APLICAÇÃO DA MATRIZ DE COVARIÂNCIA DE SEGUNDA-ORDEM DAS ESTIMATIVAS DE MÁXIMA VEROSSIMILHANÇA NA INDÚSTRIA

Rosângela Getirana SANTANA[1]
Margareth Cizuka Toyama UDO[1]
Isolde Terezinha Santos PREVIDELLI[1]
Gauss Moutinho CORDEIRO[2]

▪ RESUMO: Cordeiro e McCullagh (1991) e Cordeiro (2004) apresentam fórmulas matriciais simples para avaliar os vieses de ordem $O\left(n^{-1}\right)$ das estimativas de máxima verossimilhança em modelos lineares generalizados e a estrutura de covariância dessas estimativas até ordem $O\left(n^{-2}\right)$, respectivamente. Entretanto, essas fórmulas ainda não foram exploradas num contexto prático. Neste artigo, são analisados dados de celulose do papel através de um modelo gama com função de ligação logarítmica que mostram a adequabilidade dessas fórmulas no contexto industrial. A implementação é simples e foi feita utilizando o sistema computacional *SAS*. Alguns resultados de simulação evidenciam que para amostras reduzidas torna-se indispensável o uso da matriz de covariância de ordem $O\left(n^{-2}\right)$ das estimativas de máxima verossimilhança em modelos lineares generalizados.

▪ PALAVRAS-CHAVE: Estimativa de máxima verossimilhança; estimativa corrigida; matriz de covariância corrigida; matriz de informação; viés da estimativa.

1 Introdução

O processo de inferir a partir dos dados observados sobre parâmetros desconhecidos é parte fundamental da lógica indutiva. A inferência científica se confunde com a inferência estatística quando a conexão entre o "estado da natureza

[1]Departamento de Estatística, Universidade Estadual de Maringá – UEM, CEP: 87020-900, Maringá, PR, Brasil. E-mail: *rgsantana@uem.br / mctudo@uem.br / itsprevidelli@uem.br*
[2]Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco – UFRPE, CEP: 52171-900, Recife, PE, Brasil. E-mail: *gauss@deinfo.ufrpe.br*

desconhecido" e os fatos observados são expressos em termos probabilísticos, i.e., o mecanismo de geração dos dados é governado por uma componente especificada e um erro estocástico que varia de acordo com uma distribuição de probabilidade (conhecida ou desconhecida). Esta composição define o modelo estatístico que descreve a estrutura probabilística dos dados como função de quantidades de interesse conhecidas e de outros parâmetros possivelmente desconhecidos.

Os modelos lineares generalizados (MLGs), introduzidos por Nelder e Wedderburn (1972), representam uma metodologia muito importante para aplicações estatísticas, pois são muito flexíveis e permitem o relaxamento das hipóteses básicas de normalidade, homocedasticidade e linearidade, todas necessárias aos modelos normais lineares de regressão. Na classe dos MLGs, a componente aleatória é especificada por uma distribuição de probabilidade pertencente à família exponencial de distribuições e a componente sistemática é dada por uma função de ligação que permite modelar uma relação (possivelmente não-linear) entre a média e a estrutura linear especificada por um conjunto de variáveis explicativas que formam a matriz modelo, tornando-se, assim, uma ferramenta poderosa nas aplicações estatísticas.

Considera-se que a distribuição comum das variáveis respostas $Y_1, \ldots, Y_n$, supostas independentes, é representada pela família exponencial de distribuições dada por

$$\pi\left(y;\theta_i,\phi\right) = \exp\left\{\phi\left[y\,\theta_i - b\left(\theta_i\right)\right] + c\left(y,\phi\right)\right\}, \tag{1}$$

sendo as funções $b\left(\cdot\right)$ e $c\left(\cdot,\cdot\right)$ conhecidas. A média e a variância de Y_i para $i = 1,\ldots,n$ são, respectivamente,

$$E\left(Y_i\right) = \mu_i = \frac{db\left(\theta_i\right)}{d\theta_i} \quad \text{e} \quad Var\left(Y_i\right) = \phi^{-1}V_i$$

sendo $V_i = d\mu_i/d\theta_i$ a função de variância que corresponde à parte da variância de Y_i que depende da média μ_i. Supõe-se que o *parâmetro de precisão* ϕ seja o mesmo para todas as observações e conhecido. O inverso do parâmetro ϕ é denominado de *parâmetro de dispersão*. A função de variância define a distribuição na família exponencial e desempenha um papel importante na teoria assintótica.

