Revista Integración
Escuela de Matemáticas
Universidad Industrial de Santander
Vol. 32, No. 1, 2014, pág. 55–70

Prueba de homogeneidad de la dispersión para datos de proporción sobredispersos mediante regresión beta

Mario Morales[a,*], Jose Lozano[b]

[a] Universidad de Córdoba, Departamento de Matemáticas y Estadística, Montería, Colombia.

[b] Universidad de Antioquia, Sede de Investigación Universitaria SIU, Medellín, Colombia.

Resumen. En este artículo se propone un procedimiento para verificar la hipótesis de homogeneidad del parámetro de dispersión usando regresión beta, cuando se tienen datos de proporción sobredispersos. Se demuestra que es posible analizar este tipo de datos usando un modelo lineal generalizado usual ponderado, con pesos obtenidos mediante la regresión beta. Esta forma de proceder permite corregir el problema de la dispersión extra, manteniendo la sencillez del análisis.

Además, para algunos casos particulares, se evalúa mediante un estudio de simulación, la potencia de la prueba.

Palabras claves: Sobredispersión, datos de proporción, regresión beta, razón de verosimilitud, modelos lineales generalizados.
MSC2010: 62-07, 62J05, 62J12.

Test for homogeneity of the dispersion for overdispersed proportions data through beta regression

Abstract. In this paper we propose an approach to validate the hypothesis of homogeneity of the dispersion parameter using beta regression, when we have overdispersed proportions data. We corroborated that it is possible to analyze this type of data with an usual weighted generalized linear model, weighting the observations with weights obtained through beta regression. This procedure allows to correct the problem of overdispersion keeping the simplicity of the analysis.

Furthermore, for several cases, we made a simulation study of the power of the test.

Keywords: Overdispersion, proportion data, beta regression, likelihood ratio, generalized linear models.

* Autor para correspondencia: *E-mail*: `mamorales@correo.unicordoba.edu.co`.
Recibido: 30 de septiembre de 2013, Aceptado: 28 de marzo de 2014.
Para citar este artículo: M. Morales, J. Lozano, Prueba de homogeneidad de la dispersión para datos de proporción sobredispersos mediante regresión beta, *Rev. Integr. Temas Mat.* 32 (2014), no. 1, 55–70.

1. Introducción

En el análisis de datos de proporción es común que haya presencia de sobredispersión, situación que se presenta cuando la varianza exhibida por los datos es mucho más grande que la que predice el modelo. En la literatura usualmente se aborda el problema suponiendo un parámetro de dispersión común, como puede verse en Crowder [6] y Brooks [3]. Sin embargo es usual que la variabilidad exhibida por los datos dependa de los valores de alguna variable explicativa, debiendo ser incorporada esta situación en el modelamiento (Morales & López [16]).

Si bien es cierto que existen aplicaciones donde el supuesto de homogeneidad de la dispersión se satisface, como ocurre con los datos de germinación de las semillas de Orobanche Cernua cultivadas en tres medios, analizados por Crowder [6], existen situaciones donde, de acuerdo con la naturaleza de los datos, es posible que se justifique suponer heterogeneidad en el parámetro de dispersión. Por ejemplo, Crowder [7], refiriéndose al análisis de un conjunto de datos que reportan la presencia o ausencia de tumores en ratones expuestos a un cancerígeno, comenta que el coeficiente de correlación entre respuestas binarias está asociado al efecto de la herencia sobre la respuesta, y demuestra que, para ese caso, es apropiado suponer un coeficiente de correlación distinto para cada nivel de exposición al cancerígeno; en el citado ejemplo, la falta de independencia es la causante de la sobredispersión y por tanto se justifica el supuesto de heterogeneidad.

Morales & López [16] proponen una forma de juzgar la hipótesis (13), asumiendo la distribución betabinomial como el modelo probabilístico que genera los datos, y bajo el supuesto de que la sobredispersión se explique por el efecto de los niveles del factor en un experimento a una vía de clasificación. En este artículo se propone un método para contrastar la hipótesis de homogeneidad de la dispersión, suponiendo la distribución de probabilidad beta para la respuesta de proporción sobredispersa, sin restricción en cuanto a las variables que explican la dispersión extra.

2. Sobredispersión

Si la especificación de un Modelo Lineal Generalizado (MLG) es correcta en términos del componente aleatorio, la componente sistemática y la función de enlace, esperamos que el desvío tenga aproximadamente una distribución ji–cuadrado con los grados de libertad apropiados.

En muchas situaciones prácticas, cuando se ajusta un MLG a un conjunto de datos, se puede observar un desvío mucho más grande que el esperado si el modelo fuera correcto. La explicación más común a este fenómeno es que se tiene una componente sistemática incorrecta para el modelo, es decir, no se han incluido los predictores apropiados; no se ha hecho la transformación o combinación apropiada de efectos; o no se han incluido todos los factores de ajuste en el modelo. Otra explicación común para un desvío grande es la presencia de un número pequeño de datos atípicos (*outliers*). Es excepcional que se identifique un número grande de puntos atípicos, si eso ocurre se debe aceptar como válida una distribución inapropiada para el error. Los tamaños de muestra pequeños también conducen a desvíos grandes.

Habiendo excluido todas las posibilidades anteriores, es posible que la variación de los

datos sea mucho más grande que la predicha por el modelo. A este fenómeno se lo conoce como *sobredispersión* (Hinde & Demétrio [11]).

Las principales consecuencias de no tener en cuenta la sobredispersión son: subestimación de los errores estándar de las estimaciones de los parámetros, y por tanto su significación puede juzgarse de manera incorrecta; los cambios en el desvío asociados a los términos del modelo serían grandes y pueden conducir a la selección de un modelo demasiado complejo, lo cual lleva a que la interpretación del modelo sea incorrecta y las predicciones demasiado imprecisas (Hinde & Demétrio [11]).

Existen muchos modelos diferentes para ajustar datos en presencia de sobredispersión, los cuales surgen de los supuestos que se toman para explicar este fenómeno. Los modelos se pueden clasificar en dos grandes grupos; Modelos de media varianza y Modelos en dos etapas (Hinde & Demétrio [11]).

