FATOR DE CORREÇÃO PARA A DISTRIBUIÇÃO DA *DEVIANCE* PARA DADOS DE PROPORÇÕES

José Eduardo CORRENTE[1]
Ana Paula Gomes da Silva GIMENES[2]

■ RESUMO: A análise de dados de proporções apresenta, em geral, certas dificuldades uma vez que a distribuição subjacente a tais dados pode ser considerada binomial, que não segue algumas pressuposições básicas para o ajuste de um modelo matemático. Algumas transformações são sugeridas, mas nem sempre bons resultados são obtidos.

No enfoque de modelos lineares generalizados, a estatística que mede a qualidade do ajuste do modelo para os dados é chamada *deviance*. Ocorre que a distribuição da *deviance* é, em geral, desconhecida. No entanto, para dados com distribuição binomial, pode-se aproximar a distribuição da *deviance* por uma distribuição χ^2, mas tal aproximação não é boa para tamanhos pequenos de amostra. Para melhorar essa aproximação, alguns fatores de correção para os dados são sugeridos, mas os resultados obtidos ainda não são satisfatórios. Assim, o objetivo deste trabalho é propor um novo fator de correção para os dados seguindo uma distribuição binomial, de modo a se obter uma melhora na distribuição da *deviance* para qualquer tamanho de amostra. Para isto, adiciona-se uma constante à variável resposta e, através do valor esperado da *deviance*, calcula-se tal constante de modo a reduzir o erro cometido na aproximação. Simulações da distribuição binomial e o cálculo da *deviance* são feitos e *QQ-plots* são utilizados para a comparação com a distribuição χ^2.

[1]Departamento de Ciências Exatas, Escola Superior de Agricultura Luiz de Queiróz da Universidade de São Paulo, CEP: 13418-900, Piracicaba-SP, Brasil. E-mail: jecorren@carpa.ciagri.usp.br

[2]Departamento de Ciências Exatas, Escola Superior de Agricultura Luiz de Queiróz da Universidade de São Paulo, CEP: 13418-900, Piracicaba-SP, Brasil. E-mail: anapaula@linksat.com.br

▪ PALAVRAS-CHAVE: distribuição binomial, *deviance*, série de Taylor, modelos lineares generalizados

1 Introdução

Dados de proporções, em geral, podem ser supostos como provenientes de uma distribuição binomial e, nesses casos, é comum, quando possível, fazer uma transformação estabilizadora da variância a fim de se obter as pressuposições básicas para o ajuste do modelo a ser analisado.

Quando dados de proporções são provenientes de algum tipo de delineamento experimental, as pressuposições básicas para uma análise de variância clássica não são satisfeitas, principalmente devido à heterogeneidade de variâncias.

Cox e Snell (1968) mostram como se calcular resíduos para dados com distribuição binomial e como obter variáveis identicamente distribuídas aproximadas, e a escolha da transformação apropriada depende das exigências do modelo considerado. Já em Cox e Snell (1981), três métodos de se fazer uma transformação de dados são sugeridos.

Uma maneira de se evitar as transformações de dados é utilizar o enfoque de Modelos Lineares Generalizados, propostos por Nelder e Wedderburn (1972), para se proceder a análise. Neste caso, a medida da qualidade de ajuste do modelo proposto é a *deviance*, que compara o modelo superparametrizado com o modelo sob pesquisa, através da diferença dos logaritmos das verossimilhanças correspondentes.

A distribuição da *deviance* é desconhecida, mas para alguns casos particulares, como o caso da distribuição binomial, pode-se mostrar que sua distribuição é aproximadamente χ^2 com $n - p$ graus de liberdade, onde n é o tamanho da amostra e p é o número de parâmetros envolvidos no modelo sob pesquisa. Tal aproximação apresenta um resto da ordem de $O(m^{-1})$, onde m é o número de realizações independentes do ensaio binomial. Diversos fatores de correção têm sido propostos a fim de reduzir essa ordem de convergência. Apesar disso, o problema torna-se maior quando se trata de pequenas amostras.

Gart e Zweifel (1967) sugerem a adição de $0,5$ às frequências observadas para aperfeiçoar a aproximação do χ^2 de referência. Esse fator será bastante razoável quando todas as observações são

maiores que 5, o que é requerido também com relação aos valores esperados ao se aplicar o teste de χ^2 em tabelas de contingência.

Cordeiro (1986) relata que para examinar propriedades da *deviance* em pequenas amostras, torna-se necessária uma aproximação de ordem superior para o seu valor esperado, supondo o modelo sob pesquisa verdadeiro. Tem-se que a definição da *deviance* é

$$D_p = 2(\hat{L}_n - L) + 2(L - \hat{L}_p)$$

em que $\hat{L}_n$, $\hat{L}_p$ e L são os valores de $L(\beta)$ calculada nos pontos y, $\hat{\mu}$ e no parâmetro verdadeiro, respectivamente, dada por:

$$L(\beta) = \sum_{l=1}^{n} \phi_l [y_l \theta_l - b(\theta_l) + c(y_l, \phi_l)]$$

em que $\theta_l = q(\mu_l)$, $\mu_l = f^{-1}(\eta)$, $\eta = \sum_{i=1}^{p} x_{li}\beta_i$ e ϕ_l é suposto conhecido para cada observação.

Daí, o valor esperado da *deviance* é expresso, até termos de ordem n^{-1}, por

$$E(D_p) = 2E(\hat{L}_n - L) - (p + \epsilon_p) + O(n^{-2})$$

O termo $2(\hat{L}_n - L)$ representa a *deviance* no parâmetro verdadeiro e o seu valor esperado pode ser calculado a partir da distribuição proposta para os dados. Dado que:

$$\hat{L}_n = \sum_{l=1}^{n} \phi_l [y_l q(y_l) - b(q(y_l)) + c(y_l, \phi_l)]$$

vem, até ordem n^{-1},

$$E(D_p) = 2 \sum_{l=1}^{n} \phi_l \{E[v(Y_l)] - v(\mu_l)\} - (p + \epsilon_p) \qquad (1.1)$$

em que

$$v(y) = yq(y) - b(q(y))$$

com $b(.)$ sendo uma função conhecida e $q(.) = \theta$.