Um MLG é definido por alguma distribuição de probabilidade na família de distribuições (1) e pela componente sistemática $g(\mu) = \eta = X\beta$, sendo $g(\cdot)$ a função de ligação, $\mu = \left(\mu_1,\ldots,\mu_n\right)^T$ o vetor de médias, $\eta = \left(\eta_1,\ldots,\eta_n\right)^T$ o preditor linear, X a matriz modelo de dimensão $n \times p$ e de posto $p < n$ e $\beta = \left(\beta_1,\ldots,\beta_p\right)^T$ o vetor de parâmetros lineares desconhecidos. A formulação dos MLGs é muito flexível e engloba como casos especiais vários modelos largamente usados na literatura, tais como: modelo logístico linear, modelo log-linear, modelo probit para proporções, polinômios inversos e modelos de regressão com distribuições gama e normal inversa. O método escore de Fisher é, usualmente, adotado para estimar os parâmetros lineares nos MLGs.

O vetor β de parâmetros é usualmente estimado pelo método de máxima verossimilhança. A inferência baseada na verossimilhança visa a construir

procedimentos ou regras apropriadas de alguma natureza científica baseando-se num certo conjunto de dados, tais como: obter uma estimativa do vetor β de parâmetros desconhecidos, construir uma região de valores possíveis para β que tenha uma confiabilidade especificada ou decidir sobre um vetor especificado previamente concebido para β. Neste sentido, as atividades fim da inferência de verossimihança são: a estimação, a construção de regiões de confiança e o desenvolvimento de testes de hipóteses. Neste artigo, não são discutidas estas atividades, pois o objetivo principal é mostrar a aplicabilidade prática de resultados assintóticos recentemente publicados para os MLGs.

Seja $\widehat{\beta}$ a estimativa de máxima verossimilhança (EMV) de β para um MLG arbitrário. Os livros de Cordeiro (1986) e McCullagh e Nelder (1989) fazem uma boa revisão dos MLGs e descrevem algumas propriedades básicas das EMVs. Em especial, as propriedades de consistência, suficiência e invariância de $\widehat{\beta}$. As EMVs $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$ são, em geral, viesadas e muitas vezes ignoradas na prática, justificando-se por terem magnitudes desprezíveis quando comparadas aos erros padrão. Cordeiro e McCullagh (1991) destacam a importância do cálculo dos vieses de segunda-ordem das EMVs, mostrando que esses vieses podem ser apreciáveis em pequenas amostras. Recentemente, Cordeiro e Barroso (2007) determinaram os vieses de terceira-ordem de $\widehat{\beta}$ em MLGs, ou seja, de ordem $O\left(n^{-2}\right)$. Entretanto, em amostras moderadas, alguns estudos de simulação desses autores têm indicado que os valores desses vieses podem ser bem inferiores do que aqueles dos vieses de segunda-ordem.

A matriz de covariância de ordem $O\left(n^{-1}\right)$ de $\widehat{\beta}$, também, pode apresentar erros apreciáveis como uma aproximação da covariância exata de $\widehat{\beta}$. Para obter uma melhor aproximação para essa covariância exata, Cordeiro (2004) apresenta uma expansão matricial para a matriz de covariância até quantidades de ordem $O\left(n^{-2}\right)$.

Deve-se salientar que poucos artigos foram publicados com aplicações de resultados assintóticos nos MLGs. Com o surgimento de sistemas especialistas que possibilitam realizar operações matriciais, tais como, R, MatLab, SAS e Ox, a aplicabilidade desses resultados assintóticos tenderão a crescer no País e, espera-se, que um primeiro passo nessa direção esteja sendo dado com esse artigo aplicado à área industrial. Acredita-se que tanto nas áreas tecnológicas quanto nas áreas médicas e biológicas, essas aplicações serão rotineiras no médio prazo.

Em muitos processos industriais, em especial, no ramo de papel e celulose, as condições de operação necessitam ser caracterizadas com base em pequenas amostras que permitam a tomada de decisões em tempo real. O grau de refino, ou seja, o grau de drenabilidade da polpa celulósica, que é considerado o mais importante componente do processo da fabricação de papel, pois passa ao mesmo a parte preponderante de suas características finais, sendo esse um tratamento mecânico dado às fibras em suspensão com o intuito de modificar sua estrutura e melhorar as características das fibras do papel. Estas modificações são irreversíveis e as propriedades mais afetadas são as resistências à tração e à tensão, entre outras.