3. Distribución beta

Se dice que una variable aleatoria Y sigue una distribución beta con parámetros $p, q > 0$, denotado por $B\,(p, q)$, si su función de densidad de probabilidad está dada por

$$f\,(y; p, q) \quad = \quad \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1},\ 0 < y < 1,$$

donde $\Gamma(\cdot)$ es la función gamma. La media y la varianza de Y, de acuerdo con Mood et al [14] son, respectivamente,

$$\mathrm{E}\,(Y) = \frac{p}{p+q} \qquad \text{y} \qquad \mathrm{Var}\,(Y) = \frac{pq}{(p+q)^2\,(p+q+1)}. \tag{1}$$

Con el fin de obtener una estructura de regresión para la media de la respuesta junto con el parámetro de precisión, se trabaja con una parametrización diferente de la densidad beta. Siguiendo a Ferrari et al [9], sea $\mu = \dfrac{p}{p+q}$ y $\delta = p + q$, es decir, $p = \mu\delta$ y $q = (1-\mu)\,\delta$. Con este cambio de variable, se deduce de la ecuación (1) que

$$\mathrm{E}(Y) = \mu \qquad \text{y} \qquad \mathrm{Var}(Y) = \frac{\mu(1-\mu)}{1+\delta}; \tag{2}$$

μ es la media de la variable respuesta y δ puede ser interpretado como un parámetro de precisión, en el sentido de que, para un μ fijo, un valor grande δ implica un menor valor de la varianza de Y. Para efectos de este trabajo, $\phi = \delta^{-1}$ es un parámetro de dispersión. Bajo esta parametrización, usaremos la notación $Y \sim B\,(\mu, \delta)$. La densidad de Y se puede reescribir como

$$f(y; \mu, \delta) \quad = \quad \frac{\Gamma\,(\delta)}{\Gamma\,(\mu\delta)\,\Gamma\,((1-\mu)\,\delta)} y^{\mu\delta-1}\,(1-y)^{(1-\mu)\delta-1}\,,\ 0 < y < 1, \tag{3}$$

donde $0 < \mu < 1$ y $\delta > 0$.

Por otra parte, si X es una variable aleatoria con distribución binomial de parámetros n y π, entonces la proporción de éxitos es $P = X/n$, y se verifica que

$$\mathrm{E}(P) = \pi \qquad \text{y} \qquad \mathrm{Var}(P) = \frac{\pi(1-\pi)}{n}. \tag{4}$$

Si se iguala el parámetro μ de la distribución beta con el parámetro π de una variable aleatoria binomial, a la luz de las ecuaciones (2) y (4) es claro que

$$\mathrm{Var}(Y) = \frac{n}{1+\delta}\,\mathrm{Var}(P), \tag{5}$$

es decir, la varianza de la distribución beta es un múltiplo de la varianza de la proporción muestral. Este hecho permite usar la distribución beta para modelar datos de proporción con sobredispersión, ya que la constante $n/(1+\delta)$ permite modelar la dispersión extra presente en los datos.

Modelo de regresión beta

Sea $Y_1, \ldots, Y_n$ una muestra aleatoria, tal que $Y_i \sim B(\mu_i, \delta)$, $i = 1, \ldots, n$, es decir, Y_i sigue la densidad dada en (3), con media μ_i y parámetro de precisión δ desconocidos. El modelo de regresión beta supone que la media de Y_i se puede escribir como

$$g(\mu_i) = \boldsymbol{x}_i^t \boldsymbol{\beta} = \eta_i, \tag{6}$$

donde $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^t$ es un vector de parámetros desconocidos; $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})^t$ es el vector de k variables regresoras ($k < n$), las cuales se asumen no aleatorias y conocidas; $\eta_i = \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$ es el predictor lineal; por lo general $x_{i1} = 1$ para todo i, de manera que el modelo tiene una intersección. Por último, $g(\cdot) : (0,1) \to \mathbb{R}$ es una función de enlace estrictamente monótona y doblemente diferenciable.

Se supondrá que la respuesta está limitada al intervalo unitario $(0,1)$. Sin embargo, el modelo sigue siendo útil para situaciones donde la respuesta es restringida al intervalo (a,b), donde a y b son escalares conocidos, con $a < b$. En ese caso se modela $\dfrac{(Y_i - a)}{b-a}$ en lugar de Y_i directamente. Además, si Y_i también toma los valores extremos 0 y 1, una transformación útil en la práctica es

$$Y^* = \frac{Y(n-1) + 0{,}5}{n}, \tag{7}$$

donde n es el tamaño de la muestra (Smithson & Verkuilen [22]).

Una extensión de la anterior propuesta empleada por Smithson & Verkuilen [22] y formalmente introducida (junto con otras extensiones) por Simas et al [21], es la del modelo de regresión beta con dispersión variable. En este modelo el parámetro de precisión no es constante para todas las observaciones, sino que se modela en forma similar a la media. Concretamente, si $Y_1, \ldots, Y_n$ son variables aleatorias independientes tales que $Y_i \sim B(\mu_i, \delta_i)$, $i = 1, \ldots, n$, los modelos se definen como

$$g_1(\mu_i) = \eta_{1i} = f_1\left(\boldsymbol{x}_i^t; \boldsymbol{\beta}\right), \tag{8}$$

$$g_2(\delta_i) = \eta_{2i} = f_2\left(\boldsymbol{z}_i^t; \boldsymbol{\theta}\right), \tag{9}$$

donde, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^t$ y $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_h)^t$ son vectores correspondientes al conjunto de parámetros que se suponen funcionalmente independientes; $k + h < n$. η_{1i} y η_{2i} son los predictores, no necesariamente lineales; $\boldsymbol{x}_i^t = (x_{i1}, \ldots, x_{1q_1})$, $\boldsymbol{z}_i^t = (z_{i1}, \ldots, z_{1q_2})$ son, respectivamente, vectores de q_1 y q_2 covariables conocidas, que no necesitan ser

excluyentes. De hecho, es muy común elegir $\boldsymbol{z}_i^t$ como un subconjunto de $\boldsymbol{x}_i^t$. Las funciones $f_1(\cdot)$ y $f_2(\cdot)$ representan funciones de los parámetros y las variables explicativas; así por ejemplo, en el caso lineal

$$f_1\left(\boldsymbol{x}_i^t;\boldsymbol{\beta}\right) = \boldsymbol{x}_i^t\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \tag{10}$$

Sean $\eta_1 = (\eta_{11},\ldots,\eta_{1n})^t$ y $\eta_2 = (\eta_{21},\ldots,\eta_{2n})^t$. Se supondrá que las matrices de derivadas $\widetilde{\boldsymbol{X}} = \dfrac{\partial\eta_1}{\partial\boldsymbol{\beta}}$ y $\widetilde{\boldsymbol{Z}} = \dfrac{\partial\eta_2}{\partial\boldsymbol{\theta}}$ tienen rango k y h respectivamente (ecuaciones (15), (16) y (18)), y que las funciones de enlace $g_1 : (0,1) \to \mathbb{R}$ y $g_2 : (0,\infty) \to \mathbb{R}$ son estrictamente monótonas y doblemente diferenciables.

