pues el producto $X^T X$ resulta en una matriz cuadrada de orden p y rango máximo.

Proposición 1.2.1 (Distribución del estimador por mínimos cuadrados). *El estimador por mínimos cuadrados del vector de parámetros β , $\hat{\beta}$, sigue una distribución del tipo normal p -variante de esperanza β y matriz de covarianzas $V_{\hat{\beta}} = \sigma^2 (X^T X)^{-1}$. Es decir: $\hat{\beta} \sim N_p(\beta, \sigma^2 (X^T X)^{-1})$*

Demostración. Partiremos viendo que el estimador $\hat{\beta}$ es insesgado:

$$E[\hat{\beta}] = E[(X^T X)^{-1} X^T Y] = (X^T X)^{-1} X^T E[Y] = (X^T X)^{-1} X^T X \beta = \beta$$

Por otro lado, calculemos la matriz de covarianzas de $\hat{\beta}$, para ello primero debemos notar que como los errores aleatorios ϵ_i son independientes e idénticamente distribuidos con esperanza nula y varianza σ^2 , $\forall i \neq j$:

$$E[\epsilon_i \epsilon_j] = E[\epsilon_i] + E[\epsilon_j] + Cov(\epsilon_i, \epsilon_j) = 0$$

y, por tanto:

$$E[\epsilon \epsilon^T] = E \begin{bmatrix} \epsilon_1^2 & \epsilon_1 \epsilon_2 & \cdots & \epsilon_1 \epsilon_n \\ \epsilon_1 \epsilon_2 & \epsilon_2^2 & \cdots & \epsilon_2 \epsilon_n \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_1 \epsilon_n & \cdots & \cdots & \epsilon_n^2 \end{bmatrix} = \begin{pmatrix} E[\epsilon_1^2] & E[\epsilon_1 \epsilon_2] & \cdots & E[\epsilon_1 \epsilon_n] \\ E[\epsilon_1 \epsilon_2] & E[\epsilon_2^2] & \cdots & E[\epsilon_2 \epsilon_n] \\ \vdots & \vdots & \ddots & \vdots \\ E[\epsilon_1 \epsilon_n] & \cdots & \cdots & E[\epsilon_n^2] \end{pmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \sigma^2 \end{bmatrix}$$

Luego, utilizando que $\hat{\beta} = \beta + (X^T X)^{-1} X^T \epsilon$, se tiene que:

$$\begin{aligned} \text{Cov}(\hat{\beta}) &= E[(\hat{\beta} - E[\hat{\beta}])(\hat{\beta} - E[\hat{\beta}])^T] = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] = \\ &= E[((X^T X)^{-1} X^T \epsilon)((X^T X)^{-1} X^T \epsilon)^T] = E[(X^T X)^{-1} X^T \epsilon \epsilon^T X (X^T X)^{-1}] = \\ &= (X^T X)^{-1} X^T E[\epsilon \epsilon^T] X (X^T X)^{-1} = (X^T X)^{-1} X^T \sigma^2 I_n X (X^T X)^{-1} = \\ &= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1} \end{aligned}$$

Acabamos la prueba recordando que $\hat{\beta}$ no es más que una combinación lineal de variables aleatorias normales, concretamente de las Y_i , para decir que en efecto sigue una distribución normal p-variante como indica el enunciado. ■

Observación 1.2.3. Ahora bien, generalmente el valor de σ^2 es desconocido así que también sería preciso dar una estimación del mismo para que así los resultados anteriores fueran de alguna utilidad.

Definición 1.2.3 (Estimador de σ^2). La varianza σ^2 admite un estimador insesgado que se basa en la suma de cuadrados:

$$\hat{\sigma}^2 = \frac{S}{n-p} = \frac{\sum_{i=1}^n (Y_i - (X\beta)_i)^2}{n-p}$$

Además se tiene que $\hat{\sigma}^2$ sigue una distribución:

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{n-p} \chi_{n-p}^2$$

Observación 1.2.4. No daremos la obtención de este estimador pero sí indicaremos que su distribución se obtiene directamente de que como S es la suma de normales $N(0, \sigma^2)$ y $\frac{\epsilon_i}{\sigma} \sim N(0, 1)$, entonces: $\frac{1}{\sigma^2} S \sim \chi_{n-p}^2$. Finalmente, multiplicando y dividiendo esta expresión por $(n-p)$ y sustituyendo la definición del estimador $\hat{\sigma}^2$ obtenemos el resultado.