Para aplicação dos resultados assintóticos de Cordeiro e McCullagh (1991) e Cordeiro (2004) na área industrial, foram coletados dados da Klabin Papéis Monte

Alegre durante o processo de fabricação de papel no ano de 2003.

Com o objetivo de avaliar o viés de $\widehat{\beta}$ de ordem $O\left(n^{-1}\right)$ bem como a expansão da matriz de covariância de $\widehat{\beta}$ até ordem $O\left(n^{-2}\right)$, ajustou-se um modelo gama com função de ligação logarítmica aos dados do grau do refino das fibras de celulose, considerando como variáveis explicativas a carga do refino e a condutividade elétrica. A distribuição gama foi adotada por conta da assimetria e positividade da variável resposta e a ligação logarítmica garante a positividade da sua média. Foram coletados dados a partir de duas amostras aleatórias de tamanhos $n = 30$ e $n = 800$ durante todo o processo de produção.

A estrutura do artigo é como se segue. Na Seção 2 definem-se as variáveis e a modelagem estatística e apresentam-se fórmulas matriciais para calcular a estimativa corrigida e a expansão da matriz de covariância de $\widehat{\beta}$ até ordem $O\left(n^{-2}\right)$. Na Seção 3 calculam-se os vieses de ordem $O\left(n^{-1}\right)$ das estimativas dos parâmetros lineares do modelo gama com ligação logarítmica ajustado ao refino das fibras de celulose, além das estimativas das matrizes de covariância de $\widehat{\beta}$ de primeira e segunda ordens. Compararam-se, ainda, essas estimativas considerando os dois tamanhos amostrais. Na Seção 4 apresentam-se alguns resultados de simulação relativos às covariâncias de primeira e segunda ordens das EMVs em MLGs. O estudo de simulação mostra que a covariância de segunda-ordem é mais próxima da matriz de covariância exata da EMV do que a matriz de covariância usual dada pela inversa da matriz de informação. Finalmente, a Seção 5 apresenta algumas conclusões e sugestões para estudos a posteriori.

2 Métodos assintóticos

Adotou-se um delineamento inteiramente casualizado para amostras com tamanhos de 30 e 800 observações, que foram coletadas e registradas na Klabin Papéis Monte Alegre no ano de 2003 durante o processo de produção.

O refino da celulose é um processo aplicado à polpa de celulose que visa à hidratação das fibras, possibilitando um melhor entrelaçamento das mesmas no processo de desagregação de celulose na máquina de papel. A variável grau de refino das fibras foi considerada como resposta e as covariáveis carga do refinador (medida em percentual) e condutividade elétrica da solução que é a recíproca da resistência do papel (medida em ohms-metro) foram controladas no processo de refino. Um dos objetivos do estudo é usar o modelo ajustado para predizer a qualidade do papel em tempo real, o que significa no processo de produção observar pequenas amostras e estimar através do modelo, a qualidade do papel para decidir se essa qualidade é satisfatória ou não.

Para atingir esse objetivo, duas amostras aleatórias de tamanhos diferentes foram retiradas do processo produtivo, em que ajustou-se um modelo gama com ligação logaritmo aos dados das duas amostras e compararam-se as estimativas corrigidas com as EMVs usuais, bem como, as matrizes de covariância de $\widehat{\beta}$ de ordens $O\left(n^{-1}\right)$ e $O\left(n^{-2}\right)$, respectivamente.

Inicialmente, analisou-se a distribuição da variável resposta por meio de uma análise exploratória dos dados que evidenciou uma assimetria e, além disso, como o grau de refino das fibras só assume valores positivos, adotou-se a distribuição gama para se ajustar ao fenômeno. A função de ligação logaritmo garante a positividade da média.

O MLG tem como componente aleatória o grau do refino das fibras de celulose (^oSR), denotada por Y_i, com $i = 1, \ldots, n$, com função densidade de probabilidade gama pertencente à família (1) com média $E(Y_i) = \mu_i$ e parâmetro de dispersão ϕ^{-1}, isto é, a função de variância é especificada por $V_i = \phi^{-1}\mu_i^2$.

A componente sistemática do MLG é

$$\eta_i = \log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i},$$

com as variáveis explicativas x_{1i} e x_{2i} representando a carga do refino e a condutividade elétrica, respectivamente.

