Apéndice B

Complementos de *MLG*

El alcance de esta tesis refiere a los modelos para datos binarios —es decir, para aquellos que provengan de procesos de Bernoulli o de variables aleatorias con distribución binomial— considerados éstos como casos particulares de los MLG. Puesto que no pretendemos hacer una extensión más general de los primeros hacia los segundos, en adelante nos valdremos de los resultados para modelos binomiales, sin entrar en mayor detalle sobre los resultados generales de los MLG. La cobertura temática de los MLG tiene una envergadura muy superior a la tratada en esta parte del trabajo, con lo cual, para una profundización tanto más general como específica sobre otros modelos distintos al logístico, las referencias citadas constituyen una guía al menos imprescindible para su estudio.

B.1. Caracterización de los *MLG*

La muy buena bibliografía existente sobre los *MLG* nos hace ser prudentes a la hora de hacer una presentación los mismos, ya que sus fundamentos teócicos no sólo se han desarrollado muy apropiadamente sino que también hay muchos investigadores que se encuentran dedicados a ellos casi con exclusividad. Sin embargo, trataremos de dejar en este apéndice los puntos más importantes de forma sintética, de modo que den sustento a ciertas cuestiones que abordamos en el cuerpo principal de la tesis.

B.1.1. Presentación

Puede considerarse que los *MLG* constituyen una familia de modelos que pretenden buscar una cierta "extensión" de los dominios de aplicabilidad del modelo lineal general[1]:

$$y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon,$$

en el sentido en que la suposición de aditividad del término de error es relajada de cierto modo. En efecto, la función de densidad de una variable y puede ser escrita desde el punto de vista del modelo lineal general como:

$$f_Y(y) = f_\varepsilon(y - \mathbf{x}'\boldsymbol{\beta}),$$

[1] *Vid.* capítulo 3, de David Firth, en Hinkley *et al.* (1991).

mientras que desde los *MLG*, su generalización conduce a una familia de modelos del tipo:

$$f_Y(y) = f(y; \mathbf{x}'\boldsymbol{\beta}),$$

en donde el vector de variables explicativas $\mathbf{x}$ sólo queda manifestado en una cierta función lineal de los parámetros, $\mathbf{x}'\boldsymbol{\beta}$.

La formulación paramétrica del modelo lineal general se puede representar indicando que el valor esperado de la respuesta y es función de un cierto vector de parámetros $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$, con lo que se tiene que:

$$E(y) = \mu(\beta_0, \beta_1, \ldots, \beta_k),$$

en donde $\mu(\cdot)$ representa una función conocida de naturaleza lineal en los parámetros:

$$\mu = \beta_0 + \sum_{j=1}^{p} x_j \beta_j. \tag{B.1}$$

Para el caso de los *MLG*, se dice que en sí mismos no son funciones lineales[2], pero que en cierto sentido sí lo son, ya que es posible encontrar una cierta función $g(\cdot)$ diferenciable y monótona tal que:

$$\mu = g^{-1}\left(\beta_0 + \sum_{j=1}^{p} x_j \beta_j\right),$$

con lo cual tendremos que no ya *la respuesta* sino *su valor esperado* será una función lineal de los parámetros. La consideración de esta última afirmación nos puede llevar a escribir la ecuación anterior como:

$$g(\mu) = \beta_0 + \sum_{j=1}^{p} x_j \beta_j \tag{B.2}$$

La comparación entre las ecuaciones (B.1) y (B.2) nos permite identificar de inmediato la diferencia "estructural" entre el modelo lineal general y el modelo lineal generalizado.

B.1.2. Familia exponencial de distribuciones

Orígenes

Los *MLG* han sido formulados dentro del contexto de una familia particular de distribuciones[3], conocida comúnmente como *familia exponencial*. A tal efecto, es habitual indicar que cada respuesta Y cuya distribución pertenezca a esta familia exponencial, tendrá una distribución de la forma:

$$f_Y(y, \theta, \phi) = \exp\left\{\frac{[y\theta - b(\theta)]}{a(\phi)} + c(y, \phi)\right\}, \tag{B.3}$$

[2] *Ídem* anterior, p. 59.
[3] *Vid.* Hardin y Hilbe (2001), p. 10.

en donde $a(\cdot)$, $b(\cdot)$ y $c(\cdot)$ son funciones que toman formas específicas, pero conocidas[4]. Los parámetros θ y ϕ representan, respectivamente, a los parámetros de posición y dispersión. En este contexto, hay dos situaciones básicas[5]:

- Si el valor del parámetro ϕ es conocido, a la familia de distribuciones dada por (B.3) se la llama *familia exponencial lineal*, y la misma queda indexada por el parámetro θ, que es llamado *parámetro natural*.
- Si, en cambio, el valor de ϕ resulta desconocido, al mismo se lo llama parámetro de dispersión, y se dice que la (B.3) define la llamada *familia exponencial de modelos de dispersión*.

Ejemplos

La familia exponencial *lineal* incluye varias distrubuciones, de las que resumimos sus valores esperados y sus varianzas:

Distribucion	Valor Esperado	Varianza	Observ.
Normal	θ	ϕ	—
Poisson	$\exp(\theta)$	$\exp(\theta)$	(a)
Gamma	$\frac{-1}{\theta}$	$\frac{\phi}{\theta^2}$	(b)
Binomial	$\frac{\exp(\theta)}{1+\exp(\theta)}$	$\frac{\phi\exp(\theta)}{[1+\exp(\theta)]^2}$	(c)

Observaciones: (a) El parámetro ϕ se toma igual a la unidad, salvo casos de sobre-dispersión. (b) El parámetro ϕ es igual a la recíproca de la llamada "gamma index". (c) El parámetro ϕ es la recíproca del número de pruebas fijas del experimento binomial, en donde y es el número de éxitos obtenido.

En este mismo grupo se incluyen también otras familias, como la binomial negativa (o distribución de Pascal), las gausianas inversas y otras de uso menos frecuente.

Formulación básica

Partiendo de la expresión (B.3), la función de distribución conjunta de una muestra aleatoria dada por el vector $\mathbf{y} = (y_1, \ldots, y_n)'$ puede escribirse como[6]:

$$f_{y_1,\ldots,y_n}(y_1, \ldots, y_n \mid \theta, \phi) = \prod_{i=1}^{n} \exp\left\{\frac{[y_i\theta_i - b(\theta_i)]}{a(\phi)} + c(y_i, \phi)\right\},$$

cuya función de log-verosimilitud correspondiente es:

$$\ell(\theta \mid \mathbf{y}, \phi) = \sum_{i=1}^{n} \exp\left\{\frac{[y_i\theta_i - b(\theta_i)]}{a(\phi)} + c(y_i, \phi)\right\}, \tag{B.4}$$

[4] *Vid.* McCullagh y Nelder (1989), p. 28.
[5] *Vid.* capítulo 3, de David Firth, en Hinkley *et al.* (1991), p. 60.
[6] *Vid.* Hardin y Hilbe (2001), p. 10, o Davison (2003), pp. 166 *et seq.*, entre otros.

ya que los valores que maximizan la función de verosimilitud $L(\theta, \phi \mid y_1, \ldots, y_n)$ son los mismos que maximizan la log-verosimilitud $\ell(\cdot) = \ln L(\cdot)$.

Suponiendo que el parámetro ϕ resulte conocido, estaremos dentro de la esfera de las familias exponenciales lineales. En este sentido, y para encontrar qué valores del parámetro natural son los que maximizan la (B.4), se resuelve la siguiente ecuación:

$$\frac{\partial}{\partial \theta}[\ell(\theta \mid \mathbf{y}, \phi)] = 0,$$

cuya solución (que no demostraremos aquí), da lugar a las siguientes dos expresiones:

- $E(y) = \ b'(\theta) = \mu$
- $V(y) = b''(\theta) \cdot a(\phi) = \Upsilon(\mu) \cdot a(\phi)$

La primera solución representa el valor esperado o media de la variable, mientras que en la segunda, el factor $\Upsilon(\cdot)$, resulta del producto de dos funciones: la primera, $b''(\theta) = \Upsilon(\mu)$, constituye la llamada "función de varianza", la cual está asociada a cada distribución en particular y puede observarse que depende del parámetro θ, es decir, de la media de la distribución. Para el caso de $a(\phi)$, la misma es independiente del parámetro θ y depende sólo de ϕ. Para el caso del modelo lineal general, en el que $V(y) = \mathbf{\Sigma}$, que verifica que $\mathbf{W} = \mathbf{\Sigma}^{-1} = \sigma^2 diag\{w_1^{-1}, \ldots, w_n^{-1}\}$, la función $a(\phi)$ muchas veces tiene la forma[7]:

$$a(\phi) = \frac{\phi}{w},$$

en donde ϕ es el parámetro de dispersión, que es asumido constante entre las observaciones i, y que también es denotado por σ^2. Para el caso de w, y por lo precitado, también se lo asocia con cada uno de los elementos individuales de la matriz de "pesos", $\mathbf{W}$.

La tabla siguiente resume los valores de esta función para algunas distribuciones de uso corriente:

Distribucion	Funcion Varianza
Normal	$\Upsilon(\mu) = 1$
Poisson	$\Upsilon(\mu) = \mu$
Gamma	$\Upsilon(\mu) = \mu^2$
Binomial	$\Upsilon(\mu) = \mu(1 - \mu)$

De esta manera, y en el contexto de las distribuciones de la familia exponencial lineal, la función varianza $\Upsilon(\mu)$ caracteriza a cada distribución[8].

