Universidad de Los Andes
Facultad de Ciencias Económicas y Sociales
Instituto de Estadística Aplicada y Computación
Programa de Doctorado en Estadística

Estudio de la agregación de niveles en el modelo logit

Proyecto de tesis doctoral

Autor: Ernesto Ponsot Balaguer (Ing./MSc)
Tutor: Surendra Sinha (PhD)
Co-tutor: Arnaldo Goitía (Doctor)

29 de marzo de 2009

Resumen

Es un resultado conocido que la suma de dos variables aleatorias independientes binomiales, en general no resulta en una variable aleatoria binomial. Al postular un modelo logit, el investigador asume la distribución binomial como aquella subyacente a los datos, y además presupone la pertenencia de dicha distribución a la familia exponencial, manteniéndose en el ámbito del modelo lineal generalizado. No obstante, es práctica común de la aplicación estadística del modelo logit, agrupar niveles de los factores originalmente considerados, en sucesivos intentos por mejorar el ajuste del modelo final. Claramente este proceder puede violentar el supuesto distribucional. Este documento propone, en su caracter de proyecto de tesis doctoral del autor, explorar la afectación del modelo en tales situaciones, específicamente en cuanto al ajuste, al papel de la función de enlace, al estudio de residuos, tamaños de muestra requeridos y métodos gráficos de diagnóstico, así como a la factibilidad de recurrir a la inferencia exacta y no-paramétrica.

Índice de contenidos

Índice de tablas

Índice de figuras

Introducción

El modelo logit ha sido en las últimas décadas una herramienta de gran utilidad en el análisis estadístico de datos categóricos y tablas de contingencia. Se desprende como un caso particular del modelo lineal generalizado (MLG), en el que se utiliza la función logit como enlace entre los componentes aleatorio y sistemático del modelo, y se postulan factores o tratamientos explicativos en lugar de variables, al estilo del diseño de experimentos y análisis de varianza (ANOVA) convencionales.

Se modela el logaritmo de la posibilidad equivalente a una función predictora lineal en los parámetros. Su propósito es estimar y establecer la significación estadística de los factores, frente a una respuesta observada. En el proceso, se predicen las probabilidades de éxito asociadas con dicha respuesta en cada combinación de niveles de los factores.

Es común encontrar en la literatura aplicaciones de este modelo en las más diversas áreas de investigación. Interesa particularmente en el trabajo doctoral, la situación en que el investigador cuenta con una tabla de contingencia (obtenida en un estudio prospectivo o por muestreo, por ejemplo) y, luego de postular y ajustar un modelo logit a los datos, decide agrupar algunos de los niveles del factor y reiterar el análisis.

Este proceder es muy común y responde en una buena proporción de los casos al problema de respuestas con muy baja o ninguna representación en la tabla de contingencia. Como se sabe, cuando su ajuste se produce por medios asintóticos, el modelo logit es exigente con respecto a los tamaños de muestra. También abundan los ejemplos en que el investigador agrega niveles del factor, simplemente para disminuir la complejidad del análisis o bien por cuanto le interesa *a posteriori* concentrarse en algunos niveles y tratar los restantes de forma anónima.

Es el caso que al reiterar el ajuste de un modelo logit sobre una segunda tabla de contingencia con niveles agrupados del factor, en general se incurre en una violación del supuesto binomial original, cuyas implicaciones tocan especialmente a las estimaciones de la varianza. Este es el centro de la investigación pautada en la tesis doctoral.

Es entonces el propósito de este documento proyectar la investigación sobre la afectación del modelo logit en tales situaciones y, manteniéndose en el ámbito del MLG, sugerir nuevos cursos de acción que mejoren la precisión de los resultados, si ello es posible. En particular, se propone estudiar el problema en cuanto al ajuste del modelo, al papel de la función de enlace,

al estudio de residuos, los tamaños de muestra requeridos y métodos gráficos de diagnóstico, así como a la factibilidad de recurrir a la inferencia exacta y no-paramétrica.

La primera sección de este documento se dedica a la formulación del problema. La segunda se destina al relato general sobre sus antecedentes en la estadística, La tercera argumenta sobre la justificación del problema. La cuarta sección presenta los objetivos del trabajo doctoral, tanto generales como específicos. La quinta resume la teoría necesaria en principio para abordar el problema. La sexta propone a grandes rasgos la metodología que se seguirá. La séptima describe los aportes que se esperan en el campo de la estadística. Por último, la octava sección hace explícitos los resultados esperados en concreto. Finaliza el documento con las referencias biblio-hemerográficas citadas y un apéndice en que se proponen y demuestran todos los teoremas y corolarios referenciados en el texto.

1. Formulación del problema a investigar

Sea la tabla 1 un arreglo de las variables categóricas Y (respuesta binaria) versus los niveles del factor A (nominales u ordinales), en el cual n_{ij} ($i = 1, \cdots, a$ y $j = 1, 2$) representa la frecuencia simple de aparición del i-ésimo nivel del factor A y el j-ésimo nivel de la variable respuesta Y.

Tabla 1: $Y(1,2)$ vs. $A(1,...,a)$

	Y		
A	1	2	Total
1	n_{11}	n_{12}	$n_{1\cdot}$
2	n_{21}	n_{22}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	n_{a1}	n_{a2}	$n_{a\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

En presencia de una respuesta binaria, la tabla 1 queda completamente especificada con los valores n_{i1} y $n_{i\cdot}$, ya que $n_{i\cdot} = n_{i1} + n_{i2}$. Para simplificar la notación, sea entonces $n_i \equiv n_{i1}$ el número de éxitos observado en el i-ésimo nivel de A, y $t_i \equiv n_{i\cdot}$ el total de observaciones para dicho nivel.

La situación de interés en la tesis doctoral asume los distintos niveles de A independientes entre si frente a Y. También se supone que la frecuencia observada de un nivel del factor, proviene de una población binomial en el número de éxitos ($Y = 1$) de la variable respuesta, esto es,

$$N_i \overset{Ind}{\sim} \text{Bin}(t_i, p_i),\ \forall\, i = 1, \cdots, a$$

donde N_i es la variable aleatoria que representa el número de éxitos en la i-ésima población y p_i es la probabilidad de éxito asociada.

Entre muchos modelos que pueden formularse en la situación, es de interés particular para esta investigación el modelo logit. Dicho modelo es la versión del tipo ANOVA del modelo de regresión logística, y se desprende como un caso particular del MLG, originalmente propuesto por Nelder y Wedderburn (1972). La formulación del modelo es la siguiente:

$$\text{logit}(p_i) = x_i'\beta, \quad i = 1, \cdots, a \tag{1}$$

En (1), $\beta = \begin{bmatrix}\beta_1 & \beta_2 & \cdots & \beta_r\end{bmatrix}'$ es un vector de parámetros a estimar, x_i' es el i-ésimo vector fila de la matriz de diseño $X_{a\times r}$ y $\text{logit}(p_i)$ es la aplicación de la función $\text{logit}(x) = \log[x/(1-x)]$ a las probabilidades de éxito de cada población. Claramente las p_i son también objeto de estimación, por lo cual el ajuste del modelo se produce generalmente por la aplicación iterativa del método de Newton-Raphson.

En su acepción más simple, dado el hecho de que se cuenta con un sólo factor explicativo A, es común utilizar la parametrización de referencia para la matriz X. Sin pérdida de generalidad, sea a el nivel de referencia, entonces esta parametrización conduce al modelo siguiente:

$$\begin{bmatrix} \text{logit}(p_1) \\ \text{logit}(p_2) \\ \vdots \\ \text{logit}(p_{a-1}) \\ \text{logit}(p_a) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & & & & \\ 1 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{a-1} \\ \beta_a \end{bmatrix} = \begin{bmatrix} \beta_1 + \beta_2 \\ \beta_1 + \beta_3 \\ \vdots \\ \beta_1 + \beta_a \\ \beta_1 \end{bmatrix} \tag{2}$$

En (2), el modelo es saturado y no pueden calcularse los estadísticos *deviance* ni de Pearson, sin embargo, aún pueden ajustarse sus parámetros. Nótese además que existe X^{-1} puesto que X es no singular (teorema A.1).

Si el investigador cuenta con la tabla 1 y postula un modelo logit como en (2), tiene a su disposición todas las herramientas de estimación y docimasia

de hipótesis de la teoría. No obstante, si luego decide agrupar dos o más niveles del factor A, y reiterar el anásis sobre la base de una nueva tabla de contingencia derivada (lo cual es el proceder habitual de los investigadores), es muy probable que incurra en una violación del supuesto distribucional.

Así, sean los niveles $a-1$ y a aquellos que el investigador decide agrupar, haciendo $n^*_{a-1} = n_{a-1} + n_a$ y $t^*_{a-1} = t_{a-1} + t_a$. Claramente la situación puede extenderse a más de dos niveles, simplemente agregando los dos últimos, luego éstos con el anterior y así sucesivamente. El teorema A.2 establece que la suma de dos variables aleatorias independientes binomiales, con probabilidades de éxito no necesariamente iguales, en general no resulta en una variable aleatoria binomial. Esto es, si X_1 y X_2 son dos variables aleatorias independientes tales que $X_1 \sim \text{Bin}(n_1, p_1)$ y $X_2 \sim \text{Bin}(n_2, p_2)$ con $n_1 \leq n_2$, entonces, la variable aleatoria $Z = X_1 + X_2$ de distribuye como sigue:

$$\text{P}[Z = k] \quad = \quad \left(\frac{p_1}{1-p_1}\right)^k (1-p_1)^{n_1}(1-p_2)^{n_2} S(k) \tag{3}$$

donde

$$S(k) \quad = \quad \begin{cases} \displaystyle\sum_{i=0}^{k} \binom{n_1}{k-i}\binom{n_2}{i}\left[\frac{p_2(1-p_1)}{p_1(1-p_2)}\right]^i, k = 0, ..., n_1 \\ \displaystyle\sum_{i=k-n_1}^{k} \binom{n_1}{k-i}\binom{n_2}{i}\left[\frac{p_2(1-p_1)}{p_1(1-p_2)}\right]^i, k = n_1+1, ..., n_2 \\ \displaystyle\sum_{i=k-n_1}^{n_2} \binom{n_1}{k-i}\binom{n_2}{i}\left[\frac{p_2(1-p_1)}{p_1(1-p_2)}\right]^i, k = n_2+1, ..., n_1+n_2 \end{cases}$$

Por su parte, el corolario A.2.1 demuestra, haciendo uso del teorema A.2, que la distribución binomial se obtiene cuando las probabilidades de éxito son iguales.