A EMV foi obtida pela diretiva GENMOD do SAS que usa o algoritmo iterativo descrito em Nelder e Wedderburn (1972). Assim, a EMV é a solução de mínimos quadrados reponderados dado pela equação iterativa

$$\hat{\beta} = \left(X^T W X\right)^{-1} X^T W \mathbf{z}, \tag{2}$$

e, no problema em pauta, a matriz modelo de ordem $n \times 3$ iguala $X = [1, x_{1i}, x_{2i}]$, a matriz de pesos de ordem n é $W = diag\{w_1, \ldots, w_n\}$ com $w_i = V_i^{-1}(d\mu_i/d\eta_i)^2 = \mu_i^2$ e o vetor da variável resposta ajustada de ordem n é z cuja i-ésima componente iguala

$$z_i = \eta_i + (y_i - \mu_i)\frac{d\eta_i}{d\mu_i} = \eta_i + \frac{(y_i - \mu_i)}{\mu_i}.$$

O viés $B(\widehat{\beta})$ de $\widehat{\beta}$ de ordem $O\left(n^{-1}\right)$ em notação matricial é apresentado em Cordeiro e McCullagh (1991) como

$$B(\widehat{\beta}) = \left(X^T W X\right)^{-1} X^T W \xi, \tag{3}$$

sendo

$$F = diag\left\{V^{-1}\frac{d\mu}{d\eta}\frac{d^2\mu}{d\eta^2}\right\}$$

uma matriz diagonal de ordem n que depende das duas primeiras derivadas da função de ligação e da função de variância V e ξ é um tipo de variável resposta modificada dada por

$$\xi = \frac{-1}{2\phi} W^{-1} Z_d F 1,$$

em que $Z_d = diag\{z_{11}, \ldots, z_{nn}\}$ é uma matriz diagonal definida com os elementos da diagonal da matriz $Z = X\left(X^T W X\right)^{-1} X^T$ de ordem $n \times n$ e 1 é um vetor $n \times 1$ de uns. Observe-se que Z representa a estrutura de covariância assintótica do preditor linear estimado $\widehat{\eta}$ exceto pelo multiplicador ϕ. A equação (3) mostra que

o viés de segunda-ordem de $\widehat{\beta}$ é obtido pela simples regressão linear ponderada de ξ sobre a matriz modelo X com pesos W, sendo que tanto ξ quanto W devem ser estimados em $\widehat{\beta}$.

Logo, a estimativa de máxima verossimilhança corrigida (EMVC) pelo viés de ordem $O\left(n^{-1}\right)$ é dada por

$$\widetilde{\beta} = \widehat{\beta} - \widehat{B}(\widehat{\beta}), \tag{4}$$

em que $\widehat{B}(\widehat{\beta})$ é o valor da função paramétrica $B(\widehat{\beta})$ avaliada na EMV $\widehat{\beta}$. Espera-se que as estimativas em $\widetilde{\beta}$ tenham vieses mais reduzidos do que aquelas estimativas não-corrigidas correspondentes em $\widehat{\beta}$. Isso pode ser comprovado em resultados de simulação que constam na bibliografia citada em Cordeiro e Barroso (2007).

Cordeiro (2004), a partir de expansões deduzidas por Peers e Iqbal (1985), obteve a matriz de covariância de $\widehat{\beta}$ até termos de ordem n^{-2}, denotada aqui por $Cov_2(\widehat{\beta})$, que em notação matricial pode ser escrita como

$$Cov_2(\widehat{\beta}) = Cov_1(\widehat{\beta}) + \phi^{-2} P \, \Lambda \, P^T, \tag{5}$$

em que $Cov_1(\widehat{\beta}) = \phi^{-1} \left(X^T \, W \, X\right)^{-1}$ é a covariância de ordem $O\left(n^{-1}\right)$ obtida como a matriz inversa da informação de Fisher, e

$$\Lambda = \left\{ H Z_d + \left(\frac{3}{2} FF + GF - GG \right) Z^{(2)} + (F + G) \, Z Z_d F \right\}$$

é uma matriz diagonal de ordem n. Aqui, G e H são, também, matrizes diagonais de ordem n que dependem das duas primeiras derivadas da função de ligação e da função de variância com sua primeira derivada $V^{(1)} = dV/d\mu$, dadas por

$$G = diag \left\{ V^{-1} \frac{d\mu}{d\eta} \frac{d^2\mu}{d\eta^2} - \frac{V^{(1)}}{V^2} \left(\frac{d\mu}{d\eta} \right)^2 \right\}$$

e

$$H = diag \left\{ V^{-1} \frac{d\mu}{d\eta} \frac{d^3\mu}{d\eta^3} - V^{-1} \left(\frac{d\mu}{d\eta} \right)^2 \left(\frac{d^2\mu}{d\eta^2} \right) + \frac{\left(V^{(1)}\right)^2}{V^3} \left(\frac{d\mu}{d\eta} \right)^4 \right\}.$$