1.3. Modelos lineales generalizados

Nos adentramos ya con esta sección en la primera extensión de los modelos lineales de las que se tratan durante el trabajo. Los Modelos Lineales Generalizados (MLG) fueron originalmente formulados por John Nelder y Robert Wedderburn (1972), quienes tenían como propósito unificar varios modelos estadísticos como la regresión lineal, la logística y la de Poisson en un mismo modelo. Este tipo de modelos relajan algunas de las condiciones de los modelos lineales, como que los errores ya no deben seguir ninguna distribución específica, y además añade nuevos elementos como la función de enlace, que interviene en la relación entre las medias y la forma lineal del modelo. También se deberá tener en cuenta que seguirá la hipótesis de que las variables de respuestas siguen distribuciones de tipo exponencial. Más tarde nos adentraremos en cada uno de estos aspectos, de momento demos la estructura básica de un MLG.

Definición 1.3.1 (Estructura básica de un MLG).

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = \begin{pmatrix} E[Y_1] \\ \vdots \\ E[Y_n] \end{pmatrix} = E[Y]$$

$$g(\mu_i) = X_i\beta, \forall i = 1, \dots, n$$

Donde:

- X es la matriz modelo de dimensión $n \times p$ con $p \leq n$, que contiene a las variables predictoras, cada columna representa una variable predictora X_i .
- $\beta = (\beta_1, \dots, \beta_p)^T$ es el vector de parámetros desconocidos como en el caso de los modelos lineales. A $\eta = X\beta$ se le conoce como predictor lineal.
- $g : \mathbb{R} \rightarrow \mathbb{R}$ es la función de enlace, que debe ser diferenciable y monótona. Representa la relación entre la media de las variables de respuesta y el predictor lineal.
- $Y = (Y_1, \dots, Y_n)^T$ es un vector aleatorio, se suele suponer que las Y_i son variables aleatorias independientes y que siguen una distribución de tipo exponencial.

Observación 1.3.1. Desde esta formulación podemos ver fácilmente el por qué decimos que los MLG son una generalización de los modelos lineales, ya que basta tomar a la identidad como la función de enlace y suponer que la distribución considerada sea de tipo normal para encontrarnos ante la forma general de un modelo lineal como vimos en la sección anterior.

1.3.1. Familia de distribuciones exponenciales

Como hemos mencionado antes, la variable de respuesta de los modelos lineales generalizados deben seguir una distribución de tipo exponencial, en esta sección veremos qué significa eso y qué implicaciones tiene. Uno de los motivos más importantes por los que se supone que las variables de respuesta Y_i siguen distribuciones de esta familia se debe a que en los modelos lineales los cambios constantes en las variables predictoras implicaban cambios constantes en la variable de respuesta, pero ahora se quiere permitir que dichos cambios constantes de entrada puedan implicar también variaciones geométricas.

Definición 1.3.2 (Distribución de tipo exponencial). Una distribución se dice que es de tipo exponencial si su función de densidad es de la forma:

$$f_\theta(y) = e^{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)}$$

Donde:

- a , b y c son funciones arbitrarias.
- ϕ es conocido como parámetro de escala.

- θ es conocido como parámetro canónico de la distribución. Más adelante veremos que depende completamente de los parámetros del modelo β .

Ejemplo 1.3.1. La distribución normal es de tipo exponencial pues su función de densidad es:

$$f_{\mu}(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} = e^{\frac{-y^2+2y\mu-\mu^2}{2\sigma^2} - \log(\sigma\sqrt{2\pi})} = e^{\frac{y\mu-\mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \log(\sigma\sqrt{2\pi})}$$

Y por tanto toma los siguientes parámetros de la familia exponencial:

- $\theta = \mu$
- $b(\theta) = \frac{\theta^2}{2} = \frac{\mu^2}{2}$
- $a(\phi) = \phi = \sigma^2$
- $c(\phi, y) = \frac{-y^2}{2\phi} - \log(\sqrt{\phi 2\pi}) = \frac{-y^2}{2\sigma} - \log(\sigma\sqrt{2\pi})$

Es posible dar una forma general para la esperanza y la varianza de las variables de tipo exponencial dependiendo de los parámetros de su función de densidad. Lo vemos en el siguiente resultado.