[7] *Vid.* McCullagh y Nelder (1989), p. 29.
[8] *Vid.* capítulo 3, de David Firth, en Hinkley *et al.* (1991), p. 61.

B.1.3. Sobre la estimación de parámetros

Tanto en los estimadores mínimo-cuadráticos como en los máximo-verosímiles, se exige que para la estimación de los parámetros de un modelo lineal la varianza de los errores sea constante, con lo cual, en caso de que no lo sea, no será posible aplicar estos métodos para la estimación de los parámetros. Esta suposición de varianza constante para todos los niveles de la respuesta dentro de la región experimental, se reduce prácticamente a suponer que la distribución de la respuesta es normal[9]. En muchos casos, esta suposición no se cumple, como en el caso de distribuciones continuas y asimétricas, como es el caso de por ejemplo las distribuciones lognormal, gamma, etc. En conclusión, una forma de expresar lo anterior es indicando que *el nivel de la varianza de las observaciones es función de su media*. No obstante esta particularidad, existen también otras formas básicas de solucionar esta falta de constancia en la varianza para poder aplicar los métodos de mínimos cuadrados o de máxima verosimilitud.

Dada una muestra aleatoria particular $\mathbf{y}_0 = (y_{01}, \ldots, y_{0n})'$, de observaciones independientes, proveniente de una cierta población que pertenezca a la familia exponencial lineal que además depende de un vector de parámetros $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$, la función verosimilitud se utilizará para obtener el mayor valor que pueda tomar el vector de parámetros. En términos matemáticos, este problema se puede formular diciendo que si el vector de parámetros desconocido $\boldsymbol{\theta}$ toma el valor $\widehat{\boldsymbol{\theta}}_{ML}$ (valor que se llama *estimador máximo verosímil* de $\boldsymbol{\theta}$, abreviado habitualmente como *MLE*[10]), entonces se cumple la siguiente relación vectorial, formada por un sistema de $p = k+1$ incógnitas:

$$\frac{\partial}{\partial \boldsymbol{\theta}} \left\{ L\left(\boldsymbol{\theta} \mid \mathbf{y}_0\right) \right\}_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{ML}} = \frac{\partial}{\partial \boldsymbol{\theta}} \left[\ell\left(\boldsymbol{\theta}\right) \right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{ML}} = \mathbf{0} \tag{B.5}$$

en donde $\mathbf{0} = (0, \ldots, 0)'$ es un vector formado por tantos ceros como número de parámetros desconocidos se hayan considerado para el modelo. Resolviendo para aquellos valores de $\boldsymbol{\theta}$ que verifiquen que la segunda derivada sea negativa, $\frac{\partial^2}{\partial^2 \boldsymbol{\theta}} \left[\ell\left(\boldsymbol{\theta}\right) \right] < 0$, se llega a las llamadas "maximum likelihood score equations", cuyo resultado para datos normales coincide con el obtenido mediante el criterio de mínimos cuadrados, tanto en la expresión de $\boldsymbol{\beta}$ como en la de $\widehat{\sigma}^2$. En el caso en que se tengan soluciones múltiples, se toma aquella para la cual la verosimilitud $L(\boldsymbol{\theta})$ sea máxima.

Resolviendo (B.5), el procedimiento conduce a las llamadas *"score equations"*[11], las cuales pueden expresarse matricialmente como:

$$\mathbf{X}' \left(\mathbf{y} - \boldsymbol{\mu} \right) = \mathbf{0}$$

en donde: $\mathbf{X} = (\mathbf{x}_1', \ldots, \mathbf{x}_n')'$ es la matriz de diseño; $\mathbf{y}$ es el vector de respuestas observadas y $\boldsymbol{\mu}$ es el vector de medias, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$. La solución de esta ecuación —que será un sistema de $n \times (k+1)$ ecuaciones, con $n > k$ — dará un vector de estimaciones máximo verosímiles para el vector de parámetros, $\boldsymbol{\beta}$.

La profundización de este tema en particular corresponde a la etapa de *análisis* de experimentos. Siendo que nuestro trabajo está alineado con los aspectos del *diseño*,

[9] *Vid.* p. ej. Myers *et al.* (2002).
[10] Acrónimo del inglés "**M**aximum **L**ikelihood **E**stimators".
[11] *Vid.* p. ej. Myers *et al.* (2002).

consideramos que no es necesario seguir ahondando sobre estos temas particulares, ya que los mismos se encuentran resueltos y muy bien descritos en la literatura actual.

B.1.4. Caracterización fundamental de los *MLG*

Para cerrar esta introducción básica de los *MLG*, adaptamos a continuación una muy clara síntesis de los *MLG* que encontramos en Hardin y Hilbe (2001)[12], en donde los autores presentan en forma de listado con los detalles más importantes de esta clase de modelos. De acuerdo con ellos, los *MLG* son aquellos modelos que se distinguen por los siguientes aspectos[13]:

a. Poseen una componente aleatoria para la respuesta, $\mathbf{y}$, que tiene una distribución que sigue la familia exponencial. Las componentes de la respuesta, $(y_1, \ldots, y_n)'$, resultan todas independientes entre sí.

b. Poseen una componente sistemática, de naturaleza lineal, formada por el producto entre la matriz de diseño $\mathbf{X}$ y el vector de parámetros $\boldsymbol{\beta}$, y que no depende de ninguna otra cantidad más. La misma es llamada predictor lineal, y se la denota por η.

c. Existe una función monótona y diferenciable, $g(\cdot)$, que relaciona la componente aleatoria con la sistemática del modelo. La misma es llamada función "link", de modo que a partir del predictor lineal, es posible determinar el valor esperado de la respuesta[14]. Las características de esta función $g(\cdot)$ hacen que la misma tenga inversa, de modo que el valor esperado de la respuesta se calcula como la inversa de la función link, es decir: $\mu = g^{-1}(\eta) = E(y)$.

d. Cada distribución particular de la familia exponencial (lineal) se encuentra caracterizada por una cierta "función de varianza", $\Upsilon(\cdot)$, la cual depende exclusivamente de la media. Por lo tanto, para los *MLG*, se tiene que $V(y) = f[\text{cte}, \Upsilon(\mu)]$, con lo que la varianza de la respuesta resulta sólo función de la media y de una constante.

e. La estimación de los parámetros del modelo se realiza mediante algoritmos iterativos, los cuales conducen a estimaciones suficientes de aquellos, y que se adaptan a todas las distribuciones de la familia exponencial.

B.2. Un caso particular de los *MLG*: los *MDB*

B.2.1. Introducción

Como comentamos en la sección precedente, la distribución binomial (y su caso particular, la de Bernoulli) constituyen casos particulares de los *MLG*, caracterizada aquella por su función de varianza, $\Upsilon(\mu) = \mu(1 - \mu)$.

[12] Un enfoque muy similar puede verse, por ejemplo, en Davison (2003), pp. 480 *et seq.*

[13] Tomamos esta clasificación para lo que anteriormente llamamos *familia exponencial lineal*, es decir, para aquellos casos en que se conoce el valor del parámetro ϕ.

[14] *Vid.* p. ej. Hoffmann (2004), pp. 23 *et seq.*

Encontrar modelos para explicar la probabilidad de éxito en función de un conjunto de variables puede situarse en dos grandes contextos: (a) *en el de los estudios observacionales*, en los cuales el analista ya dispone de un conjunto de observaciones de la respuesta y de los valores que toman las variables explicativas, y que intentará encontrar relaciones funcionales adecuadas que puedan explicar su interrelación, y (b) *en el de los experimentos diseñados*, en los que de antemano el analista decidirá qué niveles asignará a las variables explicativas o factores, por medio de los cuales realizará observaciones de la respuesta siguiendo pautas específicas[15]. Los modelos que hemos utilizado persiguen más el segundo contexto que el primero, con lo cual, describiremos algunos modelos útiles para poder realizar observaciones de la probabilidad de éxito para valores de los factores decididos por el experimentador.

B.2.2. Definición de las variables

A la hora de buscar relaciones que expliquen la probabilidad de éxito en función de un conjunto de factores de variabilidad que se encuentren bajo control del experimentador, interesará no solamente considerar modelos sencillos de construir, sino también sencillos de interpretar. En este sentido, muchas veces se suelen preferir los modelos más *parsimoniosos* frente a los de complicada estructura[16].

Si se dispone de un conjunto de n variables independientes con distribución binomial, cada una de parámetros $\theta_i = (m_i, \pi_i)'$, es decir:,

$$(y_i \mid \boldsymbol{\theta}_i) \sim \text{indep.binom.}\,(m_i, \pi_i)\,,$$

para $i = 1, \ldots, n$ observaciones, los momentos de primero y segundo orden de la i-ésima observación estarán dados por[17]:

$$E\,(y_i \mid m_i, \pi_i) = m_i \pi_i,$$

y

$$V\,(y_i \mid m_i, \pi_i) = m_i \pi_i\,(1 - \pi_i)$$

Es frecuente[18] también que el interés se centre más en modelar la *proporción de éxitos*[19] , dada por $\frac{y_i}{m_i}$ que en el número y_i de éxitos de la condición experimental en cuestión. Esto puede verse muy a menudo en procesos de fabricación por lotes, por ejemplo. De este modo, lo anterior lleva a considerar las siguientes expresiones equivalentes —en notación simplificada— para cada una de las n observaciones:

[15] Ejemplo de ellas son: la aleatorización de las observaciones, la conveniencia de diseños por bloques, partir de matrices de diseño ortogonales, etc.