Ainda, $Z^{(2)} = Z \odot Z$, com $\odot$ representando o produto de Hadamard e $P = \left(X^T \, W \, X\right)^{-1} X^T$. A matriz $Cov_1(\widehat{\beta})$ é chamada de matriz de covariância de *primeira-ordem* enquanto $Cov_2(\widehat{\beta})$ é denominada matriz de covariância de *segunda-ordem*.

As fórmulas (3), (4) e (5) foram implementadas no SAS, uma vez que esse sistema computacional dispõe de operações matriciais. Entretanto, outros *software* poderiam ser utilizados como Ox, MatLab e R.

3 Aplicação

A EMV decorrente do processo iterativo de estimação por máxima verossimilhança foi obtida pela diretiva GENMOD do aplicativo SAS. O modelo

de regressão gama de média μ_i e parâmetro de precisão ϕ com as covariáveis carga do refino e a condutividade elétrica, apresentaram um nível descritivo p menor que $0,0001$. Esse fato revela um modelo bem ajustado, uma vez que a *scaled deviance* foi próxima de um (vide Hinde e Demétrio, 1998, p.38).

Na seqüência, calcularam-se os vieses de primeira ordem relativos às amostras de tamanhos 30 e 800, para evidenciar que a correção se faz necessária em pequenas amostras.

3.1 Estimativas não-corrigidas e corrigidas

Na Tabela 1 encontram-se as estimativas não-corrigidas e corrigidas dos parâmetros β_0, β_1 e β_2 do modelo ajustado para uma amostra de tamanho $n = 30$. A coluna Quociente avalia a magnitude do viés em relação ao erro-padrão, isto é, o valor da razão (viés/erro-padrão). Observa-se que os vieses são apreciáveis quando comparados aos seus respectivos erros-padrão, e apresentam intervalo de variação para o Quociente relativamente grande, ou seja, de 144% a 917%, evidenciando a necessidade da obtenção do viés.

Tabela 1 - Estimativas Não-Corrigidas e Corrigidas para n=30

Parâmetro	Estimativa Não-Corrigida	Erro-padrão	Viés	Estimativa Corrigida	Quociente (%)
β_0	2,2239	0,2211	-1,7340	3,9579	784
β_1	0,0060	0,0023	0,0211	-0,0150	917
β_2	0,0343	0,0156	0,0226	0,0118	144

A Tabela 2 faz para $n = 800$ o mesmo que a Tabela 1. Para esta amostra, observa-se da Tabela 2 que as magnitudes dos vieses são bem menores quando comparadas aos erros padrão, isto é, apresentam um intervalo de variação menor, como seria de se esperar, em torno de 7% a 33%, evidenciando que para amostras grandes as correções dos vieses de ordem $O\left(n^{-1}\right)$ são dispensáveis.

Tabela 2 - Estimativas Não-Corrigidas e Corrigidas para n=800

Parâmetro	Estimativa Não-Corrigida	Erro-padrão	Viés	Estimativa Corrigida	Quociente (%)
β_0	2,4591	0,0431	-0,0029	2,4619	7
β_1	0,0050	0,0003	0,0001	0,0049	33
β_2	0,0174	0,0041	-0,0008	0,0182	20

3.2 Matrizes de covariância de primeira e segunda ordens

As matrizes de covariância de ordens $O\left(n^{-1}\right)$ e $O\left(n^{-2}\right)$ de $\widehat{\beta}$ foram obtidas através da diretiva GENMOD e do IML do SAS, respectivamente. Para avaliar a

relação da magnitude entre $Cov_1(\widehat{\beta})$ e $Cov_2(\widehat{\beta})$ adotou-se a seguinte medida:

$$medida = abs\left[\left(\frac{Cov_1(\widehat{\beta})}{Cov_2(\widehat{\beta})}\right) * 100 - 100\right]. \quad (6)$$

Para a amostra de tamanho igual a $n = 30$, a matriz de covariância de ordem $O\left(n^{-1}\right)$ é

$$Cov_1(\widehat{\beta}) = \begin{array}{c} \\ \widehat{\beta} \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{array} \begin{array}{c} \begin{array}{ccc} \widehat{\beta} & \widehat{\beta}_1 & \widehat{\beta}_2 \end{array} \\ \begin{pmatrix} 0,04890 & 0,00040 & 0,00232 \\ & 5,30 \times 10^{-6} & 1,6 \times 10^{-6} \\ & & 0,00024 \end{pmatrix} \end{array}.$$