Proposición 1.3.1. *Sea Y una variable de tipo exponencial, entonces verifica:*

$$E[Y] = b'(\theta) \tag{1.3}$$

$$Var(Y) = b''(\theta)a(\phi) \tag{1.4}$$

Demostración. Partimos considerando la función de verosimilitud logarítmica para θ :

$$l(\theta) = \log(f_{\theta}(y)) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)$$

Y su derivada respecto de theta:

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

Ahora bien, si cambiamos la observación y por la variable Y , podemos evaluar la esperanza de esta derivada, la cual será 0 por propiedades de la función de verosimilitud logarítmica.

$$E\left[\frac{\partial l}{\partial \theta}(\theta)\right] = \frac{E[Y] - b'(\theta)}{a(\phi)} = 0$$

Y de aquí se obtiene directamente que $E[Y] = b'(\theta)$. Seguimos derivando para obtener que:

$$\frac{\partial^2 l}{\partial \theta^2}(\theta) = -\frac{b''(\theta)}{a(\phi)}$$

Y utilizando que: $E\left[\frac{\partial^2 l}{\partial \theta^2}\right] = -E\left[\left(\frac{\partial l}{\partial \theta}\right)^2\right]$, nos queda que:

$$\frac{b''(\theta)}{a(\phi)} = \frac{E[(Y - b'(\theta))^2]}{a(\phi)^2} = \frac{E[(Y - E[Y])^2]}{a(\phi)^2} = \frac{Var(Y)}{a(\phi)^2}$$

De donde se obtiene que: $Var(Y) = b''(\theta)a(\phi)$. ■

Observación 1.3.2. Cuando ϕ es conocido el manejo de la función a no tiene dificultad, pero en muchos casos ϕ suele ser desconocido, así que para agilizar los resultados escribiremos $a(\phi) = \frac{\phi}{\omega}$, donde ω es una constante conocida. De hecho, todos los casos prácticos de interés se podrán expresar así y la mayoría con $\omega = 1$. De este modo nos queda: $\text{Var}(Y) = b''(\theta) \frac{\phi}{\omega}$. Por otro lado, en secciones posteriores necesitaremos trabajar con $\text{Var}(Y)$ en función de $\mu = E[Y]$, para ello utilizaremos la relación 1.3 y definiremos una nueva función:

$$V(\mu) = \frac{b''(\theta)}{\omega} \quad (1.5)$$

pues de este modo se tiene que: $\text{Var}(Y) = V(\mu)\phi$

Podemos recoger las características de las principales distribuciones de tipo exponencial en la siguiente tabla:

	Binomial $Bi(n, p)$	Normal $N(\mu, \sigma^2)$	Poisson $Po(\lambda)$	Gamma $Ga(p, \lambda)$
$\theta(\mu)$	$\log(\frac{\mu}{n-\mu})$	μ	$\log(\mu)$	$-\frac{1}{\mu}$
ϕ	1	σ^2	1	$\frac{1}{\mu}$
$a(\phi)$	1	σ^2	1	$\frac{p}{\mu}$
$b(\theta)$	$n \log(1 + e^\theta)$	$\frac{\theta^2}{2}$	e^θ	$-\log(-\theta)$
$c(y, \phi)$	$\log(\binom{n}{y})$	$-\frac{1}{2}(\frac{y^2}{\phi} + \log(2\pi\phi))$	$-\log(y!)$	$p \log(py) - \log(y\Gamma(p))$

1.3.2. Ajuste de los modelos lineales generalizados

La estimación de parámetros e inferencia con modelos aditivos generalizados se basa en la estimación de máxima verosimilitud, ya que, gracias a la hipótesis de que las Y_i pertenezcan a la familia de distribuciones exponenciales, siempre se dispondrá de funciones de densidad. Partiremos considerando un MLG como el de la definición 1.3.1 y nuestro principal objetivo en esta sección será dar el estimador de máxima verosimilitud del vector de parámetros β . Veremos que será necesario recurrir a un algoritmo basado en mínimos cuadrados para hallar tal máximo, el método de mínimos cuadrados ponderados iterativamente.

Empezamos considerando y una observación de Y y notando que, como los Y_i son independientes entre sí, la función de verosimilitud para β viene dada por:

$$L(\beta) = \prod_{i=1}^n f_{\theta_i}(y_i)$$

Y la función de verosimilitud logarítmica:

$$l(\beta) = \sum_{i=1}^n \log(f_{\theta_i}(y_i)) = \sum_{i=1}^n \frac{y_i \theta_i - b_i(\theta_i)}{a_i(\phi)} + c_i(\phi, y_i)$$