Claramente, establecidos los supuestos de un primer modelo logit sobre la tabla de contingencia 1, un segundo modelo logit sobre una nueva tabla que agrupe los niveles $a-1$ y a, en el caso en que $p_{a-1} \neq p_a$ violenta el supuesto binomial en las poblaciones y la pertenencia de la población a la familia exponencial de distribuciones.

Consecuentemente, este trabajo persigue estudiar del problema las dos aristas siguientes:

1. Cómo se afectan las estimaciones y las pruebas de hipótesis y, si la afectación es importante, cómo pueden mejorarse.

2. Cómo puede aprovecharse la información obtenida a partir de un primer modelo logit, en el ajuste de un segundo modelo con niveles agregados de los factores.

Para ello se apelará en primer término al teorema central del límite y al método Delta (Lehmann, 1999:73,86). Conocidas las consecuencias de la agrupación de categorías en el modelo más simple, proseguirá el trabajo doctoral extendiendo o generalizando el análisis en las direcciones siguientes:

i. Cuando la variable respuesta se supone multinomial en lugar de binomial.

ii. Cuando se postulan múltiples efectos explicativos y modelos no saturados.

iii. Cuando, de resultar apropiado, se adopta otra función de enlace distinta de logit.

iv. En cuanto a sus medidas de bondad del ajuste, análisis de residuos, diagnóstico gráfico y tamaños de muestra requeridos.

v. Indagando sobre la posibilidad de aplicar al problema las técnicas de la inferencia exacta o bien de la inferencia no-paramétrica.

2. Antecedentes de la investigación

El modelo lineal general según lo define Graybill (1976:144) supone la existencia de una variable de interés o variable respuesta, cuya realización en una situación particular de datos se modela mediante la agregación de un componente no aleatorio o sistemático, y un componente aleatorio o error. El componente sistemático del modelo es una forma funcional entre un número prestablecido de parámetros desconocidos a estimar y variables de diseño o explicativas, mientras que el error es supuesto como una variable aleatoria de la cual se conocen al menos su esperanza y su varianza. La literatura especializada reporta múltiples variantes de este modelo, sin embargo, la variante más estudiada y aplicada en la práctica, supone el error distribuido según

una distribución normal de probabilidades. El planteamiento mismo induce la naturaleza aleatoria de la variable respuesta, en todo caso, es condición *sine qua non* de la técnica, la linealidad en los parámetros del componente sistemático del modelo. Si ambos requisitos son satisfechos, la estimación de los parámetros a partir de una muestra aleatoria observada, apela a las técnicas de mínimos cuadrados y máxima verosimiltud (Christensen, 2002:24), y la adecuación del modelo se prueba mediante las técnicas clásicas de ANOVA.

Nelder y Wedderburn (1972) proponen un nuevo enfoque que denominan "modelos lineales generalizados" (MLG), en el cual se supone una equivalencia entre el componente sistemático del modelo lineal general mencionado antes y una variable respuesta considerada aleatoria en si misma, a través de una función de enlace. En su trabajo seminal, los autores se centran en la respuesta distribuida según la familia exponencial uniparamétrica de distribuciones, cuyo parámetro se desconoce y debe ser estimado. La función de enlace modela la relación entre el parámetro de la distribución de la respuesta y el componente sistemático.

En particular, cuando la variable respuesta se distribuye como una normal y el enlace es la función identidad, el MLG se reduce al modelo lineal general. En cualquier otro caso, el MLG cobra su propio espacio, y tanto la estimación de parámetros, como el ANOVA, deben ser realizados por métodos diferentes a los propuestos para el modelo lineal general. Por una parte, la estimación de los parámetros requiere el uso iterativo del método de mínimos cuadrados ponderados sobre bien sea la función de verosimilitud o, cuando menos, sobre una función que denominan de "cuasi-verosimiltud" y, por la otra, proponen los autores que el ANOVA para las pruebas de adecuacidad del modelo, se generalice sobre un nuevo estadístico, *ad hoc* para cada distribución de probabilidades empleada, que denominan "*deviance*".

El marco conceptual que aporta el MLG, ha sido aprovechado notablemente en el análisis de datos categóricos y tablas de contingencia (Agresti, 2007; Rodríguez, 2008). Una variable aleatoria cuyo posible resultado es la categoría en la que se ubica un individuo u objeto de estudio (una planta de maíz que se clasifica según su variedad, un individuo al que se le clasifica según su condición socioeconómica, un paciente que se clasifica según la recuperación de una enfermedad gracias a un tratamiento médico, un producto fabril que se clasifica según si cumple o no las especificaciones de diseño, etc.), puede ser considerada proveniente de una distribución de Bernoulli (si se trata de una variable dicotómica) o de una distribución multinomial (si se trata de una variable policotómica). Luego, el número de individuos en la mues-

tra que caen en determinada categoría, puede ser modelado mediante una distribución binomial (si se trata de variables independientes dicotómicas), producto-multinomial (si se trata de variables independientes policotómicas) o bien multinomial (si se trata de variables dependientes policotómicas).

Consecuentemente, la formulación de un modelo que permita al investigador abandonar en parte las restricciones de continuidad y normalidad, pero que ofrezca la potencialidad de hacer inferencia formal, es de gran ayuda en el análisis de este tipo de datos. En particular el modelo logit, dentro del marco de los MLG, ha sido profusamente estudiado y caracterizado [véanse por ejemplo Cox (1970); McCullagh y Nelder (1989); Agresti (2007); Rodríguez (2008)]. Sin embargo, poco se ha profundizado en las consecuencias que acarrea la violación de sus supuestos. En particular, las consecuencias de la violación del supuesto binomial para las poblaciones subyacentes, es el objeto de inspección del trabajo doctoral.

La ecuación (3) es una generalización de la suma de ensayos de Poisson (Feller, 1968:218), es decir, ensayos independientes de Bernoulli con probabilidades de éxito en general diferentes. Se trata de una realización particular de la distribución de probabilidades conocida en la literatura como Poisson-Binomial (que no pertenece a la familia exponencial de distribuciones), y ha sido estudiada desde varios puntos de vista a partir de la aparición de los trabajos de Neyman (1939). No obstante, al no tener una forma análitica simple, ha recibido atención casi exclusivamente por la vía de las aproximaciones numéricas (Sprott, 1958; Hodges y Le˜Cam, 1960; Ollero˜H. y Ramos˜R., 1991; Weba, 1999; Roos, 1999; Neammanee, 2005).

La estimación de sus parámetros no presenta problemas desde el punto de vista numérico, menos aún hoy en día con la gran capacidad computacional disponible, sin embargo, su tratamiento analítico, por ejemplo, formando parte de la función de verosimiltud, es otra cuestión. Una alternativa sería entonces proceder con base en la función de cuasi-verosimilitud (Wedderburn, 1974) en lugar de la función de verosimilitud, sin embargo, lamentablemente, tampoco es posible encontrar una asociación funcional entre la media y la varianza para el caso, con lo cual queda descartada esta posibilidad (McCullagh y Nelder, 1989:337).

En síntesis, los antecedentes del problema se encuentran en los conceptos del MLG, en particular, el modelo logit, combinados con el estudio de los ensayos de Poisson y la distribución Poisson - Binomial derivada.

3. Justificación

El problema de la agrupación de niveles en el modelo logit no parece haber despertado el interés de los teóricos de la estadística, no obstante ser un problema real, de implicaciones prácticas importantes, puesto que se le ha utilizado copiosamente para el análisis de datos. Parece razonable intuir que no ha sido objeto de preocupación, debido a que las estimaciones puntuales de los parámetros no se ven afectadas con la violación del supuesto distribucional.

Ahora bien, el supuesto binomial en el número de éxitos en cada nivel implica que $\mathrm{V}[N_i] = t_i p_i(1-p_i)\ \forall\ i$. Sin embargo, cuando el investigador agrupa los niveles $a-1$ y a formando una nueva variable aleatoria, llámese N^*_{a-1}, puede encontrarse frente al problema de que la nueva variable no es binomial (si $p_{a-1} \neq p_a$). Aún así, si nuevamente ejecuta el procedimiento de ajuste del modelo logit, la varianza se supone como $t^*_{a-1}p^*_{a-1}(1-p^*_{a-1})$, donde $t^*_{a-1} = t_{a-1} + t_a$ y $p^*_{a-1} = \mathrm{E}[N^*_{a-1}]/t^*_{a-1} = (t_{a-1}p_{a-1} + t_a p_a)/(t_{a-1}+t_a)$. Pero, como establece el corolario A.2.2, la verdadera varianza para el caso es $\mathrm{V}[N^*_{a-1}] = t_{a-1}p_{a-1}(1-p_{a-1}) + t_a p_a(1-p_a)$, y el teorema A.3 establece que ambos supuestos no son equivalentes.

La figura 1 ilustra el comportamiento de ambas varianzas cuando se fijan los parámetros t_{a-1}, t_a y p_{a-1}, y se permite la variación de p_a. En la figura se observa que $\mathrm{V}_{\mathrm{Bin}}[N^*_{a-1}]$ parece encontrarse siempre por encima de $\mathrm{V}[N^*_{a-1}]$. Una demostración formal de este hecho puede consultarse en Nedelman y Wallenius (1986) y en las referencias citadas allí. Además, resalta de la figura, que mientras más lejanos están los valores de p_{a-1} y p_a, mayor es la diferencia que puede esperarse entre ambas varianzas. Esto sugiere inmediatemente, que tal vez no valga la pena realizar estas consideraciones cuando ambas probabilidades son muy próximas, sin embargo, la afectación del modelo puede llegar a ser importante a medida que ambas probabilidades sean más distantes.

Luego, es clara la existencia de un problema cuando se agrupan niveles y se insiste en el modelo logit sin variar el supuesto binomial, especialmente en lo atinente a la estimación de las varianzas, y en consecuencia de los errores estándar utilizados para las pruebas de hipótesis sobre los parámetros del modelo.