A matriz de covariância calculada incluindo o termo de ordem $O\left(n^{-2}\right)$ é

$$Cov_2(\widehat{\beta}) = \begin{array}{c} \\ \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{array} \begin{array}{c} \begin{array}{ccc} \widehat{\beta}_0 & \widehat{\beta}_1 & \widehat{\beta}_2 \end{array} \\ \begin{pmatrix} 0,03383 & 0,00024 & 0,00187 \\ & 3,50 \times 10^{-6} & 8,00 \times 10^{-7} \\ & & 0,00021 \end{pmatrix} \end{array}.$$

Comparando $Cov_1(\widehat{\beta})$ e $Cov_2(\widehat{\beta})$ pela medida (6), obtém-se uma variação entre um mínimo de 12% e um máximo de 49%, revelando, assim, uma grande variação entre essas matrizes de covariância.

O mesmo procedimento foi adotado para amostra de tamanho igual a $n = 800$, cuja matriz de covariância de $\widehat{\beta}$ de ordem $O\left(n^{-1}\right)$ é

$$Cov_1(\widehat{\beta}) = \begin{array}{c} \\ \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{array} \begin{array}{c} \begin{array}{ccc} \widehat{\beta}_0 & \widehat{\beta}_1 & \widehat{\beta}_2 \end{array} \\ \begin{pmatrix} 0,001861 & -7,86 \times 10^{-6} & -0,000147 \\ & 1,2 \times 10^{-7} & -6,654 \times 10^{-8} \\ & & 0,00017 \end{pmatrix} \end{array},$$

enquanto a matriz de covariância de $\widehat{\beta}$ determinada até a ordem $O\left(n^{-2}\right)$ é

$$Cov_2(\widehat{\beta}) = \begin{array}{c} \\ \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{array} \begin{array}{c} \begin{array}{ccc} \widehat{\beta}_0 & \widehat{\beta}_1 & \widehat{\beta}_2 \end{array} \\ \begin{pmatrix} 0,001852 & -7,64 \times 10^{-6} & -0,000146 \\ & 1,2 \times 10^{-7} & -6,606 \times 10^{-8} \\ & & 0,000017 \end{pmatrix} \end{array}.$$

Para o tamanho de amostra $n = 800$, observa-se uma variação de 0% a $0,7\%$, isto é, praticamente não há diferença entre as duas matrizes. Nestes termos, a própria inversa da matriz de informação de Fisher é uma excelente aproximação para a estrutura de covariância exata de $\widehat{\beta}$.

4 Simulação

Nesta seção realiza-se um estudo de simulação para mostrar a aplicação prática das fórmulas das matrizes de primeira e segunda ordens dadas na equação (5) no ajuste de um modelo gama com ligação logarítmo como aquele adotado na Seção 3. A estrutura linear do modelo é definida por duas covariáveis e três parâmetros

$$\log(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i},$$

em que as covariáveis x_1 e x_2 foram escolhidas aleatoriamente de uma distribuição uniforme em $(0, 1)$ e os seus valores foram mantidos constantes para amostras de mesmo tamanho. Os valores verdadeiros dos parâmetros lineares foram tomados como $\beta_0 = 2, \beta_1 = 2$ e $\beta_2 = 3$ para os casos em que $n = 20$ and $n = 40$. A variável resposta foi gerada de uma distribuição gama com média μ_i dada pela equação acima e parâmetro de precisão igual a $\phi = 2$.

As simulações foram realizadas usando o SAS com $10,000$ repetições para cada valor de n. Para a i-ésima simulação, o modelo gama com ligação logaritmo foi ajustado e calculou-se a EMV $\widehat{\beta}^{(i)}$, as suas covariâncias de primeira e segunda ordens a partir da equação (5), com ambas sendo estimadas em $\widehat{\beta}^{(i)}$. Esse processo foi repetido para $i = 1, \ldots, 10,000$. Sejam, então, $[Cov_1(\widehat{\beta})]_i$ e $[Cov_2(\hat{\beta})]_i$ as covariâncias de $\widehat{\beta}^{(i)}$ avaliadas na estimativa $\widehat{\beta}^{(i)}$ obtidas da i-ésima simulação. As duas primeiras entradas de cada cela da Tabela 3 dão para $n = 20$ $\frac{1}{10,000}\sum_{i=1}^{10,000}[Cov_1(\widehat{\beta})]_i$ e $\frac{1}{10,000}\sum_{i=1}^{10,000}[Cov_2(\widehat{\beta})]_i$, i.e., as médias amostrais de $Cov_1(\widehat{\beta})$ e $Cov_2(\widehat{\beta})$ nas $10,000$ replicações. A terceira entrada de cada cela da Tabela 3 se refere à covariância amostral de $\widehat{\beta}^{(1)}, \ldots, \widehat{\beta}^{(10,000)}$. A Tabela 4 faz o mesmo para $n = 40$.