Se pueden usar diferentes funciones de enlace: para $g_1(\cdot)$ se tiene la especificación logit, las funciones probit y log–log complementaria entre otras, las cuales también son aplicables al modelo (6) con $g(\mu) = g_1(\mu)$. Para $g_2(\cdot)$ se tienen las funciones logaritmo, raíz cuadrada y la identidad entre otras. En la Tabla 1 se muestran estas funciones; $\Phi^{-1}(\cdot)$ denota la función de distribución normal estándar inversa y $\delta > 0$. Una rica discusión de las funciones de enlace son presentadas por Atkison [1] y McCullagh & Nelder [13].

$g_1(\cdot)$	
logit	$g_1(\mu) = \log\left(\dfrac{\mu}{1-\mu}\right)$
probit	$g_1(\mu) = \Phi^{-1}(\mu)$
log–log complementaria	$g_1(\mu) = \log\{-\log(1-\mu)\}$
$g_2(\cdot)$	
logaritmo	$g_2(\delta) = \log(\delta)$
raíz cuadrada	$g_2(\delta) = \sqrt{\delta}$
identidad	$g_2(\delta) = \delta$

Tabla 1. Funciones de enlace.

La función de log–verosimilitud para esta clase de modelos de regresión beta tiene la forma

$$\ell(\boldsymbol{\beta},\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell_i(\mu_i,\delta_i), \tag{11}$$

donde

$$\begin{aligned}\ell_i(\mu_i,\delta_i) &= \log\Gamma(\delta_i) - \log\Gamma(\mu_i\delta_i) - \log\Gamma((1-\mu_i)\delta_i) + (\mu_i\delta_i - 1)\log(y_i)\\ &\quad + \{(1-\mu_i)\delta_i - 1\}\log(1-y_i);\end{aligned} \tag{12}$$

$\mu_i = g_1^{-1}(\eta_{1i})$ y $\delta_i = g_2^{-1}(\eta_{2i})$, definidos como se indica en (8) y (9), son funciones de $\boldsymbol{\beta}$ y $\boldsymbol{\theta}$ respectivamente. Las estimaciones por máxima verosimilitud de $\boldsymbol{\beta}$ y $\boldsymbol{\theta}$ se obtienen resolviendo el sistema no lineal $\boldsymbol{U}(\boldsymbol{\varsigma}) = \boldsymbol{0}$, donde $\boldsymbol{\varsigma} = \left(\boldsymbol{\beta}^t,\boldsymbol{\theta}^t\right)^t$ es el vector de $k+h$ parámetros. En la práctica estas estimaciones se pueden obtener mediante una maximización

numérica de la función de log–verosimilitud usando un algoritmo de optimización no lineal (Newton-Raphson, Fisher's scoring, quasi–Newton, BFGS (Broydon-Fletcher-Goldfarb-Shanno), etc.). Más detalles al respecto pueden verse en Press et al [18], Nocedal & Wright [12] y Gordon [10] .

Es posible mostrar que este modelo de regresión beta es regular, en el sentido de que se mantienen todas las condiciones de regularidad descritas en Cox & Hinkley [4, p. 107]; también es posible mostrar que los estimadores de máxima verosimilitud son únicos (véase Simas et al [21]).

4. Metodología para contrastar la hipótesis

Es de crucial importancia, al momento del análisis estadístico de datos de proporción sobredispersos, poder contrastar las hipótesis

$$\begin{aligned} H_0 &: \phi_1 = \cdots = \phi_h = \phi, \\ H_1 &: \phi_i \neq \phi_j \;\; i \neq j, \;\; i,j = 1,\ldots,h, \end{aligned} \tag{13}$$

en donde $\phi = \delta^{-1}$ es el parámetro de dispersión, δ se ha definido en la ecuación (2) y h es el número de grupos sugerido por la naturaleza de los datos.

Para llevar a cabo el contraste de las hipótesis dadas en (13) se usará la prueba de la razón de verosimilitudes, es decir, compararemos los logaritmos de las verosimilitudes de dos modelos, a saber: un modelo que resulta bajo el supuesto de que la hipótesis alternativa H_1 es cierta, que llamaremos ***modelo alternativo***, y otro modelo que resulta de suponer que la hipótesis nula H_0 es cierta, el cual llamaremos ***modelo nulo***; a continuación se describen estos modelos.

Modelo alternativo

El modelo bajo el supuesto que la hipótesis alternativa es verdadera supone que el parámetro de precisión δ varía a través de una estructura de regresión lineal. Más precisamente, las ecuaciones (8) y (9) toman la forma

$$\begin{aligned} g_1(\mu_i) &= \eta_{1i} = \boldsymbol{x}_i^t \boldsymbol{\beta}, \\ g_2(\delta_i) &= \eta_{2i} = \boldsymbol{z}_i^t \boldsymbol{\theta}, \end{aligned} \tag{14}$$

con $\boldsymbol{\beta} \in \mathbb{R}^k$, $\boldsymbol{\theta} \in \mathbb{R}^h$ vectores de parámetros, $k + h < n$; $\boldsymbol{x}_i^t$, $\boldsymbol{z}_i^t$ vectores de observaciones conocidas, $\boldsymbol{x}_i \in \mathbb{R}^{q_1}$ y $\boldsymbol{z}_i \in \mathbb{R}^{q_2}$ para todo $i = 1, 2 \ldots, n$; η_{1i} y η_{2i} predictores lineales, $g_1 : (0,1) \to \mathbb{R}$ y $g_2 : (0,\infty) \to \mathbb{R}$ funciones de enlace estrictamente monótonas y doblemente diferenciables, tal como se definió en la sección 3. Para este modelo la función de log–verosimilitud toma las formas (11) y (12).