A vantagem de (1.1) é que se pode deduzir um fator de correção:

$$c = (n - p)/E(D_p)$$

Daí, pode-se definir uma *deviance* modificada $D_p^* = cD_p$ para o modelo sob pesquisa, tal que $E(D_p^*)$ é melhor aproximada por $(n-p)$ do que $E(D_p)$. Isto não implica, necessariamente, que D_p^* é melhor aproximada por uma distribição χ^2_{n-p}.

Frequentemente, $E(D_p)$ é uma função das médias desconhecidas e estas são estimadas a partir de $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1 \ldots \tilde{\beta}_p)^T$ para calcular um valor numérico para c. Isto claramente não afetará a ordem da aproximação obtida. Pelo menos, com grandes amostras, os momentos de D_p^* devem ser mais próximos dos correspondentes momentos da distribuição χ^2_{n-p} do que aqueles de D_p. Mas, para n pequeno, isto, em geral, não é verdadeiro.

McCullagh e Nelder (1991) apresentam a correção de Bartlett que surge exatamente a partir de um teste de hipóteses, utilizando a razão de verossimilhanças, que basicamente originou a *deviance*. Assim, ao se obter o valor esperado da *deviance* do modelo saturado e do modelo sob pesquisa, tem-se um fator que reduz a ordem de convergência desse valor esperado. A ordem de convergência passa a ser $O(m^{-3/2})$, obtendo-se assim uma boa aproximação para a distribuição da *deviance*. A correção de Bartlett é a mais usada em problemas envolvendo ajustes em modelos lineares generalizados, e já está implementada na maioria dos *softwares* disponíveis para essas análises.

Em Botter & Cordeiro (1998), uma proposta para a redução no vício de estimadores de máxima verossimilhança para os parâmetros do modelo linear generalizado com dispersão nas covariáveis é feita. Para esta classe de modelos, foram obtidas fórmulas gerais para os viéses de tais estimadores, expandindo-os até termos de segunda ordem para os parâmetros do modelo linear e da dispersão bem como para os preditores lineares, parâmetros de precisão e valores da média. Alguns casos especiais são tratados utilizando a distribuição normal para a variável resposta. Estudos de simulação também são feitos para variáveis com distribuição gama e normal inversa. Já em Cordeiro & Cribari-Neto (1998), a mesma proposta foi feita considerando estimadores de máxima verossimilhança para os parâmetros dos modelos de regressão não-lineares em famílias não-exponenciais, em que foi estudado a redução no vício para os preditores lineares e as médias. Os comportamentos dos estimadores com vício reduzido foi estudado através de simulação. Nota-se que, apesar de reduzir o vício dos estimadores sob diversos modelos, trabalha-se com a hipótese de

que a *deviance* possui uma distribuição assintótica χ^2, embora nada seja citado a esse respeito.

Desse modo, a fim de reduzir o vício na distribuição da *deviance* para dados de proporções através do estudo da distribuição binomial e torná-la mais próxima de uma distribuição χ^2, será proposto um novo fator de correção que será adicionado à variável resposta. Esse fator será obtido através da expansão em série de Taylor da função *deviance* e do cálculo de seu valor esperado. Na Seção 2 será apresentada a metodologia de obtenção desse fator. Na Seção 3 será feito um estudo de simulação com dados binomiais com e sem correção, apresentando-se *QQ-plots* para verificar a melhora na aproximação da distribuição da *deviance* a uma distribuição χ^2. Uma discussão sobre os resultados obtidos também será apresentada.

2 Metodologia

2.1 Expansão da *d*eviance para uma variável

Para medir a qualidade de um ajuste em Modelos Lineares Generalizados, utiliza-se a *deviance*, que é definida por:

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2l(\mathbf{y}, \mathbf{y}) - 2l(\hat{\boldsymbol{\mu}}, \mathbf{y})$$

em que: $l(\mathbf{y}, \mathbf{y})$ é o logaritmo da função de verossimilhança do modelo saturado, e $l(\hat{\boldsymbol{\mu}}, \mathbf{y})$ é o logaritmo da função de verossimilhança do modelo sob pesquisa.

Assim, para uma variável aleatória $Y \sim b(m, p)$, tem-se que a *deviance* calculada para essa variável é dada por:

$$D(y, \hat{\mu}) = 2\left\{ y \ln \frac{y}{\hat{\mu}} + (m-y) \ln \frac{(m-y)}{(m-\hat{\mu})} \right\}$$

e o valor esperado da *deviance* é dado por

$$E(D(y,\hat{\mu})) = 2 \sum_{y=1}^{m-1} \left[y \ln \frac{y}{\hat{\mu}} + (m-y) \ln \frac{(m-y)}{(m-\hat{\mu})} \right] \left(\begin{array}{c} m \\ y \end{array} \right) p^y (1-p)^{(m-y)}$$

que não tem valor explícito. Note que, nesta expressão, y toma valores de 1 até $m-1$ pois, quando $y = 0$ em $\ln(y/\hat{\mu})$ e quando $y = m$ em $\ln[(m-y)/(m-\hat{\mu})]$, não existe o valor esperado. Mas, considerando diversos valores de m e de p, pode-se calcular o valor

esperado numericamente e observar o comportamento desse valor, conforme mostram as Figuras 1 e 2.