[16] *Vid.* p. ej. Kleinbaum *et al.* (1988), en donde se comenta la conveniencia sobre la utilización de modelos sencillos frente a los complejos.

[17] *Vid.* p. ej. Collett (2003).

[18] *Ibídem.*

[19] En secciones posteriores, se utilizará el resultado según el cual que la *proporción* de éxitos es el estimador máximo verosímil de la probabilidad de éxito, es decir, $\widehat{\pi} = \frac{y}{m}$. [*Vid.* p. ej. Hosmer y Lemeshow (2000) o Collett (2003), entre otros].

RESPUESTA	DISTRIBUCIÓN	MEDIA	VARIANZA
$y =$ número de éxitos	indep. binom.(m, π)	$m\pi$	$m\pi\,(1-\pi)$
$\frac{y}{m} =$ proporción de éxitos	indep. binom.$(1, \pi)$	π	$\frac{1}{m}\pi\,(1-\pi)$

La simple observación de la expresión correspondiente a la varianza de la proporción de éxitos permite no inclinarse hacia los modelos cuyos parámetros utilicen el método de los mínimos cuadrados para estimar sus parámetros. En efecto, la expresión:

$$V\left(\frac{y_i}{m_i}\right) = V\left(\widehat{\pi}_i\right) = \frac{1}{m_i}\pi_i\left(1-\pi_i\right)$$

permite ver que para cualquier nivel i de la respuesta, la proporción de éxitos observada —es decir, la probabilidad de éxitos estimada $\widehat{\pi}_i$— dependerá del verdadero valor que tenga el parámetro poblacional del modelo, es decir, la verdadera probabilidad de éxito π_i para el mismo nivel. Observado desde otro punto de vista, vemos que en los *MDB* la varianza es función de la media. Utilizando la notación general de los *MLG*, en la que $E(y) = \mu$, y en que $V(y) = \Upsilon(\mu) = \mu(1-\mu)$, lo anterior adquiere el mismo aspecto, ya que para el caso de la distribución binomial, el parámetro de dispersión ϕ es igual a la unidad.

B.3. Modelos para la probabilidad de éxito

B.3.1. Introducción

Nuestro mayor interés dentro del contexto de estudio que nos ocupa es el de poder explicar cómo varía el valor esperado de la probabilidad de éxito con los niveles de un conjunto de factores, los cuales se encuentran bajo control. Esto nos permitirá definir modelos que nos permitan conocer de qué manera se relacionan unas y otras variables, que a su vez nos servirá también para poder explorar el lugar geométrico que representen en busca de un punto máximo.

A la hora de considerar modelos para la probabilidad de éxito, resulta útil partir de las diferencias que lo distinguen del modelo lineal, acaso el más desarrollado y de uso más difundido en la práctica. Para comentar las características más importantes de estos modelos, consideraremos solamente algunos conceptos básicos[20].

Partiremos, entonces, suponiendo que es posible definir un modelo del tipo $\pi\left(\mathbf{x}, \boldsymbol{\beta}\right)$, capaz de explicar la probabilidad de éxito en función de un conjunto de factores, desde este punto de vista, el valor de la varianza será inconstante y dependerá del nivel $\mathbf{x}_i$ que tome el vector de factores.

tal como los comentamos en capítulos anteriores.

B.3.2. Génesis: el modelo de Bernoulli

Como hemos indicado en la sección anterior, los modelos de probabilidad más utilizados para los datos de naturaleza binaria imponen que la variable respuesta pueda

[20] *Vid.* Lehmann, Donald; Gupta, Sunil y Steckel, Joel (1998). *Marketing Research.* Addison-Wesley.

tomar solamente dos valores mutuamente excluyentes, *éxito* o *fracaso*, cuya definición concreta dependerá del contexto del problema estudiado. En símbolos, el espacio muestral de la variable "resultado del experimento de Bernoulli", denotada por y, será $\{0; 1\}$, en donde cada uno de los elementos corresponde a éxito o fracaso, según proceda. Si designamos como éxito al resultado "$y = 1$" del experimento anterior, resulta habitual indicar que ésta será la probabilidad de éxito del experimento, π, y que en el caso más general resultará desconocida.

Para un conjunto dado de factores de variabilidad $\mathbf{x}$, los que se presume que afectan al valor esperado de la respuesta, es natural comenzar a proponer modelos a partir de relaciones lineales del tipo:

a. *Modelo para el número de éxitos:* $y(\mathbf{x}_i) = \beta_0 + \mathbf{x}_i'\mathbf{b} + \mathbf{x}_i'\mathbf{B}\mathbf{x}_i$, para todo i,

b. *Modelo para la probabilidad de éxito:* $\pi(\mathbf{x}_i) = \beta_0 + \mathbf{x}_i'\mathbf{b} + \mathbf{x}_i'\mathbf{B}\mathbf{x}_i$, para todo i.

Teniendo en cuenta que el predictor lineal de ambos modelos, $\eta_i = \beta_0 + \mathbf{x}_i'\mathbf{b} + \mathbf{x}_i'\mathbf{B}\mathbf{x}_i$, puede tomar valores que caigan más allá del intervalo $[0; 1]$, ambos modelos resultarán inadecuados si se los considera tal cual están definidos. Queda claro que el producto escalar equivalente $\mathbf{X}\boldsymbol{\beta} = \beta_0 + \mathbf{x}_i'\mathbf{b} + \mathbf{x}_i'\mathbf{B}\mathbf{x}_i$ difícilmente quede confinado dentro del intervalo $[0, 1]$, que es el campo de validez de la probabilidad $\pi(\mathbf{x}_i, \boldsymbol{\beta})$, para cualquier i. Antes bien, este modelo implica que la función $\pi(\mathbf{x}_i, \boldsymbol{\beta})$ pueda tomar cualquier valor del intervalo $(-\infty, \infty)$, ya que cualquier componente de $\mathbf{X}$ puede variar en igual rango. Lo anterior resulta análogo para el número de éxitos y, que solamente podrá tomar ambos extremos del intervalo $[0; 1]$.

En muchas ocasiones, más que observar si el resultado de la variable y será éxito o fracaso, interesa saber más bien cuál será la probabilidad de que dicho resultado resulte éxito (o fracaso)[21]. Por ejemplo, si un grupo de características personales de un cliente de un banco quedan definidas mediante un cierto vector $\mathbf{x}_0$, se puede definir el suceso de interés[22]: "éxito = el cliente es solvente para adquirir un crédito determinado", para el cual se asigna arbitrariamente $y = 1$. De este modo, a la institución le interesará tener algún modelo que le permita evaluar diferentes clientes potenciales, cada uno con su vector de datos $\mathbf{x}_0$, de forma de encontrar qué probabilidad existiría que dicho cliente pudiera resultar solvente en caso de incurrir en un crédito y de no poder pagarlo en término. En símbolos, el foco estará puesto sobre probabilidades del tipo:

$$P(y = 1 \mid \mathbf{x}_0) = \pi(\mathbf{x}_0),$$

para las cuales interesará tener modelos confiables.

Al considerar modelos paramétricos para la probabilidad de éxito, será necesario tener en consideración un vector de parámetros $\boldsymbol{\beta}$, el cual será representativo de la importancia que tenga cada uno de los efectos de los factores que afectan al valor esperado de la respuesta. Así, en términos generales, la última expresión se puede reescribir como:

$$P(y = 1 \mid \mathbf{x}, \boldsymbol{\beta}) = \pi(\mathbf{x}, \boldsymbol{\beta})$$

[21] Para prescindir del inevitable latiguillo *"o fracaso"*, indicaremos de ahora en más —salvo que se indique lo contrario— que el resultado de interés será el "éxito", omitiendo que su definición dependerá del objetivo específico de la variable en estudio.

[22] *Vid.* p. ej. Myers *et al.* (2002).

De igual modo que en los modelos clásicos de ajustes de datos, se requerirá contar con un cierto número de observaciones de la respuesta para poder determinar estimaciones del vector desconocido $\boldsymbol{\beta}$, de forma que con ellos se tenga algún modelo útil para su fin.

No resultaría descabellado —cuanto menos como conjetura inicial— pensar que pueda existir una relación lineal de dichos coeficientes con los factores $\mathbf{x}$, de tal modo que lleven a considerar un modelo teórico para $\pi(\mathbf{x}, \boldsymbol{\beta})$ que sea de la forma:

$$P(y = 1 \mid \mathbf{x}, \boldsymbol{\beta}) = \beta_0 + \mathbf{x}'\mathbf{b} + \mathbf{x}'\mathbf{B}\mathbf{x} \tag{B.6}$$

Sin embargo, por las razones precitadas, este modelo no resultaría el más adecuado para modelar la probabilidad de éxito en función de un conjunto de factores $\mathbf{x}$. En Collett (2003) puede encontrarse una discusión más amplia sobre este asunto.

B.3.3. Extensión al modelo binomial

En el caso más general, se tendrán n variables independientes con distribución binomial, cada una con probabilidad de éxito igual a π_i y constante dentro de su condición experimental, que a su vez estará compuesta por m_i observaciones de una variable de Bernoulli. De este modo, la respuesta puede visualizarse como un vector $\mathbf{y} = (y_1, \ldots, y_n)'$, en donde cada componente estará definida como $y_i \sim$ indep. binomial(m_i, π_i)., para $i = 1, \ldots, n$. El vector de parámetros del modelo estará formado a su vez por dos vectores, cada uno con n elementos, es decir: $\boldsymbol{\theta} = (\mathbf{m}, \boldsymbol{\pi})' = (m_1, \ldots, m_n, \pi_1, \ldots, \pi_n)'$.