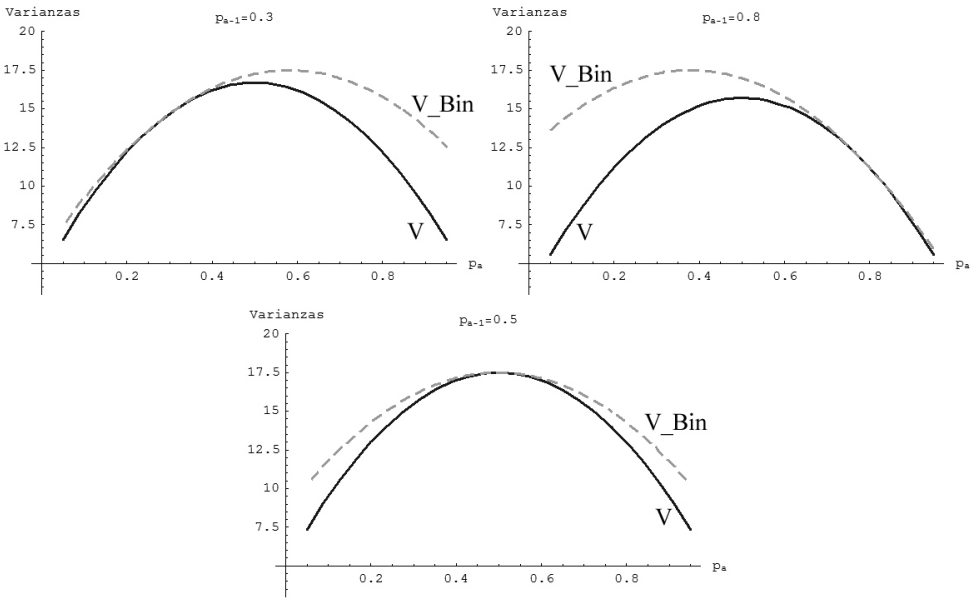

Figura 1: Comparación de $\mathrm{V}_{\mathrm{Bin}}[N^*_{a-1}]$ y $\mathrm{V}[N^*_{a-1}]$ con $t_{a-1} = 20$, $t_a = 50$

4. Objetivos de la investigación

4.1. Objetivo General

Estudiar en el marco de los MLG, la afectación que sufre el modelo logit cuando el investigador agrupa niveles de los factores explicativos. Postular cursos de acción alternativos a los procedimientos habituales, que mejoren la calidad de las estimaciones, si ello es posible.

4.2. Objetivos Específicos

1. Estudiar el comportamiento del modelo logit frente a la violación del supuesto binomial (o multinomial) de la variable respuesta. Proporcionar soluciones asintóticas y compararlas mediante simulación con los métodos establecidos.

2. Estudiar la afectación de las técnicas de bondad del ajuste, el papel de la función de enlace, los residuos, tamaños de muestra requeridos y métodos gráficos de diagnóstico, en presencia de la agrupación de

niveles.

3. Estudiar la factibilidad de recurrir a la inferencia exacta y no-paramétrica como alternativas para mejorar las estimaciones obviando el supuesto distribucional.

4. Diseñar un programa computacional que dé soporte a las técnicas que se propongan fruto del estudio del problema de agrupación de niveles.

5. Marco Teórico

En esta sección se presenta una síntesis de los conceptos estadísticos más importantes con relación al tema abordado y las razones por las cuales se inscriben en el trabajo doctoral que se propone.

5.1. El modelo lineal general

Sea $Y_{n\times 1}$ el vector de respuestas de una muestra aleatoria de tamaño n observada. Sea $X_{n\times m}$ la matriz de diseño (cuyos elementos son ceros o unos exclusivamente) o bien la matriz de covariables o variables explicativas (cuyos elementos son los valores de dichas variables dispuestos en forma matricial). Sea $\beta_{m\times 1}$ un vector cuyos elementos son m parámetros desconocidos a estimar, y sea $e_{n\times 1}$ el error (no observable), con media $\mathrm{E}[e] = 0_{n\times 1}$ y matriz de varianzas y covarianzas $V[e] = \Sigma_{n\times n}$. Se conoce como **modelo lineal general**, a la relación funcional mostrada en la ecuación (4).

$$Y = X\beta + e \tag{4}$$

En (4), $X\beta$ representa el componente sistemático del modelo y e representa el componente aleatorio, por lo tanto Y es supuesto un vector aleatorio, fruto de la suma de una constante y una variable aleatoria. Evidentemente, el propósito de este modelo es explorar las relaciones de causalidad entre la respuesta obtenida y los efectos que la ocasionan (caso de diseño de experimentos), o entre la respuesta obtenida y las variables que la explican (caso de regresión), manteniendo la forma funcional más simple, lineal, entre los parámetros desconocidos β que deben ajustarse.

Este modelo ha sido estudiado muy ampliamente desde hace ya más de un siglo. En particular, se ha estudiado y aplicado copiosamente bajo el supuesto

de normalidad de los errores y, más específicamente aún, en los casos en que dichos errores sean normales independientemente distribuidos, que se llamará "caso 1" ($e \sim \mathrm{N}(0, \sigma^2 I)$, con $I_{n\times n}$ la matriz identidad) o distribuidos normales pero no independientes que se llamará "caso 2" ($e \sim \mathrm{N}(0, \sigma^2 V)$, con $V_{n\times n}$ una matriz conocida definida positiva). En el caso 1, puede demostrarse que el estimador de mínimos cuadrados (y también máximo verosímil) del vector de parámetros β está dado por la ecuación (5).

$$\widehat{\beta} \;=\; \min_{\beta}[(Y - X\beta)'(Y - X\beta)] = (X'X)^{-}X'Y \tag{5}$$

$$\text{O equivalentemente, } \widehat{\beta} : X\widehat{\beta} = MY$$

donde $M = X(X'X)^{-}X'$ es la matriz de proyección perpendicular sobre $C(X)$, el espacio de columnas de la matrix X (Christensen, 2002:25).

Para el caso 2, Christensen (2002:34) demuestra que el estimador $\widehat{\beta}$, ahora de mínimos cuadrados generalizados de β, satisface la ecuación (6)

$$X(X'V^{-1}X)^{-}X'V^{-1}Y = X\widehat{\beta} \tag{6}$$

Así, dada una muestra observada, el procedimiento básico a seguir es el siguiente:

1. Prueba del modelo propuesto versus un modelo reducido en cuanto al número de parámetros. La idea es seleccionar aquel modelo que contenga el menor número de parámetros posibles.

   Sea el modelo original como en (4) y sea un modelo reducido $Y = X_0\gamma + e$, en el cual $C(X_0) \subset C(X)$. Es claro que en el modelo original $\mathrm{E}[Y] = X\beta$, mientras que en el modelo reducido $\mathrm{E}[Y] = X_0\gamma$. Para tamaños de muestra suficientemente grandes, la prueba de hipótesis H_0: $\mathrm{E}[Y] \in C(X_0)$ vs. H_1: $\mathrm{E}[Y] \in C(X)$ y $\mathrm{E}[Y] \notin C(X_0)$, puede ser contrastada de manera que H_0 se rechaza si se verifica la inecuación (7) para el caso 1 (Christensen, 2002:59) u (8) para el caso 2 (Christensen, 2002:86).

$$\begin{aligned} &\frac{Y'(M - M_0)Y/r(M - M_0)}{Y'(I - M)Y/r(I - M)} \\ &> \quad F(1 - \alpha; r(M - M_0), r(I - M)) \end{aligned} \tag{7}$$

$$\begin{aligned} &\frac{Y'(A - A_0)'V^{-1}(A - A_0)Y/[r(X) - r(X_0)]}{Y'(I - A)'V^{-1}(I - A)Y/[n - r(X)]} \\ &> \quad F(1 - \alpha; r(X) - r(X_0), n - r(X)) \end{aligned} \tag{8}$$

En las ecuaciones (7) y (8), M y M_0 son respectivamente las matrices de proyección perpendicular sobre $C(X)$ y $C(X_0)$, $A = X(X'V^{-1}X)^{-}X'V^{-1}$ y A_0 es una expresión equivalente sustituyendo X por X_0, F es el complemento de la función de distribución F central y $r(\cdot)$ es el rango de la matriz del argumento. Consecuentemente, si H_0 se rechaza, el modelo original prevalece frente al modelo reducido.

2. Estimación de los parámetros (o funciones de ellos) y pruebas de significación. Los parámetros del modelo se estiman según las ecuaciones (5) y (6) y se utiliza el ANOVA para contrastar las hipótesis de nulidad sobre ellos.

3. Análisis residual. Se basa en el examen de las desviaciones observadas entre los valores de la muestra y los valores predichos por el modelo. Estas desviaciones se conocen como **residuos** y son estimadores del error. Su inspección y análisis es útil para verificar si existen violaciones a los supuestos (especialmente violaciones al supuesto de normalidad, homocedasticidad y autocorrelación).

Estos y otros tantos procedimientos no mencionados, establecen las técnicas universalmente aceptadas para contrastar hipótesis de ajuste de un conjunto de datos de naturaleza aleatoria, a un modelo lineal. Se trata de un marco conceptual con una fuerte base analítica e importantes apelaciones a la teoría asintótica, por consiguiente, cualquier otro conjunto de conceptos que pretenda generalizarlo o perfeccionarlo, debe tomarlo en consideración como base de partida. Es el caso del MLG, objeto de estudio en la presente propuesta, y es por ello que habrá de ser repasado en profundidad en el transcurso del trabajo doctoral.

5.2. El modelo lineal generalizado (MLG)

Escríbase el componente sistemático del modelo lineal general en (4) por los vectores fila de la matriz X como $x_i'\beta = \begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{im} \end{bmatrix} \beta$. Sea y_i, $i = 1, 2, ..., n$, una muestra aleatoria de tamaño n de una población cuya densidad de probabilidades es $f(\cdot)$. Sea μ_i la media teórica de la i-ésima observación de la muestra. Una **función de enlace** es cualquier transformación $g(\mu_i)$, uno a uno, continua y diferenciable de la media teórica μ_i (Rodríguez, 2008:B.3).

La cantidad $\eta_i = x_i'\beta$ se denomina la i-ésima **preditora lineal**. Un **modelo lineal generalizado (MLG)**, es entonces cualquier modelo, centrado

en el estudio de la media teórica, que postule lo establecido en (9)

$$\begin{aligned} \eta_i &= g(\mu_i) \text{ o bien} \\ \mu_i &= g^{-1}(\eta_i) = g^{-1}(x_i'\beta) \end{aligned} \tag{9}$$

Nótese que el MLG describe en términos funcionales de la predictora lineal, la esperanza de la muestra ($\mu_i = \mathrm{E}[Y_i]$), en vez de la muestra en si misma (Y_i), como hace el modelo lineal general. No obstante, puede ser observado que el modelo lineal general surge como un caso particular del MLG, cuando $g(\mu_i) = \mu_i$, es decir, cuando $g(\cdot)$ es la función identidad. En efecto, si $g(\mu_i) = \mu_i$, entonces de (9) $\mu_i = g^{-1}(x_i'\beta) = x_i'\beta$, luego, desplegando en forma matricial, $\mathrm{E}[Y] = X\beta$, tal como sucede en el modelo lineal general cuando se toma la esperanza.