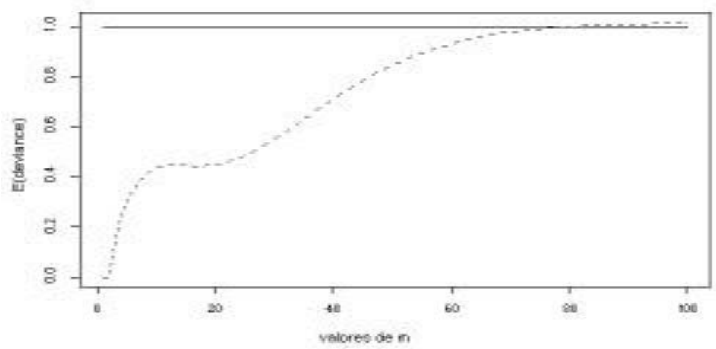


Figura - 1: Gráfico do valor esperado da *deviance* para uma variável com distribuição binomial para $m = 0, 1, \ldots, 100$ e $p = 0, 1$.

Sob a suposição de que $Y \sim b(m, p)$, espera-se que a *deviance* nesse caso se aproxime de uma distribuição χ^2 com 1 grau de liberdade. Assim, o valor esperado da *deviance* nesse caso seria próximo a 1. Como se pode observar na Figura 1, quando $p = 0, 1$ tem-se que o valor 1 é atingido para grandes valores de m, o mesmo não ocorrendo quando p aumenta, como mostra a Figura 2.

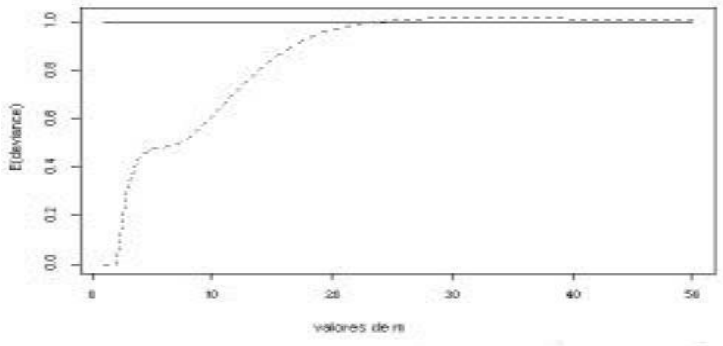


Figura - 2: Gráfico do valor esperado da *deviance* para uma variável com distribuição binomial para $m = 0, 1, \ldots, 50$ e $p = 0, 3$.

Isso pode indicar um mal ajuste do modelo sob pesquisa pela falta de ajuste da distribuição da *deviance* a uma distribuição χ^2.

Assim, para o caso de uma variável aleatória com distribuição binomial de parâmetros m e p, pode-se calcular uma aproximação do valor esperado da *deviance*, fazendo

$$D(y, m, p) = g(y)$$

e expandindo essa função em torno do ponto $\mu = mp$ em série de Taylor. Desta maneira, tem-se que

$$\begin{aligned} g(y) &= g(\mu) + g'(\mu)(y-\mu) + g''(\mu)\frac{(y-\mu)^2}{2!} \\ &+ g'''(\mu)\frac{(y-\mu)^3}{3!} + g^{iv}(\mu)\frac{(y-\mu)^4}{4!} + \cdots \end{aligned} \qquad (2.2)$$

em que

$$g(\mu) = 2\left[\mu \ln \frac{\mu}{\mu} + (m-\mu)\ln\frac{m-\mu}{m-\mu}\right] = 0$$

$$g'(y) = 2\left[\ln\frac{y}{mp} - \ln\frac{m-y}{m-mp}\right] \Rightarrow g'(\mu) = 2\left[\ln\frac{mp}{mp} - \ln\frac{m-mp}{m-mp}\right] = 0$$

$$g''(y) = \frac{2}{y} + \frac{2}{m-y} \Rightarrow g''(\mu) = \frac{2}{mp} + \frac{2}{m-mp}$$

$$g'''(y) = -\frac{2}{y^2} + \frac{2}{(m-y)^2} \Rightarrow g'''(\mu) = -\frac{2}{m^2p^2} + \frac{2}{(m-mp)^2}$$

$$g^{iv}(y) = \frac{4}{y^3} + \frac{4}{(m-y)^3} \Rightarrow g^{iv}(\mu) = \frac{4}{m^3p^3} + \frac{4}{(m-mp)^3}$$

Considerando-se agora a variável aleatória Y e calculando o valor esperado de (2.2), tem-se que:

$$\begin{aligned} E(g(Y)) &= 0 + 0 + \left(\frac{2}{mp} + \frac{2}{m(1-p)}\right)\frac{mp(1-p)}{2!} \\ &+ \left(\frac{-2}{m^2p^2} + \frac{2}{m^2(1-p)^2}\right)\left[\frac{mp - 3mp^2 + 2mp^3}{3!}\right] \\ &+ \left(\frac{4}{m^3p^3} + \frac{4}{m^3(1-p)^3}\right)\left[\frac{mp - 7mp^2 + 12mp^3 - 6mp^4 + 3m^2p^2(1-p)^2}{4!}\right] + \cdots \end{aligned}$$

Simplificando-se, segue que

$$E(g(Y)) = E(D(Y, m, p)) = 1 + \frac{1}{6m}\left[\frac{1-p+p^2}{p(1-p)}\right] + O(m^{-2})$$

Se a distribuição da *deviance* é aproximadamente χ^2 com 1 grau de liberdade, seu valor esperado deve se aproximar de 1. Mas, observa-se que existe um viés da ordem de m^{-1}, que tende a zero somente quando m torna-se grande. Assim, uma maneira de tornar esse valor esperado próximo a 1, seria anular os termos de ordem m^{-1}. Para isto, propõe-se uma correção para a variável resposta de modo que a reduzir a ordem de convergência do valor esperado. Isso garantiria uma melhor aproximação da distribuição da *deviance* a uma distribuição χ^2.