La dependencia de β viene de que θ depende de μ y este a su vez depende del predictor lineal. Ahora bien, si sustituimos en esta expresión que $a_i(\phi) = \frac{\phi}{\omega_i}$, como mencionamos

en la sección anterior, nos queda:

$$l(\beta) = \sum_{i=1}^n \omega_i \frac{y_i \theta_i - b_i(\theta_i)}{\phi} + c_i(\phi, y_i)$$

Como queremos hallar el β que la hace máxima derivamos respecto β_i e igualamos a 0:

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^n \omega_i (y_i \frac{\partial \theta_i}{\partial \beta_j} - b'_i(\theta_i) \frac{\partial \theta_i}{\partial \beta_j})$$

Donde $\frac{\partial \theta_i}{\partial \beta_j}$ por la regla de la cadena es:

$$\frac{\partial \theta_i}{\partial \beta_j} = \frac{d\theta_i}{d\mu_i} \frac{\partial \mu_i}{\partial \beta_j} = \frac{d\theta_i}{d\mu_i} \frac{\partial \eta_i}{\partial \beta_j} \frac{\partial \mu_i}{\partial \eta_i} = \frac{X_{ij}}{g'(\mu_i) b''(\theta_i)}$$

Hemos utilizado que $\frac{\partial \mu_i}{\partial \eta_i} = g'(\mu_i)$, que $\frac{\partial \eta_i}{\partial \beta_j} = X_{ij}$ y que derivando el resultado 1.3 se tiene: $\frac{d\theta_i}{d\mu_i} = \frac{1}{b''(\theta)}$. Sustituyendo también el resultado de 1.5:

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^n \frac{y_i - b'_i(\theta_i)}{g'(\mu_i) b''_i(\theta_i) / \omega_i} = \frac{1}{\phi} \sum_{i=1}^n \frac{y_i - \mu_i}{g'(\mu_i) V(\mu_i)} X_{ij} \quad (1.6)$$

Observación 1.3.3. Ahora bien, el razonamiento general que seguiríamos sería el igualar estas derivadas a 0 y resolver el sistema resultante, sin embargo, se trata de un sistema con ecuaciones no lineales, por lo que para resolverlo utilizaremos métodos numéricos, en concreto utilizaremos el método de Newton que precisa del gradiente y del Hessiano de l . Por lo que debemos volver a derivar l .

$$\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} = -\frac{1}{\phi} \sum_{i=1}^n \frac{X_{ij} X_{ik} \alpha(\mu_i)}{g'(\mu_i)^2 V(\mu_i)}$$

Donde $\alpha(\mu_i) = 1 + (y_i - \mu_i) \left(\frac{V'(\mu_i)}{V(\mu_i)} + \frac{g''(\mu_i)}{g'(\mu_i)} \right)$. Luego, si definimos la matriz $W = \text{diag}(\omega_i)$ para $\omega_i = \frac{\alpha(\mu_i)}{g'(\mu_i)^2 V(\mu_i)}$, el Hessiano de l es: $-\frac{1}{\phi} X W X$. De manera similar, si definimos $G = \text{diag}(\frac{g'(\mu)}{\alpha(\mu_i)})$, el gradiente de l puede escribirse como: $X^T W G \frac{(y - \mu)}{\phi}$.

De este modo, una actualización del método de Newton es de la forma:

$$\beta^{k+1} = \beta^k + (X W X)^{-1} X^T W G (y - \mu) = (X W X)^{-1} X^T W z$$

Donde $z_i = g'(\mu_i) \frac{y_i - \mu_i}{\alpha(\mu_i)} + \eta_i$

Observación 1.3.4. De lo anterior se podemos notar que las actualizaciones de los β^k no son más que las estimaciones de β por mínimos cuadrados ponderados, es decir, resultan de minimizar:

$$\sum_{i=1}^n \omega_i (z_i - X_i \beta)^2 \quad (1.7)$$

Podemos entonces obtener el estimador numérico de los parámetros β de los MLG mediante el siguiente algoritmo.

Definición 1.3.3 (Algoritmo de mínimos cuadrados ponderados iterativamente). 1) Inicialización: tomar $\hat{\mu}_i = y_i + \delta_i$ y $\hat{\eta}_i = g(\hat{\mu}_i)$, donde δ_i suele ser 0 o una constante que asegure que $\hat{\eta}_i$ sea finito.