La teoría para el MLG ha sido ampliamente estudiada a partir del artículo ya citado de Nelder y Wedderburn (1972), sin embargo, la mayoría de los desarrollos a la fecha se circunscriben al supuesto de que la distribución de la muestra pertenece a la **familia exponencial de distribuciones**, esto es,

$$f(y_i) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right] \tag{10}$$

En (10), θ_i es un parámetro de posición, llamado parámetro canónico, mientras que ϕ es un parámetro de escala. Además, $a_i(\cdot)$, $b(\cdot)$ y $c(\cdot,\cdot)$ son funciones conocidas y el teorema A.4 establece que $\mathrm{E}[Y_i] = \mu_i = b'(\theta_i)$ y $\mathrm{V}[Y_i] = \sigma_i^2 = b''(\theta_i)a_i(\phi)$.

Son múltiples las distribuciones que pertenecen a la familia exponencial. Por mencionar sólo algunas están la gamma (y, claro está, la exponencial), la normal, la Poisson, la binomial, la beta y otras. Cuando un MLG postula que $\eta_i = g(\mu_i) = \theta_i$ se dice que emplea una **función de enlace canónico**, y algunas simplificaciones ocurren. Funciones de enlace canónico ampliamente utilizadas para tres supuestos distribucionales se muestran en la tabla 2 (Rodríguez, 2008:B.4). Por ejemplo, para la distribución binomial, el teorema A.5 establece su pertenencia a la familia exponencial de distribuciones y el teorema A.6 establece el logit como la función de enlace canónico correspondiente.

El procedimiento de estimación de los parámetros utiliza iterativamente el método de mínimos cuadrados ponderados propuesto por McCullagh y Nelder (1989:40). En este método es necesario iterar reiteradamente, hasta conseguir una precisión preestablecida en las estimaciones, debido a que la

Tabla 2: Funciones de enlace canónico para algunas distribuciones

Distribución	Enlace
Normal	Identidad
Binomial	logit
Poisson	log

media buscada se estima a partir de los parámetros de la predictora lineal, los cuales a su vez deben ser estimados también, de modo que todos los parámetros no pueden ser estimados en una sola operación. Este método de mínimos cuadrados supone el siguiente algoritmo:

0. Inicio

1. Realice una estimación inicial del vector de parámetros β, por ejemplo haciendo $\widehat{\mu} = Y$, y calculando la solución para β del sistema $\eta = X\beta = g(\widehat{\mu})$. Llámese a esta estimación inicial $\widehat{\beta}^0$

2. Haga $j = 0$

3. Repita hasta alcanzar la precisión deseada para $\widehat{\beta}$:

   3.1. Calcule la estimación de la predictora lineal en la j-ésima iteración como $\widehat{\eta}_i^j = x_i'\widehat{\beta}^j$, la cual se utiliza para computar la estimación de la media como $\widehat{\mu}_i^j = g^{-1}(\widehat{\eta}_i^j)$.

   3.2. Calcule las componentes del vector auxiliar z^j como

$$z_i^j = \widehat{\eta}_i^j + (y_i - \widehat{\mu}_i^j) \left.\frac{\mathrm{d}\eta_i}{\mathrm{d}\mu_i}\right|_{\widehat{\mu}_i^j}$$

   Nótese que z_i^j es una suerte de linealización de la función de enlace aplicada sobre la variable respuesta original y_i en la j-ésima iteración.

   3.3. Calcule el ponderador

$$w_i^j = \left\{ \left[\left(\frac{\mathrm{d}\eta_i}{\mathrm{d}\mu_i} \right)^2 \widehat{\mathrm{V}[Y]} \right] \Bigg|_{\widehat{\mu}_i^j} \right\}^{-1} = \left\{ \left[\left(\frac{\mathrm{d}\eta_i}{\mathrm{d}\mu_i} \right)^2 b''(\theta_i) a_i(\phi) \right] \Bigg|_{\widehat{\mu}_i^j} \right\}^{-1}$$

3.4. Haga $j = j + 1$

3.4. Calcule la siguiente estimación de β mediante la regresión de los z_i^{j-1} en función de los x_i usando los pesos w_i^{j-1}, esto es:

$$\widehat{\beta}^j = (X'W^{j-1}X)^{-1}X'W^{j-1}z_i^{j-1}$$

En esta estimación de mínimos cuadrados, $W^{j-1} = \text{diag}(w_i^{j-1})$

4. Fin

McCullagh y Nelder (1989:41) demuestran que el algoritmo precedente conduce a estimaciones máximo-verosímiles. Por su parte Rodríguez (2008:B.4) desarrolla el modelo suponiendo $a_i(\phi) = \phi/p_i$ y particularizando para $p_i = 1$. El autor muestra en sus notas que la matriz de información de Fisher $I(\beta) = X'WX/\phi$ y por lo tanto, $\widehat{\beta}$ tiene una distribución asintótica normal con media β y varianza $(X'WX)^{-1}\phi$. Este resultado puede ser utilizado (en muestras grandes) para probar hipótesis sobre los parámetros mediante la χ^2 de Wald.

La prueba de bondad del ajuste del MLG, bajo el supuesto de muestras independientes de la familia exponencial de distribuciones, se conduce empleando la dócima de la razón de verosimilitud generalizada entre el modelo en cuestión (llámese M) y el modelo saturado (llámese M^*), y construyendo un estadístico bautizado como **deviance** cuya expresión es $D(\mu, y) = 2\sum_{i=1}^{n}[y_i(\theta_i^* - \theta_i) - b(\theta_i^*) + b(\theta_i)]$ (teorema A.7). $D(\mu, y)/\phi$ (suponiendo $a_i(\phi) = \phi$) se distribuye según una χ^2 con $v^* - v$ grados de libertad, donde v^* es el número de parámetros de M^* y v el número de parámetros de M.

Como ejemplo, y debido a las múltiples aplicaciones que tiene para el estudio de datos categóricos (eje central de la propuesta doctoral), se desarrolla a continuación el modelo lineal generalizado bajo el supuesto distribucional binomial, conocido como **modelo logit binario**.

5.2.1. Ejemplo: Modelo logit binario

La tabla 3 de contingencia resume una situación en que se investiga un experimento aleatorio, cuyos resultados posibles son solamente éxito (1) o fracaso (0), tomando muestras de individuos (u objetos) que pueden ser clasificados también de dos formas supuestas independientes: A y B.

La tabla 3 presenta la situación más sencilla posible para un experimento del tipo relatado, expuesta aquí sólo con el propósito de simplificar al máximo

Tabla 3: Frecuencias tipo (2×2)

Grupo	Éxitos (1)	Fracasos (0)	Total
A	k_{11}	k_{12}	$k_{1\cdot} = k_{11} + k_{12}$
B	k_{21}	k_{22}	$k_{2\cdot} = k_{21} + k_{22}$
Total	$k_{\cdot 1} = k_{11} + k_{21}$	$k_{\cdot 2} = k_{12} + k_{22}$	$k_{\cdot\cdot} = \sum_{ij} k_{ij}$

la exposición. Conceptualmente, el desarrollo puede ser extendido al caso de una tabla de contingencia de dimensiones $n \times 2$ y más allá, de dimensiones $n \times m$. Incluso estos resultados pueden extenderse de forma similar a tablas de contingencia de 3 y más entradas, con cualquier número de niveles.

Una forma de interpretar la tabla 3 es verla como la estructura de una muestra aleatoria de tamaño $n = 2$, de igual número de poblaciones binomiales con parámetros $k_{1\cdot}$, p_1 y $k_{2\cdot}$, p_2, respectivamente. El parámetro p_1 es la probabilidad de obtener un éxito en el experimento original, dado que el individuo pertenece al grupo A, y p_2 es la probabilidad de éxito dado que pertenece al grupo B. El teorema A.5 establece que la distribución binomial pertenece a la familia exponencial de distribuciones y el teorema A.6 establece que logit es la función de enlace canónico para este caso. Así, un MLG que puede ser postulado para la situación es el siguiente:

$$\begin{aligned} \eta_i &= \operatorname{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = g(\mu_i) \\ &= \log\left(\frac{\mu_i}{k_{i\cdot} - \mu_i}\right) = \log(\mu_i) - \log(k_{i\cdot} - \mu_i) \\ &= x_i'\beta, \quad i = 1, 2 \end{aligned} \tag{11}$$

Empleando la parametrización conocida como "de referencia", la matriz X contiene dos columnas solamente. La primera es una columna de 1's y la segunda es la columna asociada con los niveles del único factor explicativo con que se cuenta. Expuesto de forma matricial, el modelo se describe como sigue:

$$\eta = \begin{bmatrix}\eta_1 \\ \eta_2\end{bmatrix} = \begin{bmatrix}\operatorname{logit}(p_1) \\ \operatorname{logit}(p_2)\end{bmatrix} = \begin{bmatrix}g(\mu_1) \\ g(\mu_2)\end{bmatrix} = X\beta = \begin{bmatrix}1 & 1 \\ 1 & 0\end{bmatrix}\begin{bmatrix}\beta_1 \\ \beta_2\end{bmatrix}$$

Por otra parte, en este caso $a_i(\phi) = \phi = 1$, luego

$$\frac{\mathrm{d}\eta_i}{\mathrm{d}\mu_i} = \frac{1}{\mu_i} + \frac{1}{k_{i\cdot} - \mu_i} = \frac{k_{i\cdot}}{\mu_i(k_{i\cdot} - \mu_i)}$$

Así, para z_i^j, w_i^j y utilizando el teorema A.5, resulta

$$\begin{aligned} z_i^j &= \widehat{\eta}_i^j + (y_i - \widehat{\mu}_i^j) \left(\frac{k_{i\cdot}}{\mu_i(k_{i\cdot} - \mu_i)} \right) \Bigg|_{\widehat{\mu}_i^j} \\ &= \widehat{\eta}_i^j + \frac{k_{i\cdot}(y_i - \widehat{\mu}_i^j)}{\widehat{\mu}_i^j(k_{i\cdot} - \widehat{\mu}_i^j)} \end{aligned}$$

$$\begin{aligned} w_i^j &= \left\{ \left[\left(\frac{k_{i\cdot}}{\mu_i(k_{i\cdot} - \mu_i)} \right)^2 \mu_i(k_{i\cdot} - \mu_i) \right] \Bigg|_{\widehat{\mu}_i^j} \right\}^{-1} \\ &= \left\{ \left(\frac{k_{i\cdot}^2}{\mu_i(k_{i\cdot} - \mu_i)} \right) \Bigg|_{\widehat{\mu}_i^j} \right\}^{-1} \\ &= \left(\frac{k_{i\cdot}^2}{\widehat{\mu}_i^j(k_{i\cdot} - \widehat{\mu}_i^j)} \right)^{-1} \end{aligned}$$

Para este ejemplo, el modelo es saturado y carece de sentido el cálculo de los estadísticos *deviance* y de Pearson.