Desse modo, substituindo

$$y \to y + a$$

$$m \to m + 2a$$

ainda com a suposição de que $Y + a \sim b(m, p)$, é preciso obter o valor de a de modo a zerar o termo de $O(m^{-1})$.

A *deviance* para a variável corrigida $Y + a$ será dada por

$$D(y + a, m + 2a, p) = 2\left[(y + a)\ln\frac{y + a}{mp} + (m + a - y)\ln\frac{m + a - y}{m + 2a - mp}\right]$$

e considere novamente:

$$D(y + a, m + 2a, p) = g(y)$$

Expandindo $g(y)$ em série de Taylor em torno do ponto $\mu = mp$, tem-se que

$$g(\mu) = 0$$

$$g'(y) = 2\left[\ln\frac{y + a}{mp} - \ln\frac{m + a - y}{m + 2a - mp}\right] \Rightarrow g'(\mu) = 0$$

$$g''(y) = \frac{2}{y + a} + \frac{2}{m + a - y} \Rightarrow g''(\mu) = \frac{2}{(mp)} + \frac{2}{m + 2a - mp}$$

$$g'''(y) = -\frac{2}{(y + a)^2} + \frac{2}{(m + a - y)^2} \Rightarrow g'''(\mu) = -\frac{2}{(mp)^2} + \frac{2}{(m + 2a - mp)^2}$$

$$g^{iv}(y) = \frac{4}{(y + a)^3} + \frac{4}{(m + a - y)^3} \Rightarrow g^{iv}(\mu) = \frac{4}{(mp)^3} + \frac{4}{(m + 2a - mp)^3}.$$

Aplicando o valor esperado na expansão, considerando agora a variável aleatória Y, tem-se que:

$$
\begin{aligned}
E(D(Y \quad + \quad & a, m+2a, p)) = 0+0.E(Y-\mu)+\left[\frac{2}{(mp)}+\frac{2}{m+2a-mp}\right]\frac{E(Y-\mu)^2}{2!} \\
+ \quad & \left[-\frac{2}{(mp)^2}+\frac{2}{(m+2a-mp)^2}\right]\frac{E(Y-\mu)^3}{3!} \\
+ \quad & \left[\frac{4}{(mp)^3}+\frac{4}{(m+2a-mp)^3}\right]\frac{E(Y-\mu)^4}{4!}+\cdots \\
= \quad & \left[\frac{2}{mp}+\frac{2}{m+2a-mp}\right]\frac{mp(1-p)}{2!}+\left[-\frac{2}{m^2p^2}+\frac{2}{(m+2a-mp)^2}\right] \\
& .\frac{mp(1-p)(1-2p)}{3!}+\left[\frac{4}{m^3p^3}+\frac{4}{(m+2a-mp)^3}\right] \\
& .\frac{3m^2p^2(1-p)^2+mp(1-p)(1-6p(1-p))}{4!}+\cdots \\
= \quad & 1-p-\frac{1-3p+2p^2}{3mp}+\frac{(1-p)^2}{2mp}+\frac{mp(1-p)}{m+2a-mp} \\
+ \quad & \frac{mp(1-3p+2p^2)}{3(m+2a-mp)^2}+\frac{m^2p^2(1-p)^2}{2(m+2a-mp)^3}+O(m^{-2}) \qquad (2.3)
\end{aligned}
$$

Expandindo cada termo de (2.3) em série de Taylor em torno do ponto $a = 0$, segue que:

$$
\begin{aligned}
E(D(Y+a, m+2a, p)) \quad = \quad & 1-\frac{2ap}{m-mp}+\frac{mp(1-3p+2p^2)}{3(m-mp)^2}+\frac{m^2p^2(1-p)^2}{2(m-mp)^3} \\
& -\frac{1-3p+2p^2}{3mp}+\frac{(1-p)^2}{2mp}+O(m^{-2}) \\
= \quad & 1-\frac{1}{6}\frac{12ap^2-1+p-p^2}{mp(1-p)}+O(m^{-2})
\end{aligned}
$$

Igualando o termo da ordem de $O(m^{-1})$ a zero, tem-se que:

$$
-\frac{1}{6}\frac{12ap^2-1+p-p^2}{mp(1-p)} = 0 \quad \Rightarrow \quad a = \frac{1-p+p^2}{12p^2}
$$

obtendo-se assim o valor de a que reduz a ordem de convergência na expansão do valor esperado da *deviance*, para variáveis aleatórias com distribuição binomial de parâmetros m e p.

2.2 Expansão da Deviance para n Variáveis

Considere as variáveis aleatórias $Y_1, \ldots, Y_n$ independentes e identicamente distribuídas com distribuição $b(m, p)$. A *deviance* envolvendo as n variáveis é dada agora por:

$$
D(\mathbf{y}, m, p) = 2\sum_{i=1}^{n}\left(y_i \ln\frac{y_i}{\hat{\mu}}+(m-y_i)\ln\frac{m-y_i}{m-\hat{\mu}}\right)
$$

e o valor esperado da *deviance* é dado por:

$$E(D(\mathbf{y},m,p)) = 2\sum_{i=1}^{n}\left(y_i \ln\frac{y_i}{\hat{\mu}} + (m-y_i)\ln\frac{m-y_i}{m-\hat{\mu}}\right)\left(\begin{array}{c} m \\ y_i \end{array}\right) p^{y_i}(1-p)^{(m-y_i)}$$

Para este caso, o modelo sendo ajustado é o modelo de média e portanto $\hat{\mu} = \bar{y}$. Logo,

$$D(\mathbf{y},m,p) = 2\sum_{i=1}^{n}\left(y_i \ln\frac{ny_i}{\sum_{i=1}^{n} y_i} + (m-y_i)\ln\frac{n(m-y_i)}{nm-\sum_{i=1}^{n} y_i}\right)$$