2) Calcular: $z_i = g'(\mu_i) \frac{y_i - \hat{\mu}_i}{\alpha(\mu_i)} + \hat{\eta}_i$ y $\omega_i = \frac{\alpha(\hat{\mu}_i)}{g'(\hat{\mu}_i)^2 V(\hat{\mu}_i)}$

3) Encontrar $\hat{\beta}$ el parámetro que minimiza la función objetivo de mínimos cuadrados ponderados 1.7 y actualizar $\hat{\eta} = X\hat{\beta}$ y $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$

Observación 1.3.5. Para saber con qué iteración quedarnos debemos comprobar paso a paso si la variación entre el valor antiguo y el nuevo de $\hat{\beta}$ o las derivadas de la función de verosimilitud logarítmica están lo suficientemente cerca de 0.

1.3.3. Funciones de enlace canónicas

Definición 1.3.4 (Funciones de enlace canónicas). Se dice que una función de enlace g_c es canónica para una distribución de la familia exponencial si verifica: $g_c(\mu_i) = \theta_i$, es decir, relaciona directamente el parámetro canónico con el predictor lineal.

Proposición 1.3.2 (Propiedades de las funciones de enlace canónicas). ■ *Su uso en el modelo 1.3.1 resulta en: $\theta_i = X_i\beta$*

- *Hacen que $\alpha(\mu_i) = 1$.*
- *El Hessiano de la función de verosimilitud logarítmica coincide con su valor esperado.*
- *El sistema a resolver para obtener los estimadores de máxima verosimilitud de β , es decir, el formado por las derivadas parciales $\frac{\partial l}{\partial \beta_j}$ igualadas a 0 se reduce a: $X^T y - X^T \hat{\mu} = 0 \Rightarrow X^T y = X^T \hat{\mu}$ pues $\frac{\partial \theta_i}{\partial \beta_j} = X_{ij}$*

1.3.4. Residuos

Al igual que para los modelos lineales, el estudio de los residuos ϵ_i de los MLG es un buen método para el control de los modelos, pero en este caso la estandarización de los residuos es necesaria y más complicada. Esto se debe a que si las suposiciones hechas sobre el modelo son correctas, entonces los residuos estandarizados deben tener aproximadamente la misma varianza y se deben comportar, tanto como sea posible, como los residuos de los modelos lineales. Para ello veremos dos tipos de estandarizaciones distintas.

Definición 1.3.5 (Residuos de Pearson). Se dividen los residuos entre una cantidad proporcional a la desviación estándar dada por el modelo ajustado:

$$\hat{\epsilon}_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

Deben tener media nula y varianza ϕ si el modelo es correcto.

Observación 1.3.6. Estos residuos no deberían mostrar ninguna tendencia en la media o la varianza al representarlos frente a los valores ajustados del modelo.

En la práctica los residuos de Pearson suelen ser asimétricos en torno al 0, lo que no concuerda mucho con que se parezcan a los residuos de los modelos lineales. Por tanto, se considera también la siguiente estandarización de los residuos, que surge al comparar la desviación del MLG con la suma de los residuos al cuadrado de los modelos lineales. Veámos primero a qué nos referimos con desviación:

Definición 1.3.6 (Desviación). En el contexto de la definición 1.3.1, decimos que la desviación del modelo se define como:

$$D = 2(l(\hat{\beta}_{max}) - l(\hat{\beta}))\phi = \sum_{i=1}^n 2\omega_i(y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)) \quad (1.8)$$

Donde:

- $l(\hat{\beta}_{max})$ representa el valor máximo de la función de verosimilitud logarítmica del modelo saturado (el que tiene un parámetro por dato). Este es el valor máximo que pueden tener todas las funciones de verosimilitud logarítmicas para los datos dados.
- $\tilde{\theta}_i$ es la estimación del parámetro canónico para el modelo saturado.
- $\hat{\theta}_i$ es la estimación del parámetro canónico del modelo que estamos estudiando.

Definición 1.3.7 (Residuos de desviación). Si denotamos por d_i a la i -ésima componente de la desviación de un MLG, nos queda que $D = \sum_{i=1}^n d_i$ y se definen los residuos de desviación como:

$$\hat{\epsilon}_i^d = \text{signo}(y_i - \hat{\mu}_i)\sqrt{d_i}$$

Observación 1.3.7. Obviamente se tiene que: $D = \sum_{i=1}^n (\hat{\epsilon}_i^d)^2$. Además, notemos que: $D^* = \frac{D}{\phi} \sim \chi_n^2$ y, aunque esto no se pueda trasladar directamente componente a componente, podemos intuir que:

$$\frac{d_i}{\phi} \sim \chi_1^2 \Rightarrow \hat{\epsilon}_i^d \sim N(0, \phi)$$

Es decir, intuitivamente, se comportarán como los residuos de los modelos lineales.