Os valores nas Tabelas 3 e 4 evidenciam que os termos de segunda-ordem das covariâncias de $\widehat{\beta}$ não devem ser ignorados em amostras de tamanho moderado. Em todos os casos, as covariâncias de segunda-ordem são mais próximas das covariâncias amostrais do que aquelas covariâncias de primeira-ordem. Nestes termos, conclui-se que a expansão (5) é uma melhor aproximação para o valor verdadeiro da matriz de covariância $Cov(\widehat{\beta})$ de $\widehat{\beta}$ do que a matriz inversa da informação. Um tamanho de amostra grande é necessário para que a aproximação de primeira-ordem para a covariância exata $Cov(\widehat{\beta})$ seja adequada. Entretanto, como esperado, os valores dos termos em $Cov_2(\widehat{\beta})$ tendem a ser praticamente iguais àqueles correspondentes termos de $Cov_1(\widehat{\beta})$ para valores moderados de n. A sinalização dada pelo termo de ordem $O\left(n^{-2}\right)$ em (5) está correta em todos os casos das Tabelas 3 e 4 sendo na direção de conduzir a matriz de covariância estimada de segunda-ordem para ficar mais próxima da covariância amostral.

Em resumo, essas simulações mostram que as covariâncias de segunda-ordem das EMVs em MLGs podem ser pronunciadas em amostras pequenas e, em geral, melhoram a aproximação em relação ao valor verdadeiro de $Cov(\widehat{\beta})$.

Tabela 3 - $Cov_1(\widehat{\beta})$, $Cov_2(\widehat{\beta})$ e Covariâncias amostrais para $n = 20$

	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$
$\widehat{\beta}_1$	1.27	2.03	−1.23
	1.39	2.29	−1.05
	1.43	2.34	−0.90
$\widehat{\beta}_2$		2.32	1.67
		2.57	1.72
		2.61	1.79
$\widehat{\beta}_3$			1.54
			1.62
			1.68

Tabela 4 - $Cov_1(\widehat{\beta})$, $Cov_2(\widehat{\beta})$ e Covariâncias amostrais para $n = 40$

	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$
$\widehat{\beta}_1$	1.09	1.97	−1.00
	1.21	2.13	−0.96
	1.25	2.18	−0.91
$\widehat{\beta}_2$		2.26	1.51
		2.38	1.55
		2.43	1.60
$\widehat{\beta}_3$			0.96
			1.14
			1.16

Conclusões

Os MLGs estendem os modelos normais lineares e possuem as seguintes características: a distribuição de probabilidade da variável resposta Y pertence à família exponencial de distribuições cuja média está associada a um preditor linear representativo de variáveis explicativas através de uma função de ligação, possivelmente não-linear num conjunto de parâmetros desconhecidos. Na formulação de um MLG deve-se escolher a distribuição da probabilidade da variável resposta, as variáveis explicativas que formam a matriz modelo e a função de ligação relacionando a média da distribuição com o preditor linear.

A partir dos anos 80, alguns artigos foram publicados mostrando que a correção das EMVs pelo viés em modelos de regressão é relevante, fato esse evidenciado em estudos de simulação, principalmente, em amostras de tamanho pequeno.

Entretanto, as estimativas corrigidas, cujos vieses são de ordem $O\left(n^{-2}\right)$, têm sido ainda muito pouco utilizados na prática, uma vez que não estão disponibilizadas nos aplicativos estatísticos.

Infelizmente, o uso prático dessas fórmulas não tem sido ainda objeto de estudo. Com o objetivo de avaliar a viabilidade prática dessas fórmulas, modelou-se o grau de refino das fibras de celulose advindo do processo de fabricação do papel. Ressalte-se que a tomada de decisão sobre a qualidade do papel deve ser tomada em tempo real e, portanto, os tamanhos das amostras usualmente são pequenos. Para isto, foram analisadas dados de celulose do papel com duas amostras aleatórias de tamanhos $n = 30$ e $n = 800$ através de um modelo gama com função de ligação logarítmica.