Análogamente a lo visto anteriormente, este modelo quedará entonces definido de la siguiente manera:

$$\{y_1, \ldots, y_n \mid m_1, \ldots, m_n, \pi_1, \ldots, \pi_n\} \sim \text{indep. binomial}(m_1, \ldots, m_n, \pi_1, \ldots, \pi_n)$$

$$\text{es decir: } \{\mathbf{y} \mid \mathbf{m}, \boldsymbol{\pi}\} \sim \text{ indep. binomial } (\mathbf{m}, \boldsymbol{\pi})$$

$$P(\mathbf{y} = \mathbf{y_0}) = p(\mathbf{y_0} \mid \mathbf{m}, \boldsymbol{\pi}) = \prod_{i=1}^{n} p(y_{0i} \mid m_i, \pi_i) = \prod_{i=1}^{n} \binom{m_i}{y_{0i}} \pi_i^{y_{0i}} \ (1 - \pi_i)^{m_i - y_{0i}}$$

Al modelar la probabilidad de éxito en función de un conjunto de factores y parámetros, y teniendo fijado el parámetro restante, m, se tendrá que la (B.6) toma el siguiente aspecto:

$$P(\mathbf{y} = \mathbf{y_0} \mid \mathbf{x}, \boldsymbol{\beta}) = \beta_0 + \mathbf{x}'\mathbf{b} + \mathbf{x}'\mathbf{B}\mathbf{x}$$

B.3.4. Sobre la componente aleatoria

Antes de describir las distintas formas disponibles para modelar $\pi(\mathbf{x}, \boldsymbol{\beta})$, resulta importante dejar en claro la poca utilidad que ofrecería un modelo que no sólo considerase una cierta parte sistemática estimada $\widehat{\pi}(\mathbf{x}, \widehat{\boldsymbol{\beta}})$, sino también que contemplara una componente aleatoria, digamos, ε.

Si se hiciera una analogía con el modelo lineal, sería entonces posible expresar el valor de la respuesta tipo binaria de la forma $y_i = \pi(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i$, en donde π será una cierta función teórica que vinculará los resultados de la respuesta observada con el vector de variables explicativas, y ε representará una componente aleatoria. Esta última será representativa de la diferencia entre la respuesta observada y la esperada (o ajustada), es decir, $\varepsilon_i = y_i - \widehat{\pi}(\mathbf{x}_i, \widehat{\boldsymbol{\beta}})$. Si no se indicara lo contrario, nada impediría pensar que estas componentes aleatorias fuesen $DIIN(0, \sigma^2)$, para todo i.

Pero este razonamiento resulta inadecuado[23] para datos de naturaleza binaria, ya que la variable y_i puede tomar solamente los valores $\{0; 1\}$, lo cual conduce a que la componente de error tome uno de los siguientes dos posibles valores[24]:

- Si $y = 1$, se sigue que $\varepsilon_i = 1 - \pi(\mathbf{x}, \boldsymbol{\beta})$, con probabilidad $P(y = 1) = \pi(\mathbf{x}, \boldsymbol{\beta})$.
- Si $y = 0$, entonces ε será igual a: $\varepsilon_i = 0 - \pi(\mathbf{x}, \boldsymbol{\beta}) = -\pi(\mathbf{x}, \boldsymbol{\beta})$, suceso que ocurre con probabilidad $P(y = 0) = 1 - \pi(\mathbf{x}, \boldsymbol{\beta})$. Queda claro que una probabilidad negativa carece de sentido en este caso.

Por estos motivos, la componente aleatoria tiene naturaleza discreta y, por lo tanto, será característicamente no normal. En otras palabras, cada residuo tiene una única distribución. En el sentido de lo anterior, la varianza de dicha componente sería:

$$V(\varepsilon_i) = \pi(\mathbf{x}_i, \boldsymbol{\beta})\left[1 - \pi(\mathbf{x}_i, \boldsymbol{\beta})\right]$$

Por tanto, ya que la probabilidad $\pi(\mathbf{x}_i, \boldsymbol{\beta})$ depende de los niveles que tomen los factores $\mathbf{x}$ y de los parámetros $\boldsymbol{\beta}$, la varianza resulta no sólo no normal sino también no constante, lo cual hace que no sean válidas las hipótesis necesarias para aplicar el criterio de los cuadrados mínimos ordinarios, que caracteriza al modelo lineal normal.

Una alternativa para solucionar este problema es utilizando el procedimiento de cuadrados mínimos ponderados para el caso de la inconstancia de la varianza, aunque algunos autores señalan ciertas dificultades de aplicabilidad, ya que los pesos con los que se ponderan las varianzas son función de $\pi(\mathbf{x}, \boldsymbol{\beta})$, la cual resulta, por lo general, desconocida[25].

B.3.5. Modelos para $\pi(\mathbf{x}, \boldsymbol{\beta})$: generalidades

En el punto anterior, indicamos que un modelo lineal para la probabilidad de éxito, del tipo $\pi(\mathbf{x}, \boldsymbol{\beta}) = \beta_0 + \mathbf{x}'\mathbf{b} + \mathbf{x}'\mathbf{B}\mathbf{x}$, no resultaba adecuado para ajustar datos de naturaleza binaria, como así también indicamos la poca utilidad de considerar una componente aleatoria para el modelo. No obstante, y partiendo de la misma forma lineal, es posible sacar una valiosa ventaja de todo lo que podría ofrecer el enfoque lineal del problema, que da lugar a al menos 3 posibilidades de modelos para $\pi(\mathbf{x}, \boldsymbol{\beta})$, en los que cada uno de ellos parte de una función link distinta:

[23] *Vid.* p. ej. Myers (1990). En Collett (2003) podemos ver una justificación bien fundamentada por la cual el modelado de $\pi(\mathbf{x}, \boldsymbol{\beta})$ mediante una expresión lineal del tipo $\beta_0 + \mathbf{x}'\mathbf{b} + \mathbf{x}'\mathbf{B}\mathbf{x}$ es inadecuada.

[24] *Vid.* Janke y Tinsley (2005), p. 495 o Hosmer y Lemeshow (2000), p. 6 y 7, entre otros.

[25] *Vid.* p. ej. Myers (1990) o Collett (2003).

a. Modelo logit

Mediante la siguiente transformación logarítmica, es posible conseguir que el modelo lineal se ajuste de forma adecuada a los datos binarios:

$$\ln\left[\frac{\pi\left(\mathbf{x}_i,\boldsymbol{\beta}\right)}{1-\pi\left(\mathbf{x}_i,\boldsymbol{\beta}\right)}\right]=\beta_0+\mathbf{x}_i'\mathbf{b}+\mathbf{x}_i'\mathbf{B}\mathbf{x}_i, \text{ para todo } i$$

Dicha transformación da lugar al llamado *modelo logit de la probabilidad de éxito*,

$$\text{logit}\left[\pi\left(\mathbf{x},\boldsymbol{\beta}\right)\right]=\beta_0+\mathbf{x}'\mathbf{b}+\mathbf{x}'\mathbf{B}\mathbf{x},$$

que permitirá aprovechar la teoría y métodos existentes para el modelo lineal, puesto que tanto la parte izquierda como la derecha de la ecuación anterior, no tienen restricciones en cuanto al campo de validez, ya que ambos pueden va-riar dentro de $(-\infty,\infty)$, teniendo ambos vectores la posibilidad de tomar valores continuos cualesquiera. Por estos motivos, los principios que guían el análisis desde el enfoque de la regresión lineal, pueden ser utilizados también para analizar datos binarios a partir del modelo logístico[26].

b. Modelo logístico

Si se resuelve la ecuación anterior para $\pi\left(\mathbf{x},\boldsymbol{\beta}\right)$, se podrá llegar al siguiente resultado[27]:

$$\pi\left(\mathbf{x}_i,\boldsymbol{\beta}\right)=\frac{\exp\left(\beta_0+\mathbf{x}_i'\mathbf{b}+\mathbf{x}_i'\mathbf{B}\mathbf{x}_i\right)}{1+\exp\left(\beta_0+\mathbf{x}_i'\mathbf{b}+\mathbf{x}_i'\mathbf{B}\mathbf{x}_i\right)}, \quad \text{para todo } i$$

Esta forma de expresar la probabilidad de éxito define el llamado *modelo logístico para la probabilidad de éxito*. Una de las ventajas en el uso de este modelo radica en que la razón de exponenciales del miembro de la derecha tiene igual campo de validez que la probabilidad de la derecha, es decir, que variará entre 0 y 1, sin imponer que tanto $\mathbf{x}$ como $\boldsymbol{\beta}$ tengan que tomar valores discretos o continuos. En la abundante bibliografía disponible[28] es posible encontrar una caracterización completa sobre todas las bondades en el uso de este tipo de modelos.

c. Otros modelos

Si bien no abordaremos en detalle otros modelos para ajustar datos binarios, mencionaremos otros dos que también se indican frecuentemente en la literatura, que son los modelos *probit* y *complementary log-log*. Partiendo del hecho de que la probabilidad de éxito se encuentra acotada entre 0 y 1, es posible asociarle a la misma un modelo emparentado con la función de distribución de probabilidad de la normal estándar, como es el caso del *modelo probit*. Para dicho modelo, la relación funcional entre el valor esperado de la respuesta y las variables explicativas será:

$$\text{probit}\left[\pi\left(\mathbf{x}_i,\boldsymbol{\beta}\right)\right]=\Phi^{-1}\left[\pi\left(\mathbf{x}_i,\boldsymbol{\beta}\right)\right]=\beta_0+\mathbf{x}_i'\mathbf{b}+\mathbf{x}_i'\mathbf{B}\mathbf{x}_i, \text{ para todo } i.$$

[26] *Vid.* p. ej. Hosmer y Lemeshow (2000).
[27] *Vid.* p. ej. Kleinbaum (1994).
[28] *Vid.* p. ej. Kleinbaum (1994); Hosmer y Lemeshow (2000); Collett (2003); entre otros.