De modo análogo ao caso de uma variável, considere

$$D(\mathbf{y},m,p) = g(\mathbf{y})$$

e vê-se que $g(\boldsymbol{\mu}) = 0$. Expandindo essa função em torno de $\boldsymbol{\mu}$, segue que

$$g(\mathbf{y}) = g(\boldsymbol{\mu}) + dg(\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu}) + \frac{1}{2!}d^2g(\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})^2 + \frac{1}{3!}d^3g(\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})^3 + \cdots \quad (2.4)$$

e considerando a variável aleatória Y, o valor esperado de $g(Y)$ será dado por:

$$\begin{aligned} E(g(\mathbf{y})) &= g(\mu) + \sum_{i=1}^{n}\left.\frac{\partial g}{\partial y_i}\right|_{y_i=mp} E(Y_i-\mu) + \frac{1}{2!}\sum_{i=1}^{n}\left.\frac{\partial^2 g}{\partial y_i^2}\right|_{y_i=mp} E(Y_i-\mu)^2 \\ &+ \frac{1}{3!}\sum_{i=1}^{n}\left.\frac{\partial^3 g}{\partial y_i^3}\right|_{y_i=mp} E(Y_i-\mu)^3 + \frac{1}{4!}\sum_{i=1}^{n}\left.\frac{\partial^4 g}{\partial y_i^4}\right|_{y_i=mp} E(Y_i-\mu)^4 \\ &+ \frac{1}{2!2!}\sum_{i<j=2}^{n}\left.\frac{\partial^4 g}{\partial y_i^2 \partial y_j^2}\right|_{y_i,y_j=mp} E[(Y_i-\mu)^2(Y_j-\mu)^2] + \cdots \end{aligned} \quad (2.5)$$

Pode-se observar que as derivadas parciais mistas de 2^a ordem e 3^a ordem envolvem termos da forma $(y_1-\mu)(y_2-\mu)$, $(y_1-\mu)(y_2-\mu)^2$, que possuem valor esperado nulo. Mas isto não ocorre com a derivada parcial mista de 4^a ordem que envolve fatores da forma $(y_1-\mu)^2(y_2-\mu)^2$, cujo valor esperado é a variância ao quadrado. Desse modo a derivada parcial mista de 4^a ordem de g para y_1 e y_2 é:

$$\frac{\partial^4 g}{\partial y_1^2 \partial y_2^2} = \frac{4}{n^3}\left[-\frac{1}{(\bar{y})^3} - \frac{1}{(m-\bar{y})^3}\right]$$

Calculando os termos da série, segue que:

$$\sum_{i=1}^{n}\left.\frac{\partial g}{\partial y_i}\right|_{y_i=mp} = 2\left[\sum_{i=1}^{n}\ln\frac{mp}{mp} - \sum_{i=1}^{n}\frac{m-mp}{m-mp}\right] = 0$$

$$\sum_{i=1}^{n} \frac{\partial^2 g}{\partial y_i^2}\bigg|_{y_i=mp} = 2(n-1)\left[\frac{1}{mp} + \frac{1}{m-mp}\right]$$

$$\sum_{i=1}^{n} \frac{\partial^3 g}{\partial y_i^3}\bigg|_{y_i=mp} = 2(n-1)\left[-\frac{1}{(mp)^2} + \frac{1}{(m-mp)^2}\right]$$

$$\sum_{i=1}^{n} \frac{\partial^4 g}{\partial y_i^4}\bigg|_{y_i=mp} = 4(n-1)\left[\frac{1}{(mp)^3} + \frac{1}{(m-mp)^3}\right]$$

$$\sum_{i<j=2}^{n} \frac{\partial^4 g}{\partial y_i^2 \partial y_j^2}\bigg|_{y_i,y_j=mp} = \frac{2(n-1)}{n^2}\left[-\frac{1}{(mp)^3} - \frac{1}{(m-mp)^3}\right]$$

Daí,

$$\begin{aligned}
& E(D(\mathbf{y},m,p)) = 0 + 0 + 2(n-1)\left[\frac{1}{mp} + \frac{1}{m-mp}\right]\frac{mp(1-p)}{2!} \\
+ \quad & 2(n-1)\left[-\frac{1}{(mp)^2} + \frac{1}{(m-mp)^2}\right]\frac{mp(1-p)(1-2p)}{3!} \\
+ \quad & 4(n-1)\left[\frac{1}{(mp)^3} + \frac{1}{(m-mp)^3}\right]\frac{mp(1-p)(1-6p(1-p)) + 3m^2p^2(1-p)^2}{4!} \\
+ \quad & \frac{2(n-1)}{n^2}\left[-\frac{1}{(mp)^3} - \frac{1}{(m-mp)^3}\right]\frac{m^2p^2(1-p)^2}{2!2!} + \cdots \\
= \quad & (n-1) + \frac{1}{6}\frac{(n^2 - n^2p + n^2p^2 - 3 + 9p - 9p^2)(n-1)}{mp(1-p)n^2} + O(m^{-2})
\end{aligned}$$

Neste caso, se a distribuição da *deviance* é assintoticamente $\chi^2_{(n-1)}$, o seu valor esperado deve ser aproximadamente $(n-1)$. Novamente, observa-se que existe um viés da ordem de m^{-1}, que tende a zero somente quando m torna-se grande. Assim, para que se tenha uma melhor aproximação do valor esperado de $(n-1)$, é necessário que os termos de $O(m^{-1})$ se anulem.

A fim de reduzir a ordem de convergência da *deviance* e torná-la mais próxima a uma distribuição de χ^2, introduz-se uma correção aos valores observados, denotada por a e, do mesmo modo, substitui-se y por $y+a$ e m por $m+2a$. Assim, a fórmula da *deviance* com a correção ficará da seguinte forma:

$$D(\mathbf{y}+a, m+2a, p) = 2\sum_{i=1}^{n}\left[(y_i+a)\ln\frac{y_i+a}{\bar{y}+a} + (m+a-y_i)\ln\frac{m+a-y_i}{m+a-\bar{y}}\right]$$

e considere $D(\mathbf{y}+a, m+2a, p) = g(\mathbf{y})$.