5.2.2. Diagnóstico de la bondad del ajuste

Los residuos ($\widehat{e}$) de un modelo lineal general estiman el error y son las desviaciones entre los valores observados y los valores predichos por el modelo. En notación de Christensen (2002:321), $\widehat{e} = Y - X\widehat{\beta} = (I - M)Y$.

Se espera que satisfagan lo más apropiadamente posible los supuestos del modelo. Por ejemplo, si se ha supuesto normalidad, los $\widehat{e}_i$ deben resultar aproximadamente normales. Si se ha supuesto independencia, los residuos no deberían mostrar correlaciones marcadas. Si se ha supuesto homocedasticidad, los residuos deberían reflejar igualdad de varianzas, y así sucesivamente. El análisis residual es por consiguiente una parte importante del trabajo con modelos lineales generales.

Exite una batería de pruebas, tanto teóricas como gráficas, para explorar el comportamiento de los residuos en el modelo lineal general, no obstante, no resulta tan simple su exploración en el caso de los MLG, visto que este tipo de modelos no cuenta (explícitamente) con el componente aleatorio con que cuenta el modelo lineal general. Los residuos ahora deben ser redefinidos

en términos ya no de las observaciones sino de su media, y dependen claro está, de la función de enlace que se utilice.

En particular han sido bien desarrolladas algunas técnicas para el análisis de residuos en modelos de regresión logística binaria (Pardoe y Cook, 2002; Stokes y otros, 2000:217), como por ejemplo la curva ROC, sin embargo, el desarrollo no ha avanzado tanto en la regresión logística multinomial, ni en otras particularizaciones del MLG. Consecuentemente, estudiar y proponer métodos inspirados en la curva ROC [como por ejemplo proponen Cai y Zheng (2007)] que permitan evaluar las bondades de un modelo multinomial, ofrece buenas oportunidades de investigación.

Otra aproximación que vale la pena explorar como medida de adecuacidad, es la "profundidad de regresión" (Rousseeuw y Hubert, 1999) [RD, por sus siglas en inglés]. Aunque estos conceptos no fueron originalmente desarrollados para comparar modelos, al permitir el juego de distintas rectas de ajuste y de regresión, pueden eventualmente servir a este propósito. Esto también es una oportunidad de investigación para el trabajo doctoral, aplicando las ideas de RD a los modelos logit, y en particular al caso de la agregación de niveles, que es el interés primordial.

5.2.3. La curva ROC

Una forma de diagnóstico de mucha aceptación para el particular caso de la regresión logística binaria a dos vías, en su condición de clasificadora binaria de los datos, es el despliegue de la curva ROC (por *Receiver Operating Characteristic*, en inglés, o Característica de Operación del Receptor, en español). La curva ROC despliega la sensibilidad de la prueba en función de 1 menos la especificidad de la prueba (Agresti, 2007:143). La sensibilidad se entiende como la fracción de positivos verdaderos, esto es, la fracción de observaciones que fueron detectadas por el modelo como éxitos, siendo que en realidad son éxitos, mientras que la especificidad se refiere a 1 menos la fracción de falsos positivos, esto es, 1 menos la fracción de observaciones que fueron detectadas por el modelo como éxitos, cuando en realidad son fracasos. Así, la curva ROC se puede formar equivalentemente como la fracción de verdaderos positivos en función de la fracción de falsos positivos. Este diagnóstico es muy útil, entre otras cosas, debido a que muestra una gráfica fácil de leer y comprender.

5.2.4. Profundidad de regresión (RD)

Dado $\widehat{\theta}$ un "ajuste", esto es, vector de parámetros estimados en un modelo, y Z_n, matriz de las covariables observadas ampliada con la respuesta, también observada, se define la *profundidad* del conjunto $\text{depth}(\widehat{\theta}, Z_n)$ como el menor número de observaciones de Z_n que deben ser retiradas para que $\widehat{\theta}$ deje de ser un ajuste, o pase a ser un "no ajuste".

Intuitivamente, para el caso de la regresión lineal, es claro que en un ajuste perfecto, esto es, un ajuste tal que todos los puntos observados pertenecen a la recta estimada, la profundidad es n, pues habría que retirar todas las observaciones para lograr que el ajuste perfecto pase a ser un no ajuste. En cualquier otro caso la profundidad es un número entero $< n$. Así, la profundidad es claramente una medida de la bondad del ajuste de un modelo, de forma tal que a mayor profundidad, mayor perfección en el ajuste logrado.

Por ejemplo, en particular para la regresión lineal simple, el modelo $y_i = \widehat{\theta}_1 x_i + \widehat{\theta}_2,\ \ i = 1, 2, ..., n$ puede ser estudiado según su profundidad, ahora denominada *profundidad de regresión*, atendendiendo a los residuos calculados como $r_i = y_i - \widehat{\theta}_1 x_i - \widehat{\theta}_2$, de manera que un candidato a ajuste pasa a ser un no ajuste si $\exists\ v \neq x_i\ \forall\ i : (r_i < 0\ \forall\ x_i < v \wedge r_i \geq 0\ \forall\ x_i > v) \vee (r_i > 0\ \forall\ x_i < v \wedge r_i \leq 0\ \forall\ x_i > v)$.

En términos conceptuales esto significa que se puede determinar la RD por dos vías equivalentes, a saber:

1. En el espacio de las observaciones: Rotando la recta de ajuste candidata en torno a un eje imaginario determinado por su íntersección con la recta $x = v$, y contando el número de observaciones que son barridas hasta alcanzar la verticalidad (ellas son el número de observaciones que habría que retirar para hacer el ajuste candidato un no ajuste).

2. En el espacio de los residuos: Fijo v, contando el número de residuos que tendrían que cambiar de signo de manera que a la izquierda de la recta $x = v$, o bien todos los residuos fuesen positivos o todos negativos, y a su derecha, igualmente todos fuesen positivos o todos negativos.

Como ejemplo, véanse las figuras 2, 3 y 4. Dado un conjunto de datos fijo, en la figura 2 se despliega una recta de regresión ajustada por el método convencional de mínimos cuadrados y sus correspondientes residuos. Nótese que en este caso $RD = 3$, pues éste es el número de observaciones que habría que retirar para alcanzar un no ajuste.

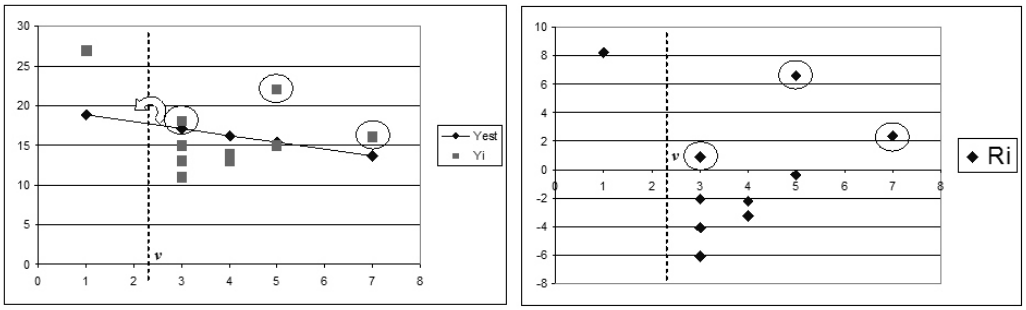

Figura 2: Regresión por mínimos cuadrados $y = -0,86x + 19,65$

En la figura 3 se despliega una recta arbitraria, candidata a ajuste para el mismo conjunto de datos, con sus correspondientes residuos. Ahora $RD = 1$, pues éste es el número de observaciones que habría que retirar para alcanzar un no ajuste.

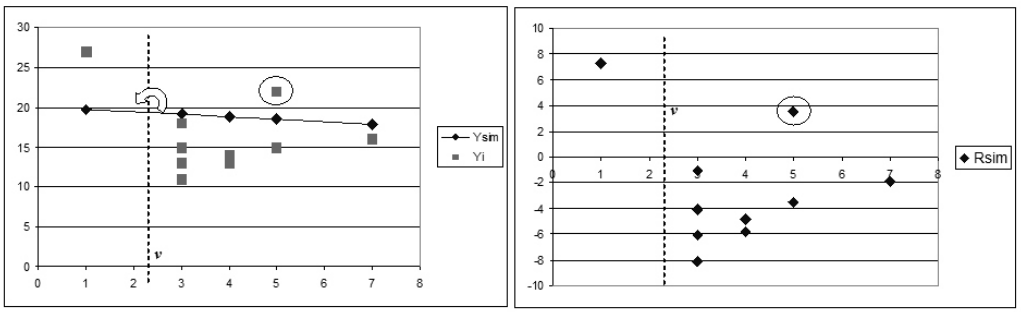

Figura 3: Ajuste arbitrario $y = -0,3x + 20$

Por último, en la figura 4 se despliega otra recta arbitraria, también candidata a ajuste para el mismo conjunto de datos, y sus correspondientes residuos. Ahora $RD = 0$, pues esta recta arbitraria es originalmente un no ajuste.

Claramente la extensión de estas ideas, del ámbito de los modelos lineales generales y la regresión clásica, al ámbito de los MLG y la regresión logística o el modelo logit, ofrecen oportunidades de investigación. Las figuras sugieren que la mejor (más alta) RD de un modelo lineal, es precisamente la recta de regresión por mínimos cuadrados. Será interesante indagar en primer término, si son aplicables estas ideas a modelos no intrínsecamente lineales (como sería

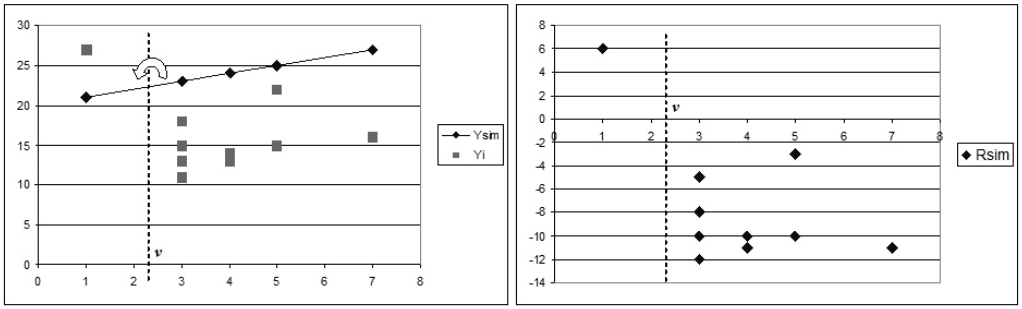

Figura 4: No ajuste $y = x + 20$

el caso del modelo logit) y, más allá, de ser aplicables, cómo se afecta la RD cuando se violenta el supuesto binomial en las poblaciones.