O modelo ajustado poderá ser utilizado na indústria para avaliação do processo de produção do papel para estimar o valor médio do grau de refino, principalmente, utilizando as estimativas corrigidas pelos vieses de ordem $O(n^{-1})$. Os resultados evidenciaram que para o tamanho de amostra igual a $n = 30$ as matrizes de covariância de primeira e segunda ordens apresentam uma grande diferença e, portanto, a matriz de covariância de segunda-ordem deve ser considerada na inferência, enquanto que para o tamanho de amostra igual a $n = 800$, as duas matrizes são praticamente iguais e a covariância de ordem superior não se faz necessária.

Diante do exposto, sugere-se sempre adotar estimativas de segunda ordem quando as amostras são de tamanho pequeno a moderado, pois isso significa na indústria minimizar tempo e custos.

A possível aplicabilidade dos métodos assintóticos apresentados no artigo em outras áreas pode ser feita "mutatis mutandis" ao desenvolvimento deste artigo.

Não foram feitas análises de diagnóstico e dos resíduos e gráficos do tipo envelope, pois o objetivo principal é mostrar a aplicabilidade de alguns resultados assintóticos à indústria de papel. Entretanto, essas análises poderão ser feitas usando as estimativas corrigidas seguindo a análise convencional descrita em Cordeiro (1986) e McCullagh e Nelder (1989), bastando substituir as EMVs pelas estimativas corrigidas.

Foram feitos estudos de simulação que comprovam a maior adequabilidade da covariância de segunda-ordem da EMV em MLGs em termos de aproximação para a covariância exata do que da covariância usual dada pela inversa da matriz de informação.

Agradecimentos

Os autores são gratos aos professores Dr. Ivo Neitzel e Dr. Carlos E. S. Alves do Departamento de Engenharia Química da Universidade Estadual de Maringá pela disponibilização dos dados e aos dois revisores por sugestões que melhoraram muito uma versão preliminar do artigo.

SANTANA, R. G.; UDO, M. C. T.; PREVIDELLI, I. T. S.; CORDEIRO, G. M. Application of the covariance matrix of second-order of the maximum likelihood estimates in the industry. *Rev. Bras. Biom.,* São Paulo, v.25, n.3, p.105-117, 2007.

▪ *ABSTRACT: Cordeiro and McCullagh (1991) and Cordeiro (2004) present simple matrix formulas to evaluate the biases of order* $O\left(n^{-1}\right)$ *of the maximum likelihood estimates in generalized linear models and the covariance matrix of these estimates up to order* n^{-2}*, respectively. However, these formulas had still not been explored in a practical context and, with this objective, we analyzed a cellulose data set through a gamma model with logarithmic link function. The implementation of the formulas of the corrected estimates and the covariance matrix is simple and was made using the computational system SAS. Some simulations show the evidence that for small samples it is important to use the covariance matrix of the maximum likelihood estimates in generalized linear models up to order* n^{-2}*.*

▪ *KEYWORDS: Bias of the estimate; corrected covariance matrix; corrected estimate; information matrix; maximum likelihood estimate*

Referências

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

CORDEIRO, G. M. Second-order covariance matrix of maximum likelihood estimates in generalized linear models. *Statist. Prob. Letters*, v.66, p.153-160, 2004.

CORDEIRO, G. M.; BARROSO, L. P. A third-order bias corrected estimate in generalized linear models. *Test*, v.16, p.76-89, 2007.

CORDEIRO, G. M.; McCULLAGH, P. Bias correction in generalized linear models. *J. R. Stat. Soc. B*, v.53, p.629-643, 1991.

HINDE, J.; DEMÉTRIO, C. G. B. Overdispersion: models and estimation. In: SINAPE, 13., 1998, Caxambú. *Anais...* Caxambú: SBMAC, 1998. 73p.

McCULLAGH, P.; NELDER, J. *Generalized linear models*. 2.ed. London: Chapman and Hall, 1989. 511p.

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

PEERS, H. A.; IQBAL, M. Asymtotic expansions for confidence limits in the presence of nuisance of parameters, whith applications. *J. R. Statistit. Soc. B*, v.47, p.547-554, 1985.

Recebido em 05.03.2007.

Aprovado após revisão em 23.12.2007.