Utilizando a mesma expansão dada em (2.4), tem-se que

$$g(\mu) = 0$$

$$\sum_{i=1}^{n} \frac{\partial g}{\partial y_i}\bigg|_{y_i=mp} = 2\left[\sum_{i=1}^{n} \ln \frac{mp}{mp} - \sum_{i=1}^{n} \ln \frac{m+2a-mp}{m+2a-mp}\right] = 0$$

$$\sum_{i=1}^{n} \frac{\partial^2 g}{\partial y_i^2}\bigg|_{y_i=mp} = 2(n-1)\left[\frac{1}{mp} + \frac{1}{m+2a-mp}\right]$$

$$\sum_{i=1}^{n} \frac{\partial^3 g}{\partial y_i^3}\bigg|_{y_i=mp} = 2(n-1)\left[-\frac{1}{(mp)^2} + \frac{1}{(m+2a-mp)^2}\right]$$

$$\sum_{i=1}^{n} \frac{\partial^4 g}{\partial y_i^4}\bigg|_{y_i=mp} = 4(n-1)\left[\frac{1}{(mp)^3} + \frac{1}{(m+2a-mp)^3}\right]$$

$$\sum_{i<j=2}^{n} \frac{\partial^4 g}{\partial y_i^2 \partial y_j^2}\bigg|_{y_i,y_j=mp} = \frac{2(n-1)}{n^2}\left[-\frac{1}{(mp)^3} - \frac{1}{(m+2a-mp)^3}\right]$$

Daí, o valor esperado da *deviance* para as variáveis corrigidas será dado por:

$$\begin{aligned} E(D(\mathbf{y} \quad + \quad & a, m+2a, p)) = 0 + 0 + 2(n-1)\left[\frac{1}{mp} + \frac{1}{m+2a-mp}\right]\frac{mp(1-p)}{2!} \\ + \quad & 2(n-1)\left[-\frac{1}{(mp)^2} + \frac{1}{(m+2a-mp)^2}\right]\frac{mp(1-p)(1-2p)}{3!} \\ + \quad & 4(n-1)\left[\frac{1}{(mp)^3} + \frac{1}{(m+2a-mp)^3}\right] \\ & \cdot\frac{[mp(1-p)(1-6p(1-p)) + 3m^2p^2(1-p)^2]}{4!} \\ + \quad & \frac{2(n-1)}{n^2}\left[-\frac{1}{(mp)^3} - \frac{1}{(m+2a-mp)^3}\right]\frac{m^2p^2(1-p)^2}{2!2!} + \cdots \end{aligned} \tag{2.6}$$

Fazendo as simplificações necessárias e expandindo novamente os termos de (2.6) em série de Taylor em torno de $a = 0$, segue que:

$$\begin{aligned} E(D(\mathbf{y} \quad + \quad & a, m+2a, p)) = (n-1) - \frac{2(n-1)p}{m(1-p)}a - \frac{1}{3}\frac{(1-2p)^2(n-1)}{mp(1-p)} \\ + \quad & \frac{1}{2}\frac{(n-1)(1-3p+3p^2)}{mp(1-p)} - \frac{1}{2}\frac{(1-3p+3p^2)(n-1)}{n^2mp(1-p)} + \cdots \\ = \quad & (n-1) - \frac{1}{6}\frac{(n-1)(n^2(12p^2a - 1 + p - p^2) - 3(1-3p+3p^2))}{n^2mp(1-p)} \\ + \quad & O(m^{-2}) \end{aligned}$$

Para se encontrar o valor de a que torna o termo da expansão da ordem de $O(m^{-1})$ nulo, faz-se:

$$0 \quad = \quad -\frac{1}{6}\frac{(n-1)(n^2(12p^2a - 1 + p - p^2) - 3(1-3p+3p^2))}{n^2mp(1-p)}$$

$$a = \frac{1-p+p^2}{12p^2} - \frac{1-3p+3p^2}{4n^2p^2} = \frac{1-p+p^2}{12p^2} + o(n^{-2})$$

que é o valor que reduz o viés da distribuição da *deviance*, aproximando-a à distribuição de χ^2, cujo valor esperado é $(n-1)$ (número de graus de liberdade para n variáveis).

3 Resultados e Discussão

De acordo com o desenvolvimento teórico, para o caso de uma variável aleatória com distribuição binomial de parâmetros m e p, o novo fator proposto para correção da distribuição da *deviance* foi

$$a = \frac{1-p+p^2}{12p^2}.$$

Como o fator de correção é uma função de p, é necessário uma análise dessa função de modo a tornar o fator constante. Assim, foram feitas várias tentativas com diferentes valores de p, e verificou-se que para $p = 0,3$ e, portanto, $a = 0,73$, foi o valor obtido que melhor aproximou a distribuição da *deviance* a uma distribuição de χ^2. Pode-se observar, a seguir, que este valor constante corrige satisfatoriamente a *deviance* para qualquer valor de p.

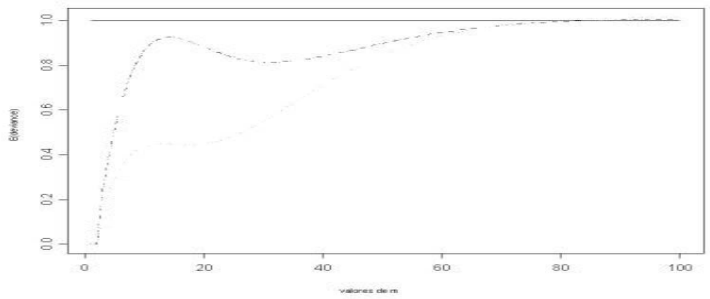


Figura - 3: Gráfico do valor esperado da *deviance* para uma variável com distribuição binomial para $m = 0, 1, \ldots, 100$ e $p = 0,1$ com e sem o fator de correção. (- - - - = deviance sem correção, — - — - = deviance com correção)

As Figuras 3 e 4 mostram o valor esperado da *deviance* para as variáveis sem correção e variáveis corrigidas para o valor de $a = 0,73$, para diversos valores de m e p.