Por otro lado, y considerando funciones logarítmicas, es posible establecer otro modelo para $\pi(\mathbf{x}, \boldsymbol{\beta})$ de la siguiente forma:

$$\operatorname{clog}[\pi(\mathbf{x}_i, \boldsymbol{\beta})] = \log\{-\log[1 - \pi(\mathbf{x}_i, \boldsymbol{\beta})]\} = \beta_0 + \mathbf{x}_i'\mathbf{b} + \mathbf{x}_i'\mathbf{B}\mathbf{x}_i, \text{ para todo } i,$$

el cual define el llamado modelo complementary log-log, también de naturaleza lineal en sus parámetros. En Collett (2003), p. 57, por ejemplo, puede encontrarse una discusión interesante sobre la convieniencia de uso de las distintas transformaciones en diferentes casos.

Tanto el modelo logístico como el logit han cobrado especial aceptación y adecuación para modelar variables binarias, por poseer propiedades estadísticas muy deseables, las cuales permiten aprovechar muchas características de modelos que gozan de un estado muy maduro de desarrollo, cual es, el modelo lineal. Por estos motivos, tomaremos al modelo logístico y al logit como herramientas de trabajo para estudiar las implicaciones que tienen los sistemas en los que intervienen variables binarias cuando se desea ganar en conocimiento acerca de su modo de funcionamiento.

B.4. Medidas de calidad de ajuste

Si bien la validación de los modelos pertenece también al ámbito del análisis de experimentos, comentaremos brevemente algunos puntos básicos sobre el tema, aplicados especialmente a posibles extensiones de los alcances de esta tesis.

Dentro de los criterios de medición de la calidad de ajuste que tengan los modelos propuestos a los datos, encontramos que los más utilizados[29] son: (a) la *razón de verosimilitudes* y (b) el *estadístico de Pearson*, cuyos resultados principales comentaremos a continuación.

B.4.1. A partir de la razón de verosimilitudes: la devianza

Una forma de comprobar la bondad del ajuste de un modelo es la de compararlo con un modelo general, de modo que éste contenga el máximo número de parámetros —e igual en número a la cantidad de observaciones disponibles de la respuesta— que puedan ser estimados. Este último suele denominarse *modelo saturado*. De este modo, cuando $n = p$, el modelo saturado será aquel que tenga una función verosimilitud máxima para la muestra dada[30]. Parece entonces razonable pensar que todo modelo "candidato" que se desee ajustar debería ser comparable con el modelo saturado en términos de su verosimilitud, para una muestra dada.

Desde el punto de vista de elección de modelos, el buen sentido estadístico sugiere preferir los modelos que contengan menor número de parámetros. Llamando $L[\boldsymbol{\pi}(\mathbf{x}_i, \widehat{\boldsymbol{\beta}})_{SAT} \mid \mathbf{y}]$ al modelo saturado y $L[\boldsymbol{\pi}(\mathbf{x}_i, \widehat{\boldsymbol{\beta}})_{MOD} \mid \mathbf{y}]$ al modelo propuesto que

[29] Existen otros criterios similares que persiguen el mismo fin, como por ejemplo, la *Prueba de Hosmer-Lemeshow* [*Vid.* p. ej. Hosmer y Lemeshow (2000)].

[30] *Vid.* p. ej. Dobson (2002).

se pretende ajustar —que será el "retador" del modelo saturado—, en virtud de lo anterior es de esperar que:

$$\dim\{L[\boldsymbol{\pi}(\mathbf{x}_i, \widehat{\boldsymbol{\beta}})_{MOD} \mid \mathbf{y}]\} \leq \dim\{L[\boldsymbol{\pi}(\mathbf{x}_i, \widehat{\boldsymbol{\beta}})_{SAT} \mid \mathbf{y}]\}.$$

A partir de esta consideración, dada una muestra de observaciones independientes, $\mathbf{y}$, se define la llamada *razón de verosimilitudes* o "likelihood ratio"[31], denotada por λ, al cociente entre la verosimilitud del modelo saturado y la del modelo propuesto:

$$\lambda = \frac{L\left[\boldsymbol{\pi}(\mathbf{x}_i, \widehat{\boldsymbol{\beta}})_{SAT} \mid \mathbf{y}\right]}{L\left[\boldsymbol{\pi}(\mathbf{x}_i, \widehat{\boldsymbol{\beta}})_{MOD} \mid \mathbf{y}\right]}$$

que proporciona un criterio para comprobar la bondad de ajuste del modelo. En la práctica, es frecuente tomar el logaritmo de la razón de verosimilitudes como estadístico de prueba, es decir:

$$\ln \lambda = \ln\left\{L\left[\boldsymbol{\pi}\left(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}\right)_{SAT} \mid \mathbf{y}\right]\right\} - \ln\left\{L\left[\boldsymbol{\pi}\left(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}\right)_{MOD} \mid \mathbf{y}\right]\right\}$$

La expresión anterior suele escribirse de forma simplificada como:

$$\ln \lambda = \ell(\widehat{\boldsymbol{\beta}})_{SAT} - \ell(\widehat{\boldsymbol{\beta}})_{MOD}$$

Por lo precitado, y a partir de este estadístico, se pueden realizar las siguientes observaciones:

- Dado que el modelo saturado es el que se supone con mayor "grado de explicación" de la respuesta para la muestra dada, es de esperar que siempre se cumpla que: $\ell(\widehat{\boldsymbol{\beta}})_{SAT} \geq \ell(\widehat{\boldsymbol{\beta}})_{MOD}$, para todo subconjunto de parámetros, $\widehat{\boldsymbol{\beta}}$, contenido en el modelo saturado.
- Cuanto mayor sea la aproximación del modelo bajo estudio hacia el modelo saturado, tanto mejor será el grado de explicación de la variabilidad de los datos mediante el primer modelo. Serán entonces deseables aquellos modelos en los que: $\ell(\widehat{\boldsymbol{\beta}})_{MOD} \to \ell(\widehat{\boldsymbol{\beta}})_{SAT}$.
- Si suponemos que para un mismo conjunto de datos la cantidad $\ell(\widehat{\boldsymbol{\beta}})_{SAT}$ permanece constante, entonces decir que el modelo ajustado será tanto más adecuado cuanto mejor se cumpla que $\ell(\widehat{\boldsymbol{\beta}})_{MOD} \to \ell(\widehat{\boldsymbol{\beta}})_{SAT}$ resulta equivalente a indicar que tanto mejor será cuanto mayor resulte $\ell(\widehat{\boldsymbol{\beta}})_{MOD}$, en términos generales[32].

Por tanto, y como primera aproximación para identificar un modelo logístico que ajuste razonablemente bien, se entiende que cuanto más pequeño sea el valor de $\ln \lambda$, tanto más se acercará la log-verosimilitud del modelo propuesto a la del saturado, lo

[31] *Ídem* anterior.

[32] En el caso de la distribución normal, esto se puede dar sin problemas puesto que la función varianza en ese caso no guarda relación con el valor que pueda tomar el valor esperado de la distribución. Sin embargo, en el caso particular de la binomial, la varianza se espera que resulte asintóticamente no muy diferente de la cantidad $\pi(1-\pi)$.

cual será un indicador de bondad de ajuste del modelo propuesto para la muestra bajo estudio.

A partir del estadístico $\ln \lambda$, se define una cantidad que resulta característica de los *MLG*, que se denomina *devianza*[33], y que se suele denotar como D. La misma se define como:

$$D = 2 \ln \lambda = 2 \left[\ell(\widehat{\boldsymbol{\beta}})_{SAT} - \ell(\widehat{\boldsymbol{\beta}})_{MOD} \right] \tag{B.7}$$

Llamando m y p al número de parámetros que tienen el modelo saturado y el propuesto, respectivamente, se demuestra que asintóticamente la devianza sigue una distribución *chi-cuadrado*[34] con $m - p$ grados de libertad[35]. En símbolos:

$$D = 2 \ln \lambda \underset{\text{(asint.)}}{\sim} \chi^2_{m-p} \tag{B.8}$$

Una de las posibilidades que permite el uso de la devianza es como medida de la bondad del ajuste del modelo bajo estudio, en cuanto a que se determina si el modelo ajustado de regresión logística es significativamente peor que el modelo saturado.

La devianza "observada" —es decir, la del modelo bajo estudio— siempre cumplirá que[36] $0 \leq D_{OBS} \leq \infty$, prefiriéndose consiguientemente aquellos modelos cuyas log-verosimilitudes "observadas" resulten lo más grandes posibles, es decir, aquellos modelos en los que D sea lo más pequeña posible. Ya que la log-verosimilitud para el modelo saturado resulta una cantidad constante para una muestra dada, *maximizar la devianza D resulta equivalente a maximizar la log-verosimilitud del modelo propuesto.*

Una regla a menudo utilizada para comprobar si la calidad del ajuste es razonablemente buena es comprobando que la cantidad $D_{OBS}/(m-p)$, llamada *devianza media*[37], no resulte apreciablemente superior a la unidad, en cuanto a que $m - p$ es la media de la distribución χ^2_{m-p}.