Recordemos que $\eta_1 = (\eta_{11}, \ldots, \eta_{1n})^t$ y $\eta_2 = (\eta_{21}, \ldots, \eta_{2n})^t$; además, $\eta_{1i} = \boldsymbol{x}_i^t \boldsymbol{\beta}$ y $\eta_{2i} = \boldsymbol{z}_i^t \boldsymbol{\theta}$ para todo i, de manera que $\dfrac{\partial \eta_{1i}}{\partial \boldsymbol{\beta}} = \boldsymbol{x}_i^t$ y $\dfrac{\partial \eta_{2i}}{\partial \boldsymbol{\theta}} = \boldsymbol{z}_i^t$ para todo i. Entonces

$$\widetilde{\boldsymbol{X}} = \frac{\partial \eta_1}{\partial \boldsymbol{\beta}} = \left(\frac{\partial \eta_{11}}{\partial \boldsymbol{\beta}}, \ldots, \frac{\partial \eta_{1n}}{\partial \boldsymbol{\beta}} \right)^t = \left(\boldsymbol{x}_i^t, \ldots, \boldsymbol{x}_n^t \right)^t = \boldsymbol{X} \tag{15}$$

y

$$\widetilde{\boldsymbol{Z}} = \frac{\partial \eta_2}{\partial \boldsymbol{\theta}} = \left(\frac{\partial \eta_{21}}{\partial \boldsymbol{\theta}}, \ldots, \frac{\partial \eta_{2n}}{\partial \boldsymbol{\theta}} \right)^t = \left(\boldsymbol{z}_i^t, \ldots, \boldsymbol{z}_n^t \right)^t = \boldsymbol{Z}, \tag{16}$$

donde $\boldsymbol{X}$ y $\boldsymbol{Z}$ son las matrices de covariables de dimensiones $n \times k$ y $n \times h$, de rangos k y h respectivamente, con filas dadas por $\boldsymbol{x}_i^t$ y $\boldsymbol{z}_i^t$ respectivamente.

Modelo nulo

El modelo bajo el supuesto que la hipótesis nula es cierta supone que el parámetro de precisión δ es constante, bajo este supuesto se tiene de (8) y (9),

$$\begin{aligned} g_1(\mu_i) &= \eta_{1i} = \boldsymbol{x}_i^t \boldsymbol{\beta}, \\ g_2(\delta_i) &= \eta_{2i} = \theta_0, \end{aligned} \tag{17}$$

donde $\theta_0 \in \mathbb{R}$ es el único parámetro a estimar en lo referente al modelo para la dispersión. La función de log–verosimilitud está dada en la forma de (11) y (12) reemplazando δ_i por δ. En este caso, la matriz de covariables $\widetilde{\boldsymbol{X}}$ tiene la forma dada en (15), con filas $\boldsymbol{x}_i^t$. Además, $\dfrac{\partial \eta_{2i}}{\partial \delta} = 1$ para todo i, de manera que

$$\widetilde{\boldsymbol{Z}} = \frac{\partial \eta_2}{\partial \boldsymbol{\theta}} = \left(\frac{\partial \eta_{21}}{\partial \delta}, \ldots, \frac{\partial \eta_{2n}}{\partial \delta} \right)^t = (1, \ldots, 1)^t = \mathbf{1}, \tag{18}$$

con $\mathbf{1} \in \mathbb{R}^h$; es decir, $\widetilde{\boldsymbol{Z}}$ es un vector de dimensión $h \times 1$ cuyas componentes son todas iguales a uno.

El modelo alternativo tiene $k + h$ parámetros a estimar (k en el modelo de medias y h en el modelo de precisión), mientras que el modelo nulo tiene $k + 1$ parámetros; así, la diferencia en el número de parámetros de los dos modelos es $h - 1$.

Para decidir si se rechaza la hipótesis nula H_0 de (13) en favor de la hipótesis alternativa H_1, se siguen los siguientes pasos:

(i) se ajustan los modelos, estimando los parámetros por el método de máxima verosimilitud y se obtienen las log–verosimilitudes;

(ii) se calcula la razón de verosimilitudes λ,

$$\lambda = \frac{L\left(\widehat{\boldsymbol{\theta}}_0; \boldsymbol{y}\right)}{L\left(\widehat{\boldsymbol{\theta}}_1; \boldsymbol{y}\right)}, \tag{19}$$

donde $L\left(\widehat{\boldsymbol{\theta}}_0; \boldsymbol{y}\right)$ y $L\left(\widehat{\boldsymbol{\theta}}_1; \boldsymbol{y}\right)$ son las verosimilitudes maximizadas bajo los modelos nulo y alternativo, respectivamente. La variable aleatoria λ, obtenida mediante la ecuación (19) converge en distribución a una ji–cuadrado con $h - 1$ grados de libertad (Dobson [8]); (iii) se rechaza la hipótesis nula en favor de la alternativa a un nivel de significación α, o equivalentemente, se concluye que el modelo alternativo ajusta significativamente mejor que el modelo nulo, si $P\left(\chi^2_{h-1} > -2\ln(\lambda)\right) < \alpha$, donde χ^2_{h-1} representa una variable aleatoria con distribución ji–cuadrado con $h - 1$ grados de libertad.

4.1. Ejemplos

A continuación se presentan dos ejemplos que ilustran el procedimiento para contrastar la hipótesis (13) descrito en esta sección.

Ejemplo 4.1. Se sometieron 58 ratas hembras a dietas deficientes en hierro, divididas en cuatro grupos. A un grupo de control se le proporcionaron semanalmente inyecciones de suplemento de hierro para mantener su ingestión de hierro en niveles normales, mientras que a otro grupo se le proporcionaron solo inyecciones de un placebo. A los otros dos grupos se les proporcionaron menos inyecciones de suplemento de hierro que a los controles. Las ratas fueron preñadas y sacrificadas tres semanas después; se registró el número total de fetos y el número de fetos muertos en cada camada. Los datos, que se muestran en la Tabla 2, fueron tomados de Moore & Tsiatis [15] y se encuentran en la librería VGAM del paquete Estadístico R [19] bajo el nombre de **lirat**.

	Grupo no tratado (bajo en hierro)											
N	10	11	12	4	10	11	9	11	10	10	12	10
R	1	4	9	4	10	9	9	11	10	7	12	9
N	8	11	6	9	14	12	11	13	14	10	12	13
R	8	9	4	7	14	7	9	8	5	10	10	8
N	10	14	13	4	8	13	12					
R	10	3	13	3	8	5	12					
	Grupo 2: inyecciones de hierro sólo en el día 7 o 10											
N	10	3	13	12	14	9	13	16	11	4	1	12
R	1	1	1	0	4	2	2	1	0	0	0	0
	Grupo 3: inyecciones de hierro en el día 0 y 7											
N	8	11	14	14	11							
R	0	1	0	1	0							
	Grupo 4: inyecciones de hierro semanalmente											
N	3	13	9	17	15	2	14	8	6	17		
R	0	0	2	2	0	0	1	0	0	0		

Tabla 2. Número de fetos (N) y números de fetos muertos (R).