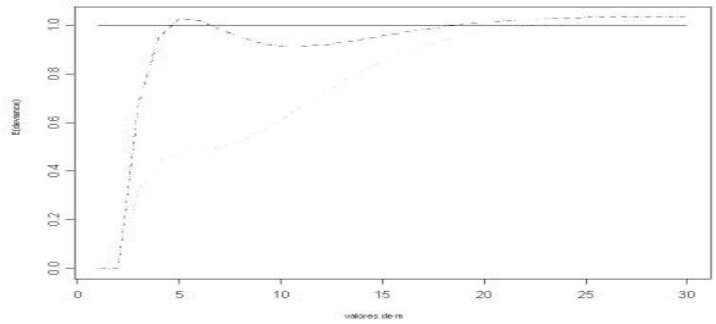


Figura - 4: Gráfico do valor esperado da *deviance* para uma variável com distribuição binomial para $m = 0, 1, \ldots, 30$ e $p = 0,3$. (- - - - = deviance sem correção, — - — - = deviance com correção)

Observando as Figuras 3 e 4, nota-se que o fator corrige o valor esperado da *deviance*, tornando-o mais próximo a 1, inclusive para pequenos valores de m, o que não ocorre com os fatores de correção propostos na literatura.

As Figuras de 5 a 8 mostram os *QQ-plots* dos valores da *deviance* contra os valores de uma distribuição de χ^2 com 1 grau de liberdade para as variáveis sem e com correção, respectivamente. A *deviance* foi obtida para dados simulados da distribuição binomial para diversos valores de m e p. Foram geradas amostras de tamanho dois com 1000 replicações. Nota-se uma melhora na distribuição da *deviance* quando corrigida pelo fator a para variáveis com m pequeno e p qualquer, e também para variáveis com m grande (até o valor de $m = 50$) e p pequeno. Para valores de $m > 50$, a correção é desnecessária.

Novamente, observa-se que o fator de correção melhora a aproximação da *deviance* a uma distribuição χ^2.

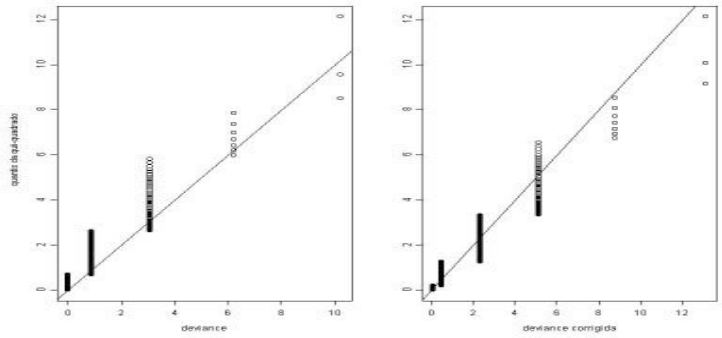
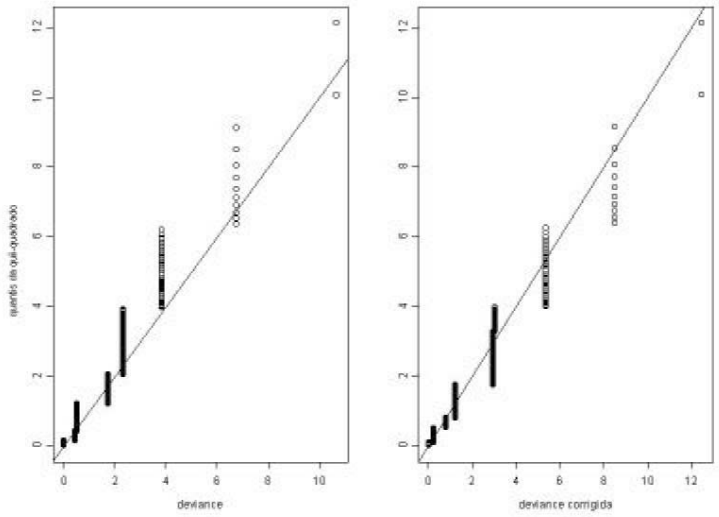


Figura - 5: *QQ-plots* dos valores da *deviance* da distribuição binomial contra os valores da χ^2 com 1 g.l. para o modelo de média, sem e com o fator de correção para $m = 10$ e $p = 0,1$.



m heightm height

Figura - 6: *QQ-plots* dos valores da *deviance* da distribuição binomial contra os valores da χ^2 com 1 g.l. para o modelo de média sem e com o fator de correção para $m = 10$ e $p = 0,3$.

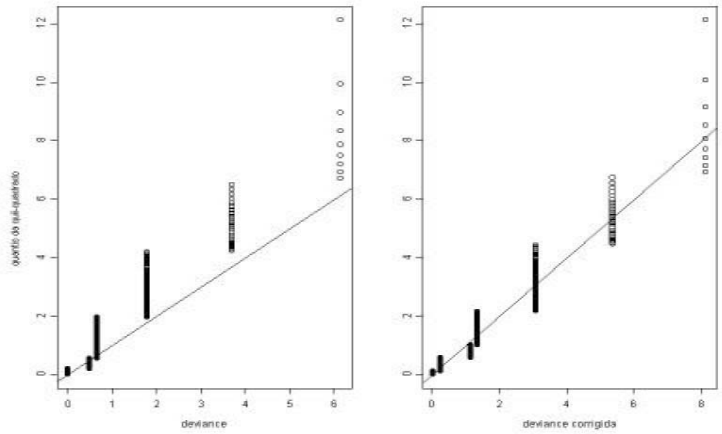


Figura - 7: *QQ-plots* dos valores da *deviance* da distribuição binomial contra os valores da χ^2 com 1 g.l. para o modelo de média sem e com o fator de correção para $m = 20$ e $p = 0,1$.