Devianza para datos con distribución binomial

Cuando se tienen n observaciones —o bien n condiciones experimentales, cada una definida por un cierto nivel $\mathbf{x}_i$ de un vector de k-dimensional de factores de variabilidad, $\mathbf{x}$— cada una de las cuales proveniente de distribuciones indep. binom.$[m_i, \pi(\mathbf{x}_i, \boldsymbol{\beta})]$, puede demostrarse[38] que para una muestra dada $\mathbf{y}$, su devianza estará dada por:

$$D = 2 \sum_{i=1}^{n} \left[y_i \ \log \left(\frac{y_i}{\widehat{\mu}_i} \right) + (m_i - y_i) \ \log \left(\frac{m_i - y_i}{m_i - \widehat{\mu}_i} \right) \right]$$

[33] *Vid.* p. ej. Wasserman (2004), p. 299 o Dobson (2002), p. 76, entre otros. Una notación alternativa utilizada con frecuencia para la devianza es: $D = -2 \ln \lambda^{-1} = -2 \ln\{L[\boldsymbol{\pi}(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}) \mid \mathbf{y}] / L[\boldsymbol{\pi}(\mathbf{x}_i, \widehat{\boldsymbol{\beta}})_{SAT} \mid \mathbf{y}]\}$.

[34] *Vid.* p. ej. Wasserman (2004), p. 299. Esta distribución también recibe el nombre alternativo de **"Distribución Gi-dos"**. [*Vid.* p. ej. Romero y Zunica (2001), pp. 152 *et seq.*].

[35] Cuando haya evidencia que las observaciones de la respuesta sigan una distribución normal y sean independientes entre sí, la devianza sigue *exactamente* una distribución ji-cuadrado con igual número de grados de libertad [*Vid.* p. ej. Dobson (2002)].

[36] Esto es así puesto que sigue el mismo campo de definición que la función logaritmo.

[37] Traducción del término en inglés "mean deviance". *Vid.* p. ej. Dobson (2002).

[38] *Vid.* p. ej. McCullagh y Nelder (1989).

en donde y_i será el número observado de éxitos e $\widehat{\mu}_i$ el número esperado de éxitos, dados por: $\widehat{\mu}_i = m_i\ \widehat{\pi}\,(\mathbf{x}_i, \boldsymbol{\beta})$, para $i = 1, \ldots, n$. Interesa tener en cuenta algunas consideraciones sobre los valores extremos que pudiera tomar la función devianza en datos binomiales, que sirven para identificar mejor si el modelo propuesto resulta adecuado para ajustar los datos.

A partir de la ecuación (B.8), es de esperar que la devianza del modelo ajustado verifique:

$$D \simeq m - p,$$

siendo m el número de parámetros del modelo que se está ajustando, y p el del modelo saturado. Reemplazando esta expresión por la (B.7), llegamos a que si el modelo ajusta adecuadamente los datos, tendremos que:

$$\ln[L(\widehat{\boldsymbol{\beta}})_{SAT}] - \ln[L(\widehat{\boldsymbol{\beta}})_{MOD}] \simeq \frac{1}{2}(m - p)$$

Ya que la log-verosimilitud del modelo saturado permanece constante para un mismo conjunto de datos, podemos definir a partir de la última ecuación:

$$\ln[L(\widehat{\boldsymbol{\beta}})_{MOD}] \simeq \ln[L(\widehat{\boldsymbol{\beta}})_{SAT}] - \frac{1}{2}(m - p)$$

$$L(\widehat{\boldsymbol{\beta}})_{MOD} \simeq \exp\left\{\ln[L(\widehat{\boldsymbol{\beta}})_{SAT}] - \frac{1}{2}(m - p)\right\} = \text{cte.}$$

Devianza para datos con distribución de Bernoulli

En el caso particular que se tenga que cada condición experimental tenga una sola observación ($m = 1$), se tendrán n observaciones del tipo binario, el criterio de la devianza no aportará mayor información de la bondad del ajuste realizado[39].

Devianza de modelos anidados

De acuerdo con Collett (2003), el método de la devianza alcanza su mayor utilidad a la hora de comparar modelos logísticos alternativos para datos binarios. Cuando un modelo contine un grupo de términos que forma parte de otro modelo, se dice que ambos están anidados. La diferencia en devianza entre dos modelos anidados mide hasta qué punto los términos adicionales mejoran o no el ajuste del modelo para la variable respuesta. La consideración de más o de menos términos en un modelo traerá cambios en el valor que tome la devianza de los respectivos modelos evaluados.

Conceptualmente, al comparar modelos anidados, interesará probar la significatividad de un estadístico en el que se comparen las verosimilitudes del modelo que no contiene la o las variables bajo estudio —que podemos llamar "modelo reducido"— contra la del que sí la o las contiene —o "modelo completo"—, que será del tipo[40]:

$$D = -2\ln\lambda = -2\ln\left[\frac{L\,(\text{modelo reducido})}{L\,(\text{modelo completo})}\right]$$

[39] *Vid.* p. ej. Collett (2003).
[40] *Vid.* p. ej. Hosmer y Lemeshow (2000).

En términos más específicos, y a partir de la definición de la función "link" $g\left[E\left(y \mid \mathbf{x}\right)\right] = \mathbf{x}'\boldsymbol{\beta}$, interesará ensayar hipótesis del tipo: $H_0 : g\left[E\left(\mathbf{y} \mid \mathbf{x}_{\mathbf{C}i}\right)\right] = \mathbf{x}'_{\mathbf{C}}\widehat{\boldsymbol{\beta}}_{\mathbf{C}}$ versus $H_a : g\left[E\left(\mathbf{y} \mid \mathbf{x}_{\mathbf{R}}\right)\right] = \mathbf{x}'_{\mathbf{R}}\widehat{\boldsymbol{\beta}}_{\mathbf{R}}$, en donde $\mathbf{x}_{\mathbf{C}}$ y $\mathbf{x}_{\mathbf{R}}$ representan dos configuraciones de datos anidados, para las cuales se tienen modelos diferentes. Si el diseño R se encuentra anidado en diseño C, resultará que la log-verosimilitud de R será menor que la de C, es decir, $\dim\left(R\right) < \dim\left(C\right)$[41]. En términos de razón de verosimilitudes, se tendrá entonces que:

$$-2\ln\left[\frac{L\left(\text{MODELO REDUCIDO}\right)}{L\left(\text{MODELO COMPLETO}\right)}\right] = \ell\left[\boldsymbol{\pi}\left(\mathbf{x}_{\mathbf{C}}, \widehat{\boldsymbol{\beta}}_{\mathbf{C}}\right) \mid \mathbf{y}\right] - \ell\left[\boldsymbol{\pi}\left(\mathbf{x}_{\mathbf{R}}, \widehat{\boldsymbol{\beta}}_{\mathbf{R}}\right) \mid \mathbf{y}\right]$$

en donde: $\ell[\boldsymbol{\pi}(\mathbf{x}_{\mathbf{C}}, \widehat{\boldsymbol{\beta}}_{\mathbf{C}}) \mid \mathbf{y}] > \ell[\boldsymbol{\pi}(\mathbf{x}_{\mathbf{R}}, \widehat{\boldsymbol{\beta}}_{\mathbf{R}}) \mid \mathbf{y}]$, para $\mathbf{x}_{\mathbf{R}}$ incluido en $\mathbf{x}_{\mathbf{C}}$. Simplificando la notación, lo anterior se resume expresando que: $\ell_{\mathbf{C}} > \ell_{\mathbf{R}}$.

Esta última condición llama a encontrar un criterio que defina cuándo esta diferencia es significativamente grande o pequeña para evaluar las hipótesis. Llamando $p_{\mathbf{C}}$ y $p_{\mathbf{R}}$ a los rangos de los diseños completo y reducido, respectivamente, cuando se tienen n puntos muestrales —o n condiciones experimentales— se puede demostrar[42] que el estadístico definido por:

$$\Delta_{C,R} = \frac{\left(\ell_{\mathbf{C}} - \ell_{\mathbf{R}}\right) / \left(p_{\mathbf{C}} - p_{\mathbf{R}}\right)}{\ell_{\mathbf{R}} / \left(n - p_{\mathbf{R}}\right)} = \frac{\left(D_{\mathbf{C}} - D_{\mathbf{R}}\right) / \left(p_{\mathbf{C}} - p_{\mathbf{R}}\right)}{D_{\mathbf{R}} / \left(n - p_{\mathbf{R}}\right)}$$

seguirá asintóticamente una distribución F de Fisher-Snedecor, con $p_{\mathbf{C}} - p_{\mathbf{R}}$ grados de libertad en el numerador y $n - p_{\mathbf{R}}$ grados de libertad en el denominador, cuyo cumplimiento será aproximado si la hipótesis nula es cierta. En síntesis, cuanto más se parezcan los modelos, tanto más pequeña será la diferencia $p_{\mathbf{C}} - p_{\mathbf{R}}$, y por ello, tanto más fiable será la prueba para detectar diferencias entre dos modelos anidados.