6. Metodología

Para el cumplimento de los objetivos previstos se apelará al método científico en su sentido clásico, esto es, planteamiento del problema, hipótesis, pruebas, experimentación y conclusiones. Se propone un trabajo de carácter deductivo y experimental, con mayor inclinación teórica que práctica. El planteamiento de los problemas específicos que se abordarán, surgirá de la revisión en profundidad de la literatura especializada, conducida por los lineamientos entregados en este documento. Debe contener una fuerte dósis de creatividad y originalidad.

La formulación de hipótesis se refiere a la proposición de nuevas teorías, técnicas o métodos, útiles para ser aplicadas sobre datos categóricos, susceptibles de ser comprobadas bien sea mediante demostraciones matemáticas y estadísticas, o bien comparadas mediante experimentos de simulación.

La fase de pruebas y experimentación se refiere a la contrastación de las hipótesis, como se señaló, en forma teórica y/o experimental.

Por último, las conclusiones se refieren a los resultados obtenidos sobre las hipótesis contrastadas, en el sentido de su verificación o negación, así como a la puesta a punto de la metodología necesaria para la aplicación práctica de aquéllas que se verifiquen.

6.1. Cronograma de actividades

Se propone el cronograma de actividades que se muestra en la tabla 4.

Tabla 4: Cronograma tentativo

Actividad	Duración
Revisión biblio-hemerográfica	6 meses
Planteamiento de hipótesis	3 meses
Pruebas teóricas	8 meses
Pruebas de simulación	4 meses
Redacción del trabajo	3 meses
Total	24 meses

7. Aportes en el campo de la Estadística

Como producto del trabajo doctoral, se espera realizar aportes originales al concimiento estadístico en cuanto al problema de la agregación de niveles en el modelo logit. Específicamente se esperan resultados originales en cuanto a los siguientes aspectos del problema:

1. Demostrar teóricamente o mediante simulación la afectación del modelo logit en presencia de la agrupación de niveles.

2. Proponer métodos de ajuste del modelo logit, alternativos a los habituales, que tomen en consideración el problema de la agregación de niveles.

3. Deducir el papel de la función de enlace en tal circunstancia.

4. Proponer métodos apropiados para el estudio de residuos, cálculos de los tamaños de muestra requeridos y diagnóstico gráfico.

5. Establecer la factibilidad de la inferencia exacta y no-paramétrica para el ajuste del modelo logit con niveles agrupados.

6. Diseñar y programar un sistema computacional que dé soporte a las distintas técnicas propuestas en el trabajo doctoral.

8. Resultados esperados

Se espera producir, además del manuscrito contentivo de la tesis doctoral y los programas computacionales que la acompañarán, cuando menos tres artículos científicos con el nivel suficiente como para ser sometidos a arbitraje en revistas científicas reconocidas. Estos artículos versarán sobre resultados parciales de la tesis y, al menos uno de ellos, deberá ser sometido a la consideración de alguna revista especializada en estadística, preferentemente de carácter internacional.

Referencias

Agresti, Alan (2007). *An Introduction to Categorical Data Analysis.* John Wiley & Sons, Inc., NJ, EEUU, 2ª edición.

Cai, Tianxi y Zheng, Yingye (2007). «Model checking for ROC regression analysis». *Biometrics*, **(63)**, pp. 152–163.

Christensen, Ronald (2002). *Plane Answers to Complex Questions. The Theory of Linear Models.* Springer-Verlag, NY, EEUU, 3ª edición.

Cox, D.R. (1970). *Analysis of Binary Data.* Methuen and Co Ltd., London, 1ª edición.

Feller, William (1968). *An Introduction to Probability Theory and Its Applications. Volumen I.* John Wiley & Sons. Inc., NY, EEUU, 3ª edición.

Graybill, Franklin (1976). *Theory and Application of the Linear Model.* Duxbury Press, CA, EEUU, 1ª edición.

Hodges, J. L. y Le Cam, Lucien (1960). «The Poisson approximation to the Poisson Binomial distribution». *The Annals of Mathematical Statistics*, **31(3)**, pp. 737–740.

Lehmann, E.L. (1999). *Elements of large-sample theory.* Springer-Verlag, NY/EEUU, 1ª edición.

McCullagh, P. y Nelder, J. (1989). *Generalized Linear Models.* Chapman & Hall, London, UK, 2ª edición.

Neammanee, K. (2005). «A refinement of Normal approximation to Poisson Binomial». *International Journal of Mathematics and Mathematical Sciences*, **(5)**, pp. 717–728.

Nedelman, Jerry y Wallenius, Ted (1986). «Bernoulli trials, Poisson trials, surprising variances, and Jensen's inequality». *The American Statistician*, **40(4)**, pp. 286–289.

Nelder, J.A. y Wedderburn, R.W.M. (1972). «Generalized Linear Models». *Journal of the Royal Statistical Society. Serie A*, **(135)**, pp. 370–384.

Neyman, J. (1939). «On a new class of 'contagious' distributions, applicable in entomology and bacteriology». *The Annals of Mathematical Statistics*, **10(1)**, pp. 35–57.

Ollero H., J. y Ramos R., H. M. (1991). «La distribución Hipergeométrica como Binomial de Poisson». *Trabajos de Estadística*, **6(1)**, pp. 35–43.

Pardoe, Iain y Cook, R. Dennis (2002). «A graphical method for assessing the fit of a logistic regression model». *The American Statistician*, **56(4)**, pp. 263–272.

Rodríguez, Germán (2008). «Lectures Notes about Generalized Linear Models». EEUU. Http://data.princeton.edu/wws509/notes.

Rohatg, V. y Ehsanes, A. (2001). *An Introduction to Probability and Statistics*. John Wiley and Sons, Inc., NY, EEUU, 2ª edición.

Roos, Bero (1999). «Asymptotics and sharp bounds in the Poisson approximation to the Poisson-Binomial distribution». *Bernoulli*, **5(6)**, pp. 1021–1034.

Rousseeuw, Peter J. y Hubert, Mia (1999). «Regression depth». *Journal of The American Statistical Association*, **94(446)**, pp. 388–402.

Sprott, D. A. (1958). «The method of maximum likelihood applied to the Poisson Binomal distribution». *Biometrics*, **14(1)**, pp. 97–106.

Stokes, Maura E.; Davis, Charles S. y Koch, Gary G. (2000). *Categorical data analysis using the SAS system*. SAS Publishing, NC, EEUU, 2ª edición.

Weba, Michael (1999). «Bounds for the total variation distance between the Binomial and the Poisson distribution in case of medium-sized success probabilities». *Journal of Applied Probability*, **(36)**, pp. 497–104.

Wedderburn, R.W.M. (1974). «Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method». *Biometrika*, **61(3)**, pp. 439–447.

Apéndices

A. Teoremas y demostraciones

Teorema A.1. *La matriz $X_{a\times r}$ dada por*

$$X = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & & & & \\ 1 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

es no singular.

Demostración. El cálculo del determinante por la descomposición en cofactores (X_{ij}), pivotando los elementos de la columna 1 resulta

$$\begin{aligned} |X| &= \sum_{i=1}^{a} x_{i1}(-1)^{i+1}|X_{i1}| \\ &= (-1)^2|X_{11}| + (-1)^3|X_{21}| + \cdots + (-1)^a|X_{(a-1)1}| + (-1)^{a+1}|X_{a1}| \\ &= (-1)^2(0) + (-1)^3(0) + \cdots + (-1)^a(0) + (-1)^{a+1}|I| = (-1)^{a+1} \neq 0 \end{aligned}$$

luego X es no singular. □

Teorema A.2. *Sean X_1 y X_2 dos variables aleatorias independientes tales que $X_1 \sim Bin(n_1, p_1)$ y $X_2 \sim Bin(n_2, p_2)$ con $n_1 \leq n_2$. Entonces, la variable aleatoria $Z = X_1 + X_2$ de distribuye como sigue:*

$$P[Z = k] = \left(\frac{p_1}{1-p_1}\right)^k (1-p_1)^{n_1}(1-p_2)^{n_2} S(k) \qquad (12)$$

donde

$$S(k) = \begin{cases} \sum_{i=0}^{k} \binom{n_1}{k-i} \binom{n_2}{i} \left[\frac{p_2(1-p_1)}{p_1(1-p_2)}\right]^i, k = 0, ..., n_1 \\ \sum_{i=k-n_1}^{k} \binom{n_1}{k-i} \binom{n_2}{i} \left[\frac{p_2(1-p_1)}{p_1(1-p_2)}\right]^i, k = n_1+1, ..., n_2 \\ \sum_{i=k-n_1}^{n_2} \binom{n_1}{k-i} \binom{n_2}{i} \left[\frac{p_2(1-p_1)}{p_1(1-p_2)}\right]^i, k = n_2+1, ..., n_1+n_2 \end{cases}$$

Demostración.

$$\begin{aligned} \mathrm{P}[Z=0] &= \mathrm{P}[X_1=0, X_2=0] = \sum_{i=0}^{0} \mathrm{P}[X_1=0-i, X_2=i] \\ \mathrm{P}[Z=1] &= \mathrm{P}[X_1=1, X_2=0] + \mathrm{P}[X_1=0, X_2=1] = \sum_{i=0}^{1} \mathrm{P}[X_1=1-i, X_2=i] \\ &\vdots \\ \mathrm{P}[Z=n_1] &= \mathrm{P}[X_1=n_1, X_2=0] + \cdots + \mathrm{P}[X_1=0, X_2=n_1] \\ &= \sum_{i=0}^{n_1} \mathrm{P}[X_1=n_1-i, X_2=i] \\ \mathrm{P}[Z=n_1+1] &= \mathrm{P}[X_1=n_1, X_2=1] + \cdots + \mathrm{P}[X_1=0, X_2=n_1+1] \\ &= \sum_{i=1}^{n_1+1} \mathrm{P}[X_1=n_1+1-i, X_2=i] \\ &\vdots \\ \mathrm{P}[Z=n_2] &= \mathrm{P}[X_1=n_1, X_2=n_2-n_1] + \cdots + \mathrm{P}[X_1=0, X_2=n_2] \\ &= \sum_{i=n_2-n_1}^{n_2} \mathrm{P}[X_1=n_2-i, X_2=i] \\ \mathrm{P}[Z=n_2+1] &= \mathrm{P}[X_1=n_1, X_2=n_2-n_1+1] + \cdots + \mathrm{P}[X_1=1, X_2=n_2] \\ &= \sum_{i=n_2-n_1+1}^{n_2} \mathrm{P}[X_1=n_2+1-i, X_2=i] \end{aligned}$$

$$\vdots$$

$$\mathrm{P}[Z = n_2 + n_1] \;=\; \mathrm{P}[X_1 = n_1, X_2 = n_2] = \sum_{i=n_2}^{n_2} \mathrm{P}[X_1 = n_1 + n_2 - i, X_2 = i]$$

$$\therefore\ \mathrm{P}[Z = k] \;=\; \begin{cases} \sum_{i=0}^{k} \mathrm{P}[X_1 = k - i, X_2 = i], & k = 0, ..., n_1 \\ \sum_{i=k-n_1}^{k} \mathrm{P}[X_1 = k - i, X_2 = i], & k = n_1 + 1, ..., n_2 \\ \sum_{i=k-n_1}^{n_2} \mathrm{P}[X_1 = k - i, X_2 = i], & k = n_2 + 1, ..., n_1 + n_2 \end{cases}$$