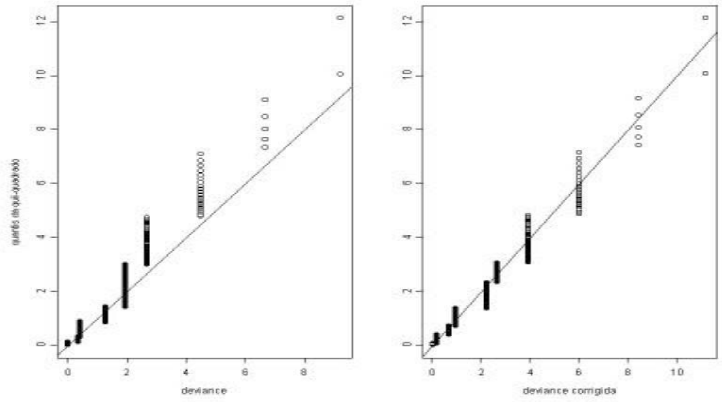


Figura - 8: *QQ-plots* dos valores da *deviance* da distribuição binomial contra os valores da χ^2 com 1 g.l. para o modelo de média sem e com o fator de correção para $m = 30$ e $p = 0,1$.

4 Conclusão

Considerando os desenvolvimentos para o ajuste do modelo de médias para n variáveis, o fator obtido foi

$$a = \frac{1-p+p^2}{12p^2} - \frac{1-3p+3p^2}{4n^2p^2} = \frac{1-p+p^2}{12p^2} + o(n^{-2})$$

e nota-se que para n variáveis, a correção depende novamente de p e de n^{-2}. Mas quando n aumenta, n^{-2} tende a zero rapidamente. Assim, o valor da correção para n variáveis se equipara ao valor da correção para uma variável e as mesmas considerações sobre o fator $a = 0,73$ podem ser feitas em relação aos valores de m e p.

Desse modo, conclui-se que o novo fator de correção para variáveis binomiais independentes e identicamente distribuídas que melhor aproxima a distribuição da *deviance* a uma distribuição χ^2, pode ser considerado $a = 0,73$ para o ajuste de um modelo de médias. Estudos mais aprofundados devem ser feitos no sentido de verificar também se esse fator corrige modelos para variáveis independentes somente, supondo que as médias sigam um modelo linear ou um modelo linear generalizado.

Agradecimentos

Os autores agradecem aos comentários e sugestões dos revisores que muito contribuíram para enriquecer este trabalho.

CORRENTE, J.E., GIMENES, A.P.G.S. Correction Factor for the Deviance Distribution of Proportion Data. *Rev. Mat. Estat.* (São Paulo), v. 20, p. 175-193, 2002.

■ *ABSTRACT: The analysis of proportion data generally presents certain difficulties since the distribution underlying such data can be considered to be binomial, that is, it does not follow some basic assumptions for the fit of a mathematical model. A few transformations are suggested; however, good results are not always obtained.In the approach given by generalized linear models, the statistic that measures the quality of the model's fit for such data is called deviance. It so happens that the deviance distribution is generally unknown. However, it can be approximated by a* χ^2 *distribution for data with a binomial distribution. Nevertheless, that is not good for small sample sizes. In order to improve this approximation, some data correction factors are suggested, but the obtained results are still unsatisfactory. Therefore, this work aims at proposing a new correction factor for data following a binomial distribution so as to obtain an improvement in the deviance distribution for any sample size. To that end, a constant is added to the response variable and, through the expected deviance value, such constant is calculated so as to reduce the error made in the approximation. Simulations of the binomial distribution and deviance calculation are made and QQ-plots are used to compare with distribution* χ^2.

■ *KEYWORDS: Binomial distribution, deviance, Taylor series, generalized linear models.*

Referências

BOTTER, D.A.; CORDEIRO, G.M. Improved estimators for generalized linear models with dispersion covariates. *J. Stat. Comput. Simulat.*, v. 62, p. 91-104, 1998.

COLOSIMO, E.A.; SILVA, A.F.; CRUZ, F.R.B. Bias evaluation in the proportional hazards model. *J. Stat. Comput. Simulat.*, v. 65, p. 191-201, 2000.

CORDEIRO, G.M. it Modelos lineares generalizados. In: SINAPE, 7., 1986, Campinas. 286p.

CORDEIRO, G.M.; CRIBARI-NETO, F. On bias reduction in exponential and non-exponential family regression models. *Commun. Stat. Simulat.*, v. 27, n. 2, p. 485-500, 1998.

COX, D.R.; SNELL, E.J. *Applied Statistics*. London: Chapman and Hall, 1981. 189p.

COX, D.R.; SNELL, E.J. A general definition of residuals. *J. R. Stat. Soc., Series B*, v.30, n.2, p.248-275, 1968.

GART, J.J.; ZWEIFEL, J.R. On the bias of various estimators of the logit and its variance with application to quantal bioassay. *Biometrika*, v.54, n.3, p.181-187, 1967.

MCCULLAGH, P.; NELDER, J.A. *Generalized linear models*. London: Chapman and Hall, 1991. 511p.

NELDER, J.A.; WEDDERBURN, R.W.M. Generalized linear models. *J. R. Stat. Soc., Series A*, v.135, n.3, p.370-384, 1972.

Statistical science S-plus for windows: user's manual. Seattle: StatiSci, 1993. 2v.

Recebido em 16.08.2000.