Realizando el ajuste de los datos a un MLG usual usando la distribución binomial y la función logit $(\cdot)$ como enlace, se obtienen las siguientes estimaciones para las proporciones de éxito en cada tratamiento: $\widehat{\pi}_1 = 0{,}758$, $\widehat{\pi}_2 = 0{,}102$, $\widehat{\pi}_3 = 0{,}034$ y $\widehat{\pi}_4 = 0{,}048$, con errores estándar de estimación 0,1292, 0,3046, 0,7196 y 0,4584, respectivamente. Se observa un desvío residual (*residual deviance*) de 173,45 con 54 grados de libertad, es decir, se tiene un desvío tres veces mayor que los grados de libertad, lo cual evidencia sobredispersión.

Para modelar la proporción de ratas muertas por medio de regresión beta, debido a que estos valores están en el intervalo $[0, 1]$, se realiza la trasformación (7). En este caso el

modelo alternativo considera el parámetro δ como función de los grupos; en el **modelo nulo** se asume que el parámetro de precisión es constante.

El ajuste del modelo alternativo (14) produce $\widehat{\pi}_1 = 0{,}770$, $\widehat{\pi}_2 = 0{,}108$, $\widehat{\pi}_3 = 0{,}040$, $\widehat{\pi}_4 = 0{,}050$, con errores estándar de estimación 0,2377, 0,4203, 0,5243 y 0,4600 respectivamente. Las estimaciones para los parámetros δ son: $\widehat{\delta}_1 = 1{,}945$, $\widehat{\delta}_2 = 6{,}090$, $\widehat{\delta}_3 = 22{,}566$, $\widehat{\delta}_4 = 12{,}565$; el valor de la log–verosimilitud a partir de este modelo con 8 parámetros es 68,73977. El ajuste del modelo nulo, dado por la ecuación (17), produce las siguientes estimaciones: $\widehat{\pi}_1 = 0{,}802$, $\widehat{\pi}_2 = 0{,}156$, $\widehat{\pi}_3 = 0{,}120$, $\widehat{\pi}_4 = 0{,}117$ con errores estándar de estimación 0,2102, 0,4134, 0,5539 y 0,4452 respectivamente. La estimación para el parámetro δ es $\widehat{\delta} = 2{,}802$. El valor de la log–verosimilitud para este modelo es 61,499 con 5 parámetros.

Así que el valor del estadístico para contrastar las hipótesis dadas por (13), es $-2 \times (61{,}499 - 68{,}73977) = 14{,}482$ con $8 - 5 = 3$ grados de libertad, y $P(\chi^2_3 > 14{,}482) = 0{,}002317$ lo que nos conduce a rechazar la hipótesis nula con un nivel de significación α menor que 0,01 (1 %). En conclusión, hay una fuerte evidencia que la dispersión de la respuesta no es homogénea, varía con los grupos y por tanto en el análisis se debe considerar un parámetro de dispersión ϕ diferente para cada grupo. Este resultado está de acuerdo con lo obtenido por Morales & López [16]. Las estimaciones para las proporciones obtenidas ajustando un MLG usual son bastantes similares a las obtenidas ajustando el modelo de regresión beta; sin embargo, los errores estándar estimados difieren. Esto es así porque el modelo de regresión beta supone un parámetro de dispersión distinto para cada grupo y los errores estándar se corrigen por ese factor, mientras que el MLG usual supone que el parámetro de dispersión es igual a uno ($\phi = 1$) para todos los grupos.

Ejemplo 4.2. Una de las principales preocupaciones en la industria de alimentos es la conservación de los productos en buen estado para el consumo humano. El ñame (*dioscorea alata*), como cualquier otro tipo de alimento, sufre un proceso de descomposición que ocasiona, después de transcurrido cierto tiempo, la imposibilidad de su consumo. Existen ciertas características que evidencian la descomposición de este producto, entre ellas el color, que se torna más oscuro al pasar el tiempo. Teniendo en cuenta esta característica, se realizó un estudio en la Universidad de Córdoba, sede Berástegui, que consiste en la evaluación (pruebas afectivas) del color del ñame por treinta jueces cada cinco días durante 20 días. Para efectos de este ejemplo usaremos los datos del día 15. Estamos interesados en modelar la proporción de calificaciones buenas, $y/30$, donde y es la cantidad de jueces que otorgaron una calificación por encima de seis al color del ñame en cada uno de los días del estudio, en función de dos niveles de temperatura de cocción. Los datos fueron proporcionados por Barrios & Peña [2], y se muestran en la Tabla 3.

Temperatura 1								
4	2	2	5	10	7	4	7	5
3	3	1	3	2	18	6	5	4
Temperatura 2								
0	0	12	0	0	0	8	0	0
0	0	2	3	0	4	3	0	0

Tabla 3. Evaluación del color del ñame en el día 15.

Al ajustar un MLG usual usando la distribución binomial y la función logit $(\cdot)$ como enlace, se tiene $\widehat{\pi}_1 = 0{,}169$ y $\widehat{\pi}_1 = 0{,}059$ con errores estándar estimados 0,115 y 0,216; el desvío residual es 143,35 con 34 grados de libertad, es decir, se tiene un desvío cuatro veces mayor que los grados de libertad, lo cual evidencia sobredispersión.

Los valores de la variable respuesta (proporción) están en el intervalo $[0, 1)$. Con el fin de ajustar un modelo de regresión beta se realiza la trasformación dada por (7), donde $n = 30$ es el número de jueces. En este caso bajo el modelo alternativo se asume que el parámetro de dispersión varía con el nivel de temperatura; por lo tanto, se considera el modelo lineal de regresión beta con el parámetro de precisión δ_i en función de los niveles de `temperatura`. El modelo nulo asume que el parámetro de precisión es el mismo para ambos niveles.

Del ajuste del modelo alternativo se tiene $\widehat{\pi}_1 = 0{,}184$ y $\widehat{\pi}_2 = 0{,}077$ con errores estándar estimados 0,1678 y 0,3204 respectivamente. Se estima, para la dispersión $\widehat{\delta}_1 = 12{,}087$ y $\widehat{\delta}_2 = 9{,}448$, con errores estándar estimados 0,3291 y 0,4948. El logaritmo de la verosimilitud es 45,006 con 4 grados de libertad. Del modelo nulo se obtiene $\widehat{\pi}_1 = 0{,}187$ y $\widehat{\pi}_2 = 0{,}073$, con errores estándar estimados 0,1711 y 0,2720 respectivamente. Se estima, para dispersión $\widehat{\delta} = 10{,}755$, con error estándar estimado 0,245. El logaritmo de la verosimilitud es 44,881 con 3 grados de libertad. El valor del estadístico de razón de verosimilitud obtenido es $-2 \times (44{,}881 - 45{,}006) = 0{,}2501$ con un grado de libertad, $p-$valor 0,617; por lo tanto, no hay suficiente evidencia en los datos para rechazar la hipótesis de homogeneidad del parámetro de dispersión, es decir, debemos adoptar un parámetro común ϕ para todas las observaciones.