Aplicación: pruebas de hipótesis acerca de $\boldsymbol{\beta}$

Como mencionáramos anteriormente, otra forma de realizar pruebas de hipótesis para el vector de coeficientes $\boldsymbol{\beta}$ es mediante las devianzas anidadas[43]. Tal es el caso de dos modelos, digamos M_0 y M_1, cuyos datos tienen igual distribución de probabilidad y la misma función "link" —binomial y logit, respectivamente— pero la componente lineal de M_0 es un caso especial de la del modelo más general, M_1. De este modo, los vectores de coeficientes de ambos modelos serán de la forma $\boldsymbol{\beta}_0 = (\beta_1, \ldots, \beta_q)'$ y $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_p)'$, respectivamente, en donde $q < p < N$, siendo N el número total de puntos muestrales del experimento. La prueba de hipótesis que nos interesa en este momento será entonces del tipo $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ versus $H_1 : \boldsymbol{\beta} = \boldsymbol{\beta}_1$.

En términos de la diferencia de devianza de ambos modelos, y por lo visto más atrás, puede plantearse que:

$$\Delta D = D_0 - D_1 = 2(\ell_C - \ell_0) - 2(\ell_C - \ell_1) = 2(\ell_1 - \ell_0)$$

[41] Queda claro que si se comparan dos modelos reducidos, que se encuentren anidados en un mismo modelo saturado, la diferencia de log-verosimilitudes resulta equivalente a la diferencia de devianzas entre el modelo completo y el reducido. En efecto, $2\ln\lambda = 2\left(\ell_{\mathbf{C}} - \ell_{\mathbf{R}}\right) = 2\left(\ell_{\mathbf{C}} - \ell_{\mathbf{R}} + \ell_{\mathbf{SAT}} - \ell_{\mathbf{SAT}}\right) = 2\left(\ell_{\mathbf{SAT}} - \ell_{\mathbf{C}}\right) - 2\left(\ell_{\mathbf{SAT}} - \ell_{\mathbf{R}}\right) = D_C - D_R$

[42] *Vid.* p. ej. Krzanowski (1998).

[43] *Vid.* p. ej. Dobson (2002), p. 80.

Si ambos modelos son adecuados para describir los datos, entonces:

$$D_0 \sim \chi^2_{N-q} \text{ y también } D_1 \sim \chi^2_{N-p}, \text{ es decir, } \Delta D \sim \chi^2_{p-q}$$

Así, si el valor ΔD_{obs} es consistente con la distribución χ^2_{p-q}, entonces será preferible el modelo M_0 al M_1 por mayor simpleza del primero. Si el valor de ΔD_{obs} se encuentra dentro de una cierta región crítica (por ejemplo, $\Delta D_{obs} > \chi^2_{p-q,\alpha}$), se rechazará la hipótesis nula $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ en favor de H_1, con lo cual, el modelo M_1 proveerá una descripción de los datos significativamente mejor que M_0.

B.4.2. A partir del estadístico de Pearson

Otra forma usual de medir la adecuación de un modelo ajustado resulta a partir del llamado *estadístico generalizado de Pearson*[44], X^2. Partiendo del hecho que $g\left[E\left(y_i \mid \mathbf{x}_i\right)\right] = \mathbf{x}_i'\boldsymbol{\beta}$ o, su equivalente, $E\left(y_i \mid \mathbf{x}_i\right) = g^{-1}\left(\mathbf{x}_i'\boldsymbol{\beta}\right)$, dicha función quedará definida mediante:

$$X^2 = \sum_{i=1}^{n} \frac{\left[y_i - E(\widehat{y}_i \mid \mathbf{x}_i)\right]^2}{\left[\mathbf{V}\left(\widehat{y}_i \mid \mathbf{x}_i\right)\right]^{1/2}} = \sum_{i=1}^{n} \frac{\left[y_i - g^{-1}\left(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}\right)\right]^2}{\left\{\phi\ \boldsymbol{\Upsilon}\left[g^{-1}\left(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}\right)\right]\right\}^{1/2}}$$

en donde ϕ es el parámetro de escala del modelo y $\boldsymbol{\Upsilon}[g^{-1}(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})]$ es la función de varianza, que dependerá de la distribución de la familia exponencial que se esté estudiando. Para el caso de variables que provengan de distribuciones binomiales independientes, de índice m_i y parámetro $\pi\left(\mathbf{x}_i, \boldsymbol{\beta}\right)$, se tiene que $\phi = 1$, y que $\text{cov}[(y_i, y_j) \mid \mathbf{x}_i] = 0$, para $i \neq j$, lo que lleva a considerar dos consecuencias importantes:

a. $\mathbf{V}\left(\widehat{y}_i \mid \mathbf{x}_i\right) = diag\left\{V\left(\widehat{y}_i \mid \mathbf{x}_i\right)\right\} = diag\left\{\sigma_i^2\right\}$, $i = 1, \ldots, n$, y

b. $\phi\ \boldsymbol{\Upsilon}[g^{-1}(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})] = \sigma_i^2 = m_i\ \widehat{\pi}\left(\mathbf{x}_i, \boldsymbol{\beta}\right)\left[1 - m_i\ \widehat{\pi}\left(\mathbf{x}_i, \boldsymbol{\beta}\right)\right]$

A partir de estas consideraciones, el estadístico de Pearson para variables binomiales quedará de la forma:

$$X^2 = \sum_{i=1}^{n} \frac{\left[y_i - m_i\ \widehat{\pi}\left(\mathbf{x}_i, \boldsymbol{\beta}\right)\right]^2}{m_i\ \widehat{\pi}\left(\mathbf{x}_i, \boldsymbol{\beta}\right)\left[1 - m_i\ \widehat{\pi}\left(\mathbf{x}_i, \boldsymbol{\beta}\right)\left(\mathbf{x}_i, \boldsymbol{\beta}\right)\right]}$$

Este criterio, por consiguiente, pretgende elegir aquellos valores de $\boldsymbol{\beta}$ que minimicen el estadístico X^2.

B.4.3. Validación: análisis residual

Una forma corriente y no por ello menos efectiva de validar las hipótesis del modelo es mediante el *análisis residual*, que en el caso de los *MLG*, se los define a partir de los dos criterios precitados, aunque también existen otras formas [*Vid.* p. ej. McCullagh y Nelder (1989)]. Los lineamientos básicos del análisis residual siguen un esquema análogo al que se sigue en los modelos lineales, y que resumiremos a continuación.

[44] *Vid.* p. ej. McCullagh y Nelder (1989).

Algunas analogías con el modelo lineal

Para el modelo lineal, puede demostrarse que tanto el criterio de minimizar el estadístico de Pearson como el de maximizar la devianza —o la log-verosimilitud del modelo propuesto—, coinciden con el criterio de minimizar la suma de los residuos al cuadrado. Desde este punto de vista, una medida de la calidad del ajuste son justamente esos residuos, que definen la distancia euclídea entre los valores observados de la respuesta, $\mathbf{y}$, y sus valores esperados, $E\left(\widehat{\mathbf{y}} \mid \mathbf{x}\right)$:

$$SS_{res} = \sum_{i=1}^{n} e_i^2 = \text{dist } \left[\mathbf{y}, E\left(\widehat{\mathbf{y}} \mid \mathbf{x}\right)\right] = \left\|\mathbf{y} - E\left(\widehat{\mathbf{y}} \mid \mathbf{x}\right)\right\|$$

Para el caso de los *MLG*, se proponen dos distancias alternativas, que para el caso del modelo lineal, coinciden con el criterio de la suma de cuadrados de los residuos: a) una distancia basada en el *estadístico de Pearson*, denotada como $X^2\left[\mathbf{y}, E\left(\widehat{\mathbf{y}} \mid \mathbf{x}\right)\right]$, y b) otra distancia llamada *devianza*, denotada por $D\left[\mathbf{y}, E\left(\widehat{\mathbf{y}} \mid \mathbf{x}\right)\right]$, que se basa en el criterio de máxima verosimilitud.

Residuos a partir del estadístico de Pearson

A partir de lo anterior, se define el *estadístico generalizado de Pearson*, o simplemente *estadístico de Pearson*, como una medida de discrepancia entre $\mathbf{y}$ e $E\left(\widehat{\mathbf{y}} \mid \mathbf{x}\right)$ medida de la siguiente manera:

$$X^2\left[\mathbf{y}, E\left(\widehat{\mathbf{y}} \mid \mathbf{x}\right)\right] = \sum_{i=1}^{n} \frac{\left[y_i - E(\widehat{y}_i \mid \mathbf{x}_i)\right]^2}{\left[\mathbf{V}\left(\widehat{y}_i \mid \mathbf{x}_i\right)\right]^{1/2}} = \sum_{i=1}^{n} \frac{\left[y_i - g^{-1}\left(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}\right)\right]^2}{\left\{\phi\ \boldsymbol{\Psi}\left[g^{-1}\left(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}\right)\right]\right\}^{1/2}}$$

expresión que resulta equivalente a la llamada *suma de residuos de Pearson al cuadrado*, que para el caso del modelo logístico, será igual a:

$$X^2\left[\mathbf{y}, E\left(\widehat{\mathbf{y}} \mid \mathbf{x}\right)\right] = \sum_{i=1}^{n} \frac{\left[y_i - m_i\ \widehat{\pi}\left(\mathbf{x}_i, \boldsymbol{\beta}\right)\right]^2}{m_i\ \widehat{\pi}\left(\mathbf{x}_i, \boldsymbol{\beta}\right)\left[1 - m_i\ \widehat{\pi}\left(\mathbf{x}_i, \boldsymbol{\beta}\right)\right]} = \sum_{i=1}^{n} e_{iP}^2$$

De modo análogo al modelo lineal, el estadístico X^2 se distribuirá asintóticamente según una ji-cuadrado con $n - p$ grados de libertad y proporcional al parámetro de escala ϕ, es decir:

$$X^2\left[\mathbf{y}, E\left(\widehat{\mathbf{y}} \mid \mathbf{x}\right)\right] \underset{\text{(asint.)}}{\sim} \phi\ \chi^2_{n-p}$$

en donde n es el número de puntos muestrales y p el número de parámetros del modelo propuesto. La estimación puntual e insesgada del parámetro de escala[45] será directamente proporcional a este estadístico e inversamente proporcional a sus grados de libertad, es decir:

$$\widehat{\phi}_P = \frac{1}{n-p}\ X^2\left[\mathbf{y}, E\left(\widehat{\mathbf{y}} \mid \mathbf{x}\right)\right]$$

[45] Para el caso binomial, se tiene que el parámetro de escala es igual a la unidad ($\phi = 1$), salvo que se evidencie el fenómeno de *sobredispersión*, en donde dicho valor supera la unidad.