Ahora bien, como X_1 y X_2 son independientes,

$$\mathrm{P}[X_1 = r, X_2 = s] = \binom{n_1}{r} p_1^r (1 - p_1)^{n_1 - r} \binom{n_2}{s} p_2^s (1 - p_2)^{n_2 - s}$$

Luego, para un k fijo,

$$\sum_{i(k)} \mathrm{P}[X_1 = k - i, X_2 = i] \;=\; \sum_{i(k)} \binom{n_1}{k - i} p_1^{k-i} (1 - p_1)^{n_1 - k + i} \binom{n_2}{i} p_2^i (1 - p_2)^{n_2 - i}$$

$$= \left(\frac{p_1}{1 - p_1}\right)^k (1 - p_1)^{n_1} (1 - p_2)^{n_2} \sum_{i(k)} \binom{n_1}{k - i} \binom{n_2}{i} \left[\frac{p_2(1 - p_1)}{p_1(1 - p_2)}\right]^i$$

Consecuentemente,

$$\mathrm{P}[Z = k] \;=\; \left(\frac{p_1}{1 - p_1}\right)^k (1 - p_1)^{n_1} (1 - p_2)^{n_2} S(k)$$

donde

$$S(k) \;=\; \begin{cases} \sum_{i=0}^{k} \binom{n_1}{k - i} \binom{n_2}{i} \left[\frac{p_2(1 - p_1)}{p_1(1 - p_2)}\right]^i, k = 0, ..., n_1 \\ \sum_{i=k-n_1}^{k} \binom{n_1}{k - i} \binom{n_2}{i} \left[\frac{p_2(1 - p_1)}{p_1(1 - p_2)}\right]^i, k = n_1 + 1, ..., n_2 \\ \sum_{i=k-n_1}^{n_2} \binom{n_1}{k - i} \binom{n_2}{i} \left[\frac{p_2(1 - p_1)}{p_1(1 - p_2)}\right]^i, k = n_2 + 1, ..., n_1 + n_2 \end{cases}$$

□

Corolario A.2.1. *Si* $p_1 = p_2 = p \Rightarrow Z \sim Bin(n = n_1 + n_2, p)$.

Demostración.

$$\begin{aligned} \mathrm{P}[Z = k] &= \left(\frac{p_1}{1-p_1}\right)^k (1-p_1)^{n_1}(1-p_2)^{n_2} S(k) \\ &= p^k (1-p)^{n-k} S(k) \end{aligned}$$

Y ya que $\{[p_2(1-p_1)]/[p_1(1-p_2)]\}^i = \{[p(1-p)]/[p(1-p)]\}^i = 1 \; \forall \; i$,

$$S(k) = \begin{cases} \sum_{i=0}^{k} \binom{n_1}{k-i}\binom{n_2}{i}, k = 0, ..., n_1 \\ \sum_{i=k-n_1}^{k} \binom{n_1}{k-i}\binom{n_2}{i}, k = n_1 + 1, ..., n_2 \\ \sum_{i=k-n_1}^{n_2} \binom{n_1}{k-i}\binom{n_2}{i}, k = n_2 + 1, ..., n_1 + n_2 \end{cases}$$

Ahora, haciendo uso de propiedades combinatorias (Rohatg y Ehsanes, 2001:191) y Feller (1968:46), se tiene:

a) Para $k = 0, ..., n_1$

$$S(k) = \sum_{i=0}^{k} \binom{n_1}{k-i}\binom{n_2}{i} = \binom{n_1 + n_2}{k} = \binom{n}{k}$$

b) Para $k = n_1 + 1, ..., n_2$

$$\begin{aligned} S(k) &= \sum_{i=k-n_1}^{k} \binom{n_1}{k-i}\binom{n_2}{i} \\ &= \binom{n_1 + n_2 - n_2}{n_1}\binom{n_2}{k-n_1} + \binom{n_1 + n_2 - n_2}{n_1 - 1}\binom{n_2}{k-(n_1-1)} \\ &\quad + \cdots + \binom{n_1 + n_2 - n_2}{0}\binom{n_2}{k} = \binom{n_1 + n_2}{k} = \binom{n}{k} \end{aligned}$$

c) Para $k = n_2 + 1, ..., n_1 + n_2$

$$\begin{aligned} S(k) &= \sum_{i=k-n_1}^{n_2} \binom{n_1}{k-i}\binom{n_2}{i} \\ &= \binom{n_1+n_2-n_2}{n_1}\binom{n_2}{k-n_1} + \binom{n_1+n_2-n_2}{n_1-1}\binom{n_2}{k-(n_1-1)} \\ &\quad + \cdots + \binom{n_1+n_2-n_2}{k-n_2}\binom{n_2}{n_2} \\ &= \binom{n_1+n_2}{k} = \binom{n}{k} \end{aligned}$$

$$\therefore P[Z=k] = p^k(1-p)^{n-k}S(k) = \binom{n}{k}p^k(1-p)^{n-k}$$

□

Corolario A.2.2. *$E[Z] = n_1p_1 + n_2p_2$ y $V[Z] = n_1p_1(1-p_1) + n_2p_2(1-p_2)$.*

Demostración. Por una parte,

$$\mathrm{E}[Z] = \mathrm{E}[X_1 + X_2] = \mathrm{E}[X_1] + \mathrm{E}[X_2] = n_1p_1 + n_2p_2$$

y dado que X_1 y X_2 son independientes,

$$\mathrm{V}[Z] = \mathrm{V}[X_1 + X_2] = \mathrm{V}[X_1] + \mathrm{V}[X_2] = n_1p_1(1-p_1) + n_2p_2(1-p_2)$$

□

Teorema A.3. *Con $N^*_{a-1} = N_{a-1} + N_a$, la varianza bajo el supuesto binomial ($V_{Bin}[N^*_{a-1}]$) en general es diferente de la verdadera varianza ($V[N^*_{a-1}]$). Ambas varianzas coinciden cuando $p_{a-1} = p_a = p$.*

Demostración. Por una parte,

$$\begin{aligned} \mathrm{V}_{\mathrm{Bin}}[N^*_{a-1}] &= t^*_{a-1}p^*_{a-1}(1-p^*_{a-1}) \\ &= (t_{a-1}+t_a)\left(\frac{t_{a-1}p_{a-1}+t_ap_a}{t_{a-1}+t_a}\right)\left(1-\frac{t_{a-1}p_{a-1}+t_ap_a}{t_{a-1}+t_a}\right) \\ &= \frac{(t_{a-1}p_{a-1}+t_ap_a)(t_{a-1}+t_a-t_{a-1}p_{a-1}-t_ap_a)}{t_{a-1}+t_a} \\ &= \frac{t^2_{a-1}p_{a-1}(1-p_{a-1}) + t_{a-1}t_a[p_{a-1}(1-p_a)+p_a(1-p_{a-1})] + t^2_ap_a(1-p_a)}{t_{a-1}+t_a} \\ &\neq t_{a-1}p_{a-1}(1-p_{a-1}) + t_ap_a(1-p_a) = \mathrm{V}[N^*_{a-1}] \end{aligned}$$

Por otra parte, si $p_{a-1} = p_a = p$

$$\begin{aligned} \mathrm{V}_{\mathrm{Bin}}[N^*_{a-1}] &= \frac{t^2_{a-1}p(1-p) + t_{a-1}t_a[p(1-p)+p(1-p)] + t^2_a p(1-p)}{t_{a-1}+t_a} \\ &= \frac{(t^2_{a-1} + 2t_{a-1}t_a + t^2_a)p(1-p)}{t_{a-1}+t_a} \\ &= \frac{(t_{a-1}+t_a)^2 p(1-p)}{t_{a-1}+t_a} \\ &= (t_{a-1}+t_a)p(1-p) = \mathrm{V}[N^*_{a-1}] \end{aligned}$$

□

Teorema A.4. *Sea Y_i una variable aleatoria cuya distribución pertenece a la familia exponencial de distribuciones, esto es:*

$$f(y_i) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right]$$

donde θ_i es el parámetro canónico, ϕ es un parámetro de escala y $a_i(\cdot)$, $b(\cdot)$ y $c(\cdot,\cdot)$ son funciones conocidas, entonces $E[Y_i] = \mu_i = b'(\theta_i)$ y $V[Y_i] = \sigma^2_i = b''(\theta_i)a_i(\phi)$.