5. Modelo lineal generalizado ponderado para ajustar datos con sobredispersión

De la ecuación (5) se deduce que, para datos de proporción sobre dispersos, el término $n/(1 + \delta)$ es el responsable de la dispersión extra. Por otra parte, el procedimiento de prueba de la hipótesis (13) exige la estimación por máxima verosimilitud de los parámetros de los modelos. En particular, se obtienen estimaciones por máxima verosimilitud de δ. Esta sección se dedica a demostrar que es posible usar estas estimaciones, de tal forma que el análisis de los datos de proporción con sobredispersión se puede llevar a cabo ajustando un modelo lineal generalizado usual, pero ponderando las observaciones usando los pesos

$$w_i = 1 + \frac{n}{1 + \delta_i} \tag{20}$$

si se rechaza la hipótesis (13), $i = 1, \ldots, h$; es decir, se obtienen los pesos para las observaciones en cada grupo; o en caso contrario,

$$w_i = 1 + \frac{n}{1 + \delta} \tag{21}$$

para $i = 1, \ldots, n$ (δ es el mismo para todos los grupos).

La estimación por máxima verosimilitud de los parámetros de un modelo lineal generalizado es, de acuerdo con Nelder & Wedderburn [17], equivalente a un proceso de Mínimos

Cuadrados Ponderados Iterativos (MCPI) con función de pesos

$$p_i = \frac{1}{V(\mu_i)} \left(\frac{d\mu_i}{d\eta_i} \right)^2 \tag{22}$$

y una *variable dependiente modificada*

$$Z_i = \eta_i + (Y_i - \mu_i) \left(\frac{d\eta_i}{d\mu_i} \right). \tag{23}$$

Los pesos p_i dados por la ecuación (22) corresponden al inverso de la varianza de Z_i; en efecto,

$$\begin{aligned} \mathrm{Var}(Z_i) &= \mathrm{Var}\left[(Y_i - \mu_i) \left(\frac{d\eta_i}{d\mu_i} \right) \right] \\ &= \mathrm{Var}(Y_i) \left(\frac{d\eta_i}{d\mu_i} \right)^2 \\ &= V(\mu_i) \left(\frac{d\eta_i}{d\mu_i} \right)^2 \\ &= \frac{1}{p_i}. \end{aligned} \tag{24}$$

La estimación por mínimos cuadrados ponderados implica la solución del sistema de ecuaciones (Dobson [8])

$$\boldsymbol{X}^t \boldsymbol{P} \boldsymbol{X} \boldsymbol{\beta} = \boldsymbol{X}^t \boldsymbol{P} \boldsymbol{z}, \tag{25}$$

donde $\boldsymbol{P} = \mathrm{diag}\{p_i\}$, $\boldsymbol{X}$ es la matriz de diseño y $\boldsymbol{z}$ es un vector cuyos elementos se obtienen mediante la ecuación (23). Resolver el sistema de ecuaciones (25) es equivalente a resolver el sistema

$$\boldsymbol{A}^t \boldsymbol{A} \boldsymbol{\beta} = \boldsymbol{A}^t \boldsymbol{z}^*, \tag{26}$$

donde $\boldsymbol{A} = \boldsymbol{P}^{\frac{1}{2}} \boldsymbol{X}$ para $\boldsymbol{P}^{\frac{1}{2}} = \mathrm{diag}\{\sqrt{p_i}\}$ y $\boldsymbol{z}^* = \boldsymbol{P}^{\frac{1}{2}} \boldsymbol{z}$. El $i-$ésimo elemento de $\boldsymbol{z}^*$ es de la forma $Z_i^* = \sqrt{p_i} Z_i$, y $\mathrm{Var}(Z_i^*) = p_i \, \mathrm{Var}(Z_i) = 1$ por (24); sin embargo, si la variable respuesta tiene una función de varianza de la forma $\mathrm{Var}(Y_i) = w_i V(\mu_i)$, entonces, del desarrollo de (24) es claro que $\mathrm{Var}(Z_i) = w_i / p_i$, y por tanto $\mathrm{Var}(Z_i^*) = p_i \, \mathrm{Var}(Z_i) = p_i \frac{w_i}{p_i} = w_i$

Se ha demostrado que con una función de varianza de la forma $w_i V(\mu_i)$ la varianza de la variable modificada (23) es proporcional a w_i, factor que es el responsable de la varianza extra, lo que sugiere, por analogía con el modelo lineal clásico (Ravishanker & Dey [20]), que en el ajuste del modelo (26) se usen los valores $1/w_i$ como ponderaciones, de tal forma que la variable transformada Z_i^* quede con varianza 1. Usar pesos $1/w_i$ en la solución de las ecuaciones (26) es equivalente a usar $\boldsymbol{P} = \mathrm{diag}(p_i/w_i)$ como matriz de ponderaciones en el sistema (25) del proceso de MCPI.

Así que con datos de proporción con sobredispersión, el razonamiento anterior conduce a que el proceso MCPI se realice con pesos $p_i^* = p_i/w_i$ donde p_i son los pesos que se calculan, para cada observación, mediante la ecuación (22) y que resultan del modelo binomial usual.

Los parámetros δ_i de las ecuaciones (20) y (21), y por tanto los pesos w_i, no se conocen, pero después del proceso de prueba de la hipótesis de interés (13), se tienen las estimaciones por máxima verosimilitud que se usarían en su lugar. La mayoría de los paquetes para el cálculo estadístico permiten la inclusión de ponderaciones en el ajuste de un modelo lineal generalizado.

Ejemplo 5.1. Aplicando el procedimiento descrito anteriormente a los Ejemplos 4.1 y 4.2, se obtiene los siguientes resultados:

A partir de los datos del Ejemplo 4.1, después de ajustar el MLG usual usando los pesos dados por la ecuación (20) y estimados mediante el modelo de regresión beta, se tiene un desvío de 56,057 con 54 grados de libertad, evidenciando que el ajuste del modelo mejora significativamente. Las estimaciones para las proporciones de éxito en cada tratamiento son $\widehat{\pi}_1 = 0{,}758$, $\widehat{\pi}_2 = 0{,}102$, $\widehat{\pi}_3 = 0{,}034$ y $\widehat{\pi}_4 = 0{,}048$ con errores estándar estimados iguales a 0,4321, 0,6966, 1,0849 y 0,8214 respectivamente.