Siguiendo la analogía con el modelo normal, se tendrá que el i-ésimo residuo de Pearson —es decir, el de la i-ésima condición experimental, dada por $\mathbf{x}_i$— será de la forma:

$$e_{iP} = \frac{y_i - m_i\, \widehat{\pi}\,(\mathbf{x}_i, \boldsymbol{\beta})}{m_i\, \widehat{\pi}\,(\mathbf{x}_i, \boldsymbol{\beta})\,[1 - m_i\, \widehat{\pi}\,(\mathbf{x}_i, \boldsymbol{\beta})]}$$

cuya distribución aproximada será normal, es decir, $e_{iP} \underset{\text{(asint.)}}{\sim} N[0, V(e_{iP})]$, en donde $V(e_{iP}) \simeq \phi\,(1 - h_{ii})$, siendo h_{ii} el i-ésimo elemento de la diagonal de la llamada *matriz sombrero* o "hat matrix", $\mathbf{H}$, que en el caso particular del modelo logístico, será:

$$\begin{aligned} \text{logit}\,(\widehat{\pi}) &= \mathbf{X}\boldsymbol{\beta} = \mathbf{H}\mathbf{y} = \mathbf{X}\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}\mathbf{y} \\ \therefore \quad \mathbf{H} &= \mathbf{X}\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W} \end{aligned}$$

A partir de esto, se definen también los llamados residuos de Pearson *estandarizados*, que se calcularán como:

$$e^*_{iP} = \frac{e_{iP}}{[\phi\,(1 - h_{ii})]^{1/2}}$$

cuya distribución será aproximadamente una t-Student, con $n - p$ grados de libertad.

Residuos a partir de la devianza

Del mismo modo que se hiciera con el criterio del estadístico de Pearson, la devianza también se puede expresar como medida de discrepancia entre los valores observados y esperados de la respuesta. A partir de dicha diferencia, y extendiéndola a lo largo de todo el conjunto de puntos observacionales, se puede definir la suma al cuadrado de los llamados *residuos devianza*, que será de la forma:

$$D\,[\mathbf{y}, E\,(\widehat{\mathbf{y}} \mid \mathbf{x})] = \sum_{i=1}^{n} e^2_{iD}$$

en donde:

$$e_{iD} = \text{signo}\,[y_i - E\,(\widehat{y}_i \mid \mathbf{x}_i)] \cdot \{\text{dist}\,[y_i, E\,(\widehat{y}_i \mid \mathbf{x}_i)]\}^{1/2}$$

Análogamente con lo precitado, la devianza también se distribuirá asintóticamente como una ji-cuadrado con $n - p$ grados de libertad y proporcional al parámetro de escala, es decir,

$$D\,[\mathbf{y}, E\,(\widehat{\mathbf{y}} \mid \mathbf{x})] \underset{\text{(asint.)}}{\sim} \phi\ \chi^2_{n-p}$$

cuya estimación puntual insesgada será:

$$\widehat{\phi}_D = \frac{1}{n - p}\ D\,[\mathbf{y}, E\,(\widehat{\mathbf{y}} \mid \mathbf{x})]$$

Haciendo una analogía con la suma de cuadrados residual, la devianza es un indicador antiparsimonioso, ya que frente al agregado de términos al modelo, ésta se hace cada vez más pequeña.

En cuanto a la distribución de los residuos devianza, los mismos lo harán asintóticamente de acuerdo con una normal, es decir, $e_{iD} \underset{(\text{asint.})}{\sim} N[0, V(e_{iD})]$, en donde $V(e_{iD}) \simeq \phi\,(1 - h_{ii})$, y cuya forma estandarizada será:

$$e^*_{iP} = \frac{e_{iP}}{[\phi\,(1 - h_{ii})]^{1/2}}$$

cuya distribución será aproximadamente una t-Student, con $n - p$ grados de libertad.

Una descripción más detallada sobre la definición y análisis de los residuos en *MLG* puede encontrarse, por ejemplo, en el artículo de Pierce y Schafer (1986)[46].

B.4.4. Otras comprobaciones

Una vez que se haya realizado el análisis residual, será necesario verificar a continuación el grado de cumplimiento de las hipótesis asumidas para los *MLG*, que serán:

a. $g\,[E\,(y_i \mid \mathbf{x}_i)] = \mathbf{x}_i'\boldsymbol{\beta}$: se puede hacer un análisis gráfico de los residuos estandarizados contra $E\,(\widehat{y}_i \mid \mathbf{x}_i)$ y contra el predictor lineal $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$, comprobando que el mismo siga una relación aproximadamente lineal.

b. $V\,(y_i \mid \mathbf{x}_i) = \phi\ \boldsymbol{\Psi}[g^{-1}(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})] = \phi\ \boldsymbol{\Psi}\,[E\,(y_i \mid \mathbf{x}_i)]$: gráficamente, se comprueba que los residuos estandarizados permiten ver que la varianza de las observaciones cambiará con el nivel que tome el valor esperado.

c. *Independencia de las observaciones:* para el caso de los *MLG*, una forma muy habitual de comprobar la independencia de los valores observados de la respuesta es mediante los llamados residuos parciales, e_i^K, cuya expresión genérica para cada una de las variables explicativas del modelo es $e^j_{i\odot} = e_{i\odot} + \widehat{\beta}_j x_{ij}$, en donde el símbolo $\odot$ representa cualquiera de los dos residuos definidos anteriormente.

Por otro lado, en los *MLG* se pueden aplicar todos los mismos criterios que para el modelo lineal, como por ejemplo:

- Detección de anomalías: cuando el modelo no puede explicar adecuadamente ciertos puntos cuyos residuos estandarizados superan un cierto valor,
- Comprobación de la correcta elección de la función "link", en cuyo caso, existen pruebas de hipótesis estadísticas a tal efecto (p.ej.: prueba de Aranda-Ordaz),
- ΔD_i, o disminución de la devianza cuando se escogen un cierto número de observaciones con el regresor i-ésimo,
- ΔX_i^2, o disminución del estadístico de Pearson cuando se escogen un cierto número de observaciones con el regresor i-ésimo,
- Detección de puntos especialmente influyentes, para los cuales la distancia de Cook es una herramienta útil,

[46] *Vid.* p. ej. Pierce, Donald A. y Schafer, Daniel W. "Residuals in Generalized Linear Models". *Journal of the American Statistical Association,* dic. de 1986, vol. 81, N^o 396, p. 977 *et seq.*

- Sobredispersión: cuando la varianza de la respuesta excede el valor nominal que correspondería, se está en presencia del fenómeno de *sobredispersión*[47]. Para el caso del modelo logístico, esto equivale a indicar que en la expresión de la varianza, se deberá tener en cuenta un cierto factor de heterogeneidad[48], ϕ, mayor que la unidad, de tal modo que:

$$V(y_i) = \phi\, m_i \pi(\mathbf{x}_i, \boldsymbol{\beta}) [1 - \pi(\mathbf{x}_i, \boldsymbol{\beta})], \qquad \phi > 1$$

  el cual se puede estimar mediante observaciones repetidas[49].

B.4.5. Selección de modelos

Dos problemas básicos que ocurren cuando se agregan variables al modelo son los siguientes:

a. El ajuste "se mejora", ya que se reduce la distancia entre y , utilizando cualquiera de los criterios presentados: suma residual al cuadrado, estadístico de Pearson o devianza, pero de forma "estructural", es decir, sin distinguir si la variable agregada es realmente útil o si no lo es.

b. Desde el punto de vista del principio de parsimonia, agregar nuevos términos al modelo puede traer complicaciones para su interpretación eficiente.

Una forma de evitar estos problemas es mediante la definición de un criterio que penalice el grado de complejidad de los modelos, como por ejemplo el *Criterio de Información de Akaike* (AIC), en donde se agrega un término que es directamente proporcional al número de parámetros del modelo:

- Para la devianza: $AIC = \frac{D}{\widehat{\phi}_D} + 2p$, siendo p el número de parámetros del modelo en estudio,
- Para el estadístico de Pearson: $AIC = \frac{X^2}{\widehat{\phi}_P} + 2p$.

Otro criterio análogo al anterior es el llamado *Criterio de Información Bayesiano* (BIC), cuyo término de penalización es $p \cdot \log n$, que tiene además en cuenta el tamaño de las muestras empleadas.

[47] *Vid.* p. ej. McCullagh y Nelder (1989). Sobre las posibles causas de sobredispersión, puede leerse la misma obra o bien Dobson (2002) o Fahrmeir y Tutz (2001).
[48] *Vid.* p. ej. Lindsey (1997).
[49] *Vid.* p. ej. Montgomery *et al.* (2001).