Demostración. Si $f(\cdot)$ es una función de densidad de probabilidades, entonces se verifica que

$$\int_{Y_i} f(t)\mathrm{d}t = \int_{Y_i} \exp\left[\frac{t\theta_i - b(\theta_i)}{a_i(\phi)} + c(t,\phi)\right]\mathrm{d}t = 1$$

Luego, derivando en ambos miembros y empleando la regla de Leibniz,

$$\frac{\partial}{\partial\theta_i}\left(\int_{Y_i} f(t)\mathrm{d}t\right) = \int_{Y_i}\frac{\partial f(t)}{\partial\theta_i}\mathrm{d}t = 0$$

Ahora bien,

$$\begin{aligned} \frac{\partial f(t)}{\partial\theta_i} &= \frac{\partial}{\partial\theta_i}\left\{\exp\left[\frac{t\theta_i - b(\theta_i)}{a_i(\phi)} + c(t,\phi)\right]\right\} \\ &= \frac{t - b'(\theta_i)}{a_i(\phi)}\exp\left[\frac{t\theta_i - b(\theta_i)}{a_i(\phi)} + c(t,\phi)\right] \\ &= \frac{t - b'(\theta_i)}{a_i(\phi)} f(t) \end{aligned}$$

Así,

$$\begin{aligned}
\int_{Y_i} \frac{\partial f(t)}{\partial \theta_i} \mathrm{d}t &= \int_{Y_i} \frac{t - b'(\theta_i)}{a_i(\phi)} f(t) \mathrm{d}t \\
&= \int_{Y_i} \frac{t}{a_i(\phi)} f(t) \mathrm{d}t - \frac{b'(\theta_i)}{a_i(\phi)} \int_{Y_i} f(t) \mathrm{d}t \\
&= \frac{E[Y_i]}{a_i(\phi)} - \frac{b'(\theta_i)}{a_i(\phi)} = 0 \\
&\therefore \quad E[Y_i] = b'(\theta_i)
\end{aligned}$$

Con relación a la varianza se tiene:

$$\begin{aligned}
\frac{\partial^2 f(t)}{\partial \theta_i^2} &= \frac{\partial}{\partial \theta_i} \left\{ \frac{t - b'(\theta_i)}{a_i(\phi)} \exp\left[\frac{t\theta_i - b(\theta_i)}{a_i(\phi)} + c(t, \phi) \right] \right\} \\
&= \frac{\partial}{\partial \theta_i} \left\{ \frac{t - b'(\theta_i)}{a_i(\phi)} \right\} \exp\left[\frac{t\theta_i - b(\theta_i)}{a_i(\phi)} + c(t, \phi) \right] + \\
&\quad + \left(\frac{t - b'(\theta_i)}{a_i(\phi)} \right) \frac{\partial}{\partial \theta_i} \left\{ \exp\left[\frac{t\theta_i - b(\theta_i)}{a_i(\phi)} + c(t, \phi) \right] \right\} \\
&= -\left(\frac{b''(\theta_i)}{a_i(\phi)} \right) \exp\left[\frac{t\theta_i - b(\theta_i)}{a_i(\phi)} + c(t, \phi) \right] + \\
&\quad + \left(\frac{t - b'(\theta_i)}{a_i(\phi)} \right)^2 \exp\left[\frac{t\theta_i - b(\theta_i)}{a_i(\phi)} + c(t, \phi) \right] \\
&= -\left(\frac{b''(\theta_i)}{a_i(\phi)} \right) f(t) + \left(\frac{t - b'(\theta_i)}{a_i(\phi)} \right)^2 f(t)
\end{aligned}$$

Luego,

$$\begin{aligned}
\int_{Y_i} \frac{\partial^2 f(t)}{\partial \theta_i^2} \mathrm{d}t &= -\int_{Y_i} \left(\frac{b''(\theta_i)}{a_i(\phi)} \right) f(t) \mathrm{d}t + \int_{Y_i} \left(\frac{t - b'(\theta_i)}{a_i(\phi)} \right)^2 f(t) \mathrm{d}t \\
&= -\frac{b''(\theta_i)}{a_i(\phi)} \int_{Y_i} f(t) \mathrm{d}t + \frac{1}{a_i(\phi)^2} \int_{Y_i} (t - b'(\theta_i))^2 f(t) \mathrm{d}t \\
&= -\frac{b''(\theta_i)}{a_i(\phi)} + \frac{1}{a_i(\phi)^2} \int_{Y_i} (t - \mathrm{E}[Y_i])^2 f(t) \mathrm{d}t \\
&= -\frac{b''(\theta_i)}{a_i(\phi)} + \frac{1}{a_i(\phi)^2} \mathrm{V}[Y_i] = 0 \\
&\therefore \quad \mathrm{V}[Y_i] = b''(\theta_i) a_i(\phi)
\end{aligned}$$

□

Teorema A.5. *La distribución binomial con parámetro p_i, dado n_i constante fija conocida, esto es,*

$$f(y_i) = \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i-y_i}, y_i = 1, 2, ..., n_i \tag{13}$$

pertenece a la familia exponencial de distribuciones.

Demostración. La ecuación (13) puede también expresarse como sigue:

$$f(y_i) = \binom{n_i}{y_i} \left(\frac{p_i}{1-p_i}\right)^{y_i} (1-p_i)^{n_i}$$

Ahora, tomando logaritmos

$$\log f(y_i) = \log \binom{n_i}{y_i} + y_i \log\left(\frac{p_i}{1-p_i}\right) + n_i \log(1-p_i)$$

Exponenciando nuevamente,

$$f(y_i) = \exp\left\{\log \binom{n_i}{y_i} + y_i \log\left(\frac{p_i}{1-p_i}\right) + n_i \log(1-p_i)\right\}$$

Sea $\theta_i = \log[p_i/(1-p_i)] = \text{logit}(p_i)$, entonces

$$\begin{aligned} b(\theta_i) &= -n_i \log(1-p_i) = n_i \log\left(\frac{1}{1-p_i}\right) \\ &= n_i \log\left(\frac{1-p_i+p_i}{1-p_i}\right) = n_i \log\left(1+\frac{p_i}{1-p_i}\right) \\ &= n_i \log\left\{1+\exp\left[\log\left(\frac{p_i}{1-p_i}\right)\right]\right\} \\ &= n_i \log(1+e^{\theta_i}) \end{aligned}$$

Por otra parte, sea $a_i(\phi) = 1$ y $\phi = 1$, entonces $c(y_i, \phi) = \log \binom{n_i}{y_i}$. Luego, (13) pertenece a la familia exponencial de distribuciones (dado n_i fijo). También, $\mu_i = b'(\theta_i) = n_i p_i$ y $\sigma_i^2 = b''(\theta_i)a_i(\phi) = n_i p_i(1-p_i)$ como se demuestra a continuación:

$$
\begin{aligned}
\mu_i &= b'(\theta_i) = \frac{\partial}{\partial \theta_i}[n_i \log(1 + e^{\theta_i})] = \frac{n_i \exp(\theta_i)}{1 + \exp(\theta_i)} \\
&= \frac{n_i \exp(\log[p_i/(1-p_i)])}{1 + \exp(\log[p_i/(1-p_i)])} = \frac{n_i p_i/(1-p_i)}{1 + p_i/(1-p_i)} \\
&= \frac{n_i p_i/(1-p_i)}{(1-p_i+p_i)/(1-p_i)} = n_i p_i \\
\sigma_i^2 &= b''(\theta_i) a_i(\phi) = b''(\theta_i) = \frac{\partial}{\partial \theta_i}\left(\frac{n_i \exp(\theta_i)}{1 + \exp(\theta_i)}\right) \\
&= \frac{n_i \exp(\theta_i)[1 + \exp(\theta_i)] - n_i \exp(\theta_i) \exp(\theta_i)}{[1 + \exp(\theta_i)]^2} \\
&= \frac{n_i[p_i/(1-p_i)][1 + p_i/(1-p_i)] - n_i[p_i/(1-p_i)][p_i/(1-p_i)]}{[1 + p_i/(1-p_i)]^2} \\
&= \frac{[n_i p_i/(1-p_i)][1/(1-p_i)] - [n_i p_i/(1-p_i)][p_i/(1-p_i)]}{[1/(1-p_i)]^2} \\
&= \frac{n_i p_i/(1-p_i)^2 - n_i p_i^2/(1-p_i)^2}{1/(1-p_i)^2} = n_i p_i - n_i p_i^2 = n_i p_i (1-p_i)
\end{aligned}
$$

□

Teorema A.6. *La función de enlace canónico para la densidad binomial con media $\mu = np$ es logit(p).*

Demostración. El teorema A.5 demuestra que para una variable aleatoria binomial con parámetros n y p, el parámetro de enlace canónico $\theta = \text{logit}(p)$. Resta probar que $g(\mu) = \text{logit}(p)$ es efectivamente una función de enlace según la definición (ver Sección 2.2).

$$
\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \log\left(\frac{np}{n(1-p)}\right) = \log\left(\frac{\mu}{n-\mu}\right) = g(\mu)
$$

Y $g(\mu)$ es claramente una función uno a uno, continua y diferenciable de μ, previsto $\mu \neq n$. Luego $g(\cdot)$ es una función de enlace según la definición. □

Teorema A.7. *La razón de verosimilitud generalizada de un modelo lineal generalizado saturado M^* y cualquier otro modelo lineal generalizado propuesto M, se construye como el deviance dividido por ϕ para el caso de que $a_i(\phi) = \phi$.*

Demostración. En M^*, $\mu_i = y_i$ y sea θ_i^* el valor de θ_i en tal caso. Así, la función de verosimilitud L de M^* es

$$L_{M^*}(\mu, y) \;=\; L_{M^*}(y, y) = \prod_{i=1}^{n} \exp\left[\frac{y_i\theta_i^* - b(\theta_i^*)}{a_i(\phi)} + c(y_i, \phi)\right]$$

De forma similar, para M se tiene que

$$L_M(\mu, y) \;=\; \prod_{i=1}^{n} \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right]$$

La razón de verosimilitud es entonces

$$\lambda \;=\; \frac{\prod_{i=1}^{n} \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right]}{\prod_{i=1}^{n} \exp\left[\frac{y_i\theta_i^* - b(\theta_i^*)}{a_i(\phi)} + c(y_i, \phi)\right]}$$

Tomando logaritmos, resulta

$$\begin{aligned}
\log\lambda &= \log\left\{\frac{\prod_{i=1}^{n} \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right]}{\prod_{i=1}^{n} \exp\left[\frac{y_i\theta_i^* - b(\theta_i^*)}{a_i(\phi)} + c(y_i, \phi)\right]}\right\} \\
&= \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + \sum_{i=1}^{n} c(y_i, \phi) - \sum_{i=1}^{n} \frac{y_i\theta_i^* - b(\theta_i^*)}{a_i(\phi)} - \sum_{i=1}^{n} c(y_i, \phi) \\
&= \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} - \sum_{i=1}^{n} \frac{y_i\theta_i^* - b(\theta_i^*)}{a_i(\phi)} \\
&= \sum_{i=1}^{n} \frac{y_i(\theta_i - \theta_i^*) - b(\theta_i) + b(\theta_i^*)}{a_i(\phi)}
\end{aligned}$$

Multiplicando ambos miembros por -2, resulta

$$-2\log\lambda \;=\; 2\sum_{i=1}^{n} \frac{y_i(\theta_i^* - \theta_i) - b(\theta_i^*) + b(\theta_i)}{a_i(\phi)}$$

Suponiendo $a_i(\phi) = \phi$, la cantidad $D(\mu, y) = 2\sum_{i=1}^{n}[y_i(\theta_i^* - \theta_i) - b(\theta_i^*) + b(\theta_i)]$ es precisamente el *deviance*, y la razón de verosimilitud generalizada es $D(\mu, y)/\phi$. □