De manera similar ocurre con los datos del Ejemplo 4.2; como la prueba no rechazó la hipótesis nula H_0, suponemos que $w_1 = w_2$. Al ajustar un modelo lineal generalizado ponderado con los pesos dados por la ecuación (21) y estimados a partir del ajuste del modelo de regresión beta, se obtuvo un desvío de 40,356 con 34 grados de libertad, evidenciando una mejora en el ajuste del modelo. Con el modelo lineal generalizado ponderado, las estimaciones de las proporciones son $\widehat{\pi}_1 = 0{,}169$ y $\widehat{\pi}_2 = 0{,}059$, con errores estándar de las estimaciones iguales a 0,2167 y 0,4061 respectivamente.

Se observa en ambos ejemplos que las proporciones de éxito estimadas con el modelo lineal generalizado ponderado no difieren significativamente de las obtenidas mediante el modelo lineal generalizado usual. Sin embargo, los errores estándar de estimación son más grandes, lo cual evidencia que se ha tenido en cuenta la dispersión extra presente en los datos.

6. Análisis de la potencia de la prueba mediante simulación

Dado que el procedimiento de prueba propuesto para la hipótesis (13) involucra métodos numéricos iterativos, no es posible obtener una expresión explícita para la función de potencia de la prueba. Para obtener una aproximación gráfica como la que se presenta en la figura 1, se llevó a cabo un estudio de simulación. Se generan los datos de acuerdo con el modelo betabinomial, el cual permite controlar el grado de sobredispersión de los datos simulados, usando la función **rbetabinom()** de la librería VGAM del paquete estadístico R [19]. Para tal efecto se consideraron los modelos $g_1(\mu_i) = \beta_0 + \beta_i$ para la media y $g_2(\delta_i) = \theta_0 + \theta_i$ para la precisión (modelos de una vía de clasificación) con $i = 1, \ldots, K$. Las observaciones de cada tratamiento fueron generadas de tal forma que $\Delta\phi = \phi_{k+1} - \phi_k$ fuera constante, es decir, que la diferencia entre la dispersión de dos niveles de tratamiento consecutivos fuera constante. Para los valores de los parámetros ϕ se tomó $\phi_1 = 0{,}10$, y para $i = 2, \ldots, K$, los valores de ϕ_i se establecieron de tal forma que la diferencia $\Delta\phi = \phi_{k+1} - \phi_k$ tome los valores 0, 0,01, 0,03, 0,04, 0,06, 0,07, 0,09, 0,10, 0,11, 0,13, 0,14, 0,16, 0,17, 0,19, 0,20.

Se empieza simulando datos bajo el supuesto de que la hipótesis H_0 es verdadera, y luego se van variando los parámetros de tal forma que cada vez sea más marcado el alejamiento de la hipótesis nula y sea la hipótesis alternativa la que es verdadera. Con

los datos obtenidos de esa forma se lleva a cabo el contraste de la hipótesis a un nivel de significación del 5 % y se registra si se rechaza o no la hipótesis. Estos pasos se repiten mil veces, y luego se calcula la proporción de rechazos de H_0; en la Tabla 4 se muestran las proporciones de rechazo de la hipótesis nula con datos sobre dispersos generados bajo un modelo a una vía de clasificación con $K = 2, 3, 4$ y 5 niveles del factor y 5, 15, 20, 30 repeticiones, y cuando $\phi_1 = \cdots = \phi_K = 0{,}10$. Se observa en todos los casos que la proporción de rechazos disminuye con el número de repeticiones; para dos vías de clasificación ($K = 2$), la proporción de rechazos es 0,054, muy cercano a 0,05, el valor esperado; para $K > 2$ la convergencia de la proporción de rechazos a 0,05 es más lenta, requiriendo tamaños de muestras más grandes. En todo caso, es claro el comportamiento asintótico de la prueba, es decir, a medida que el tamaño de la muestra aumenta, el tamaño del test se acerca al valor nominal 0,05.

	Número de repeticiones			
Nro. Trat	5	15	20	30
2	0,120	0,064	0,072	0,054
3	0,117	0,068	0,072	0,066
4	0,141	0,064	0,078	0,073
5	0,141	0,075	0,069	0,088

Tabla 4. Proporción de rechazos bajo H_0 cierta.

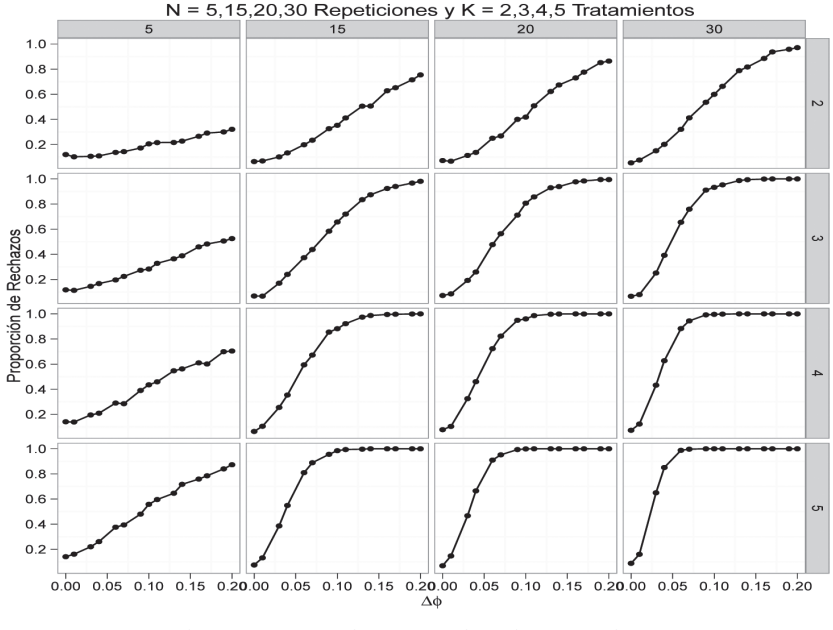

Figura 1. Función de potencia obtenida por simulación.

7. Análisis de residuales

Luego de ajustar el modelo lineal generalizado con ponderaciones, como se describe en la sección 5, se llevó a cabo el análisis de residuales con el fin de verificar el ajuste de los datos; en ambos ejemplos se pudo verificar la adecuación de las funciones de varianza y enlace. Una medida de influencia de cada observación sobre los parámetros de regresión estimados es la distancia de Cook (Cook [5]). El gráfico medio normal (*half normal plot*) para la distancia de Cook muestra que no hay evidencia de observaciones influyentes y datos atípicos, evidenciando que el procedimiento propuesto para corregir el problema de sobredispersión es apropiado, como lo corrobora el gráfico de los residuales [1] con bandas simuladas, usado en el análisis de residuales que se muestra en la Figura 2.

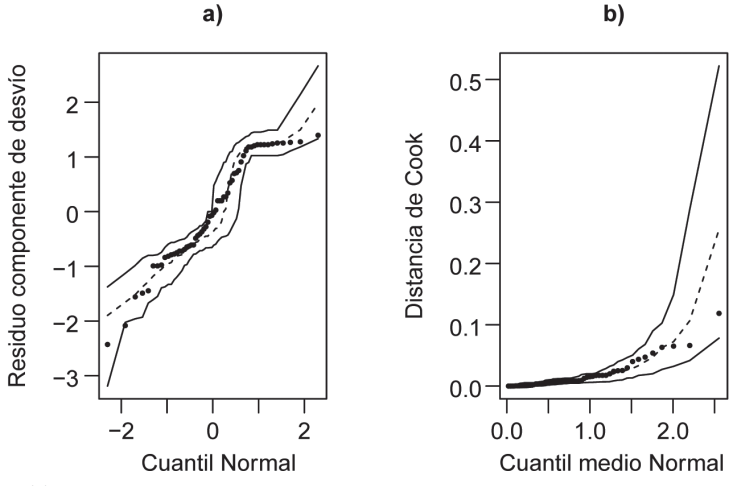

Figura 2. (a): Gráfico de probabilidad normal con banda de confianza simulada para los residuos y (b): Gráfico medio normal con banda de confianza simulada al 95 % para la distancia de Cook, Ejemplo 4.1.

8. Conclusiones

En presencia de datos de proporción sobre dispersos:

La regresión beta proporciona una forma sencilla y natural de verificar el supuesto de homogeneidad en el parámetro de dispersión mediante la prueba de la razón de verosimilitud.

Una forma de corregir el problema de la sobredispersión es mediante el ajuste de un modelo lineal generalizado ponderado; las ponderaciones o pesos se obtienen como un subproducto de los modelos de regresión beta que se ajustan para contrastar las hipótesis (13). La ventaja de esta aproximación es que la tarea de interpretar y obtener conclusiones a partir de los resultados del ajuste del modelo es más sencilla para el analista no experto en estadística. Al mismo tiempo, el investigador se protege de las consecuencias causadas por la sobredispersión.

[1] Se usaron los residuales de desvío (*deviance residual*), véase McCullagh & Nelder [13], sección 2.4 y capítulo 12.

Bajo los escenarios propuestos la potencia de la prueba basada en la razón de verosimilitud presenta un comportamiento asintótico: a medida que el tamaño de la muestra aumenta, se incrementa la capacidad de la prueba para rechazar la hipótesis nula H_0 cuando es falsa, y la proporción de rechazos bajo el supuesto de que H_0 es cierta tiende al valor nominal α previamente fijado.

Referencias

[1] Atkison A., *Plots, Tranformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*, Oxford University Press, New York, 1985.

[2] Barrios D.E. y Peña E., "Obtención de un empaque antimicrobiano a partir de almidon de ñame con adición de sorbato de potasio para su aplicación en la conservación de ñame minimamente procesado", Tesis, Programa de Ingeniería de Alimentos, Universidad de Córdoba, 2009.

[3] Brooks R.J., "Approximate Likelihood-Ratio Test in the Analysis of Beta-Binomial Data", *Stat. Appl.* 12 (1978), 1589–1596.

[4] Cox D.R. and Hinkley D.V., *Theoretical Statistics*, Chapman & Hall, London, 1974.

[5] Cook R.D., "Detection of Influencial Observations in Linear Regression", *Technometrics* 19 (1977), 15–18.

[6] Crowder M.J. "Beta-Binomial Anova for Proportions", *Appl. Stat.* 27 (1978), no. 1, 34–37.

[7] Crowder M.J., "Inference About the Intraclass Correlation Coefficient in the Beta–binomial ANOVA for Proportions", *J. R. Stat. Soc.* 41 (1979), no. 2, 230–234.

[8] Dobson A.J., *An Introduction to Generalized Linear Models*, Chapman & Hall/CRC, New York, 2 ed., 2002.

[9] Ferrari S.L.P. and Cribari-Neto F., "Beta Regression for Modelling Rates and Proportions", *J. Appl. Stat.* 31 (2004), no. 7, 799–815.

[10] Gordon K.S., *Optimization and Nonlinear Equations*, Encyclopedia of Biostatistics, Chichester, 1998.

[11] Hinde J. and Demétrio C., *Overdispersion: Models and Estimation*, ABE, São Pablo, 1998.

[12] Nocedal J. and Wright S.J., *Numerical Optimization*, Springer–Verlag, New York, 1999.

[13] McCullagh P. and Nelder J., *Generalized Linear Models*, Chapmann & Hall/CRC, New York, 1989.

[14] Mood A., Graybill F., and Boes D., *Introduction to the theory of statistics*, McGraw–Hill, 3 ed., 1974.

[15] Moore D.F. and Tsiatis A., "Robust Estimation of the Variance in Moment Methods for Extra-binomial and Extra-Poisson Variation", *Biometrics* 47 (1991), 383–401.

[16] Morales M.A. & López L.A., "Estudio de homogeneidad de la dispersión en diseño a una vía de clasificación para datos de proporciones y conteos", *Rev. Colombiana Estadíst.* 32 (2009), no. 1, 59–78.

[17] Nelder J.A. and Wedderburn D.W.M., "Generalized Linear Models", *J. R. Stat. Soc. Ser. A* 135 (1972), no. 3, 370–384.

[18] Press W.H., Teukolsky S.A., Vetterling W.T., and Flannery B.P., *Numerical recipes in C: the art of scientific computing*, Cambridge University Press, New York, 2 edition, 1992.

[19] R Development Core Team. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2012, URL `http://www.r-project.org`.

[20] Ravishanker N. and Dey D.K., *A First Course in Linear Model Theory*, Chapman & Hall/CRC, New York, 2001.

[21] Simas A., Barreto-Souza W., and Rocha A., "Improved Estimators for a General Class of Beta Regression Models", *Comput. Statist. Data Anal.* 54 (2010), no. 2, 348–366.

[22] Smithson M. and Verkuilen J., "A Better Lemon Squeezer? Maximum–Likelihood Regression with Beta–Distributed Depedent Variables", *Psychological Methods* 11 (2006), no. 1, 54–71.