Um estudo computacional do efeito das componentes obtidas na decomposição dos valores singulares com predições de modelos GEE

Haiany Aparecida Ferreira[1], Matheus Feres Freitas[2], Carla Regina Guimarães Brighenti[3], Marcelo Ângelo Cirillo[4], Kelly Pereira de Lima[5]

1. Introdução

A análise de dados categóricos, em geral, é feita considerando dados agrupados em estrutura de uma tabela de contingência, a qual possibilita ajustar diferentes modelos. A considerar a distribuição conjunta entre duas ou mais variáveis, respectivamente, caracterizando as tabelas bidimensionais e multidimensionais.

O ajuste de modelos para essas tabelas poderá ser feito considerando as distribuições marginais. Nesse caso, verifica-se o uso do método de equações de estimação generalizadas, considerando a distribuição binomial. Entretanto, a modelagem da proporção poderá ser dada em duas possibilidades de expressar a proporção marginal, isto é, em relação ao total de cada linha ou coluna.

Dada a motivação da aplicabilidade do modelo binomial, naturalmente utiliza-se as funções de ligação, dadas pelas funções logit e complemento log-log. A seleção de uma destas funções é uma etapa importante para que o modelo seja bem ajustado e as estimativas dos coeficientes sejam interpretáveis.

Para a organização dos dados em uma tabela de contingência uma possível estrutura é na forma de um delineamento em blocos ao acaso com respostas multivariadas, a qual possibilita um estudo quanto ao comportamento nos efeitos entre os blocos.

Diante disso, o objetivo deste trabalho consiste em investigar o efeito do grau de correlação nas estimativas das somas de quadrados das componentes de aditividade e diferença entre os blocos, obtidas através de simulações Monte Carlo.

2. Material e métodos

As simulações foram feitas no software R Core Team (2016) considerando uma tabela de contingência, estruturada por dois blocos, cinco tratamentos e três variáveis respostas, o delineamento envolvido no procedimento de simulação é dado pela matriz com a seguinte codificação:

[1]Doutoranda em Estatística e Experimentação Agropecuária na DES/UFLA. email: *haianyferreira@yahoo.com.br*.
[2]Doutorando em Estatística e Experimentação Agropecuária na DES/UFLA. email: *matheus712@hotmail.com*.
[3]Professora Associada II, DZOO/UFSJ. email: *carlabrighenti@ufsj.edu.br.*
[4]Professor Associado III, DES/UFLA. email: *macufla@dex.ufla.br.*
[5]Doutoranda em Estatística e Experimentação Agropecuária no DES/UFLA. email: *kelly_limaadm@hotmail.com.*

$$X = \begin{pmatrix} 1 & -2 & 1 \\ 1 & -1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & -2 & 0 \\ 1 & -1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & -2 & -1 \\ 1 & -1 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & -1 \\ 1 & 2 & -1 \\ -1 & -2 & 1 \\ -1 & -1 & 1 \\ -1 & 0 & 1 \\ -1 & 1 & 1 \\ -1 & 2 & 1 \\ -1 & -2 & 0 \\ -1 & -1 & 0 \\ -1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 2 & 0 \\ -1 & -2 & -1 \\ -1 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & 1 & -1 \\ -1 & 2 & -1 \end{pmatrix}.$$

O vetor de variáveis respostas para cada tratamento codificado em -2, -1, 0, 1, 2 foi gerado seguindo a distribuição binomial correlacionada, mencionada por Cirillo e Ramos (2014) e adaptada nesse trabalho, com as seguintes especificações descritas a seguir.

Dado o vetor $Y = (Y_1, \ldots, Y_5)$, em que cada componente representa o número de ocorrências no j-ésimo tratamento $(j = 1, ..., 5)$ relacionado ao vetor $\underset{\sim}{\pi} = (\pi_1, \ldots, \pi_J)$ em que, π_j correspondeu a

probabilidade de sucesso do modelo binomial correlacionado. De modo que, $y_j \sim BC\left(\mathrm{n}_j, \pi_j, \rho\right)$ com distribuição de probabilidade dada por:

$$\mathrm{P}\left(\mathrm{Y}_j \middle| \mathrm{n}_j, \pi_j, \rho\right) = \binom{\mathrm{n}_j}{y_j} \pi_j^{y_j} \left(1-\pi_j\right)^{\mathrm{n}_j - y_j} \left(1-\rho\right) I_{\theta_1}\left(y_j\right) + \pi^{y_j/\mathrm{n}_j} \rho I_{\theta_2}\left(y_j\right),$$

sendo ρ a taxa de mistura entre as distribuições binomial $\left(n_j, \pi_j\right)$ com probabilidade $\left(1-\rho\right)$ e uma distribuição bernoulli modificada representada pela variável $BeM\left(\pi\right)$ assumindo 0 ou n_j valores com probabilidade ρ, fixado o vetor π = (0,5; 0,5; 0,5; 0,5; 0,5). Assim, os valores paramétricos utilizados na geração das observações binominais encontram-se na Tabela 1.

Tabela 1 - Valores paramétricos utilizados como cenários para gerar as tabelas de contingência.

Cenário	n	Grau de correlação (ρ)
1		0,2
2	50	0,5
3		0,8
4		0,2
5	100	0,5
6		0,8
7		0,2
8	150	0,5
9		0,8

Fonte: Do autor (2019).

Com a realização deste procedimento, a correlação ρ entre as variáveis "linha" (Tabela 1), justifica supor o efeito de superdispersão (STONER; LEROUX, 2002). Assim, torna-se inapropriado o uso de modelos generalizados, e para tal problema, como alternativa utilizou-se o modelo GEE para as funções de ligação logit e complemento log-log. Dado pela solução do sistema

$$\sum_{i=1}^{N} \left(\frac{\partial \mu_i\left(\beta\right)}{\partial \beta} \right) V_i^{-1} \left(Y_i - \hat{\mu}_i\left(\beta\right)\right) = 0,$$

em que, $\hat{\mu}_i\left(\beta\right)$ correspondeu ao vetor de médias ajustados considerando os modelos binomiais com as funções de ligação logit (1) e cloglog (2)

$$\log\left(\frac{\mu_{ijk}}{1-\mu_{ijk}}\right)=\eta_{ijk}\ ;\ \mu_{ijk}=\frac{\exp\left(\eta_{ijk}\right)}{1+\exp\left(\eta_{ijk}\right)} \quad (1)$$

$$\log\left\{-\log\left(1-\mu_{ijk}\right)\right\}=\eta_{ijk};\ \ \mu_{ijk}=1-\exp\left\{-\exp\left(\eta_{ijk}\right)\right\}. \quad (2)$$

Diante disso, após a execução de 1000 realizações Monte Carlo em ambos os modelos, em que cada realização corresponde a uma tabela de contingência estruturada pelo delineamento blocos ao acaso computou-se as médias das somas de quadrados do efeito de blocos das componentes A+B e A-B com o propósito de investigar o efeito dessas componentes em relação a amostras binomiais correlacionadas.

3. Resultados e discussão

Os resultados descritos na Tabela 2 correspondem à média das estimativas das componentes A+B e A-B em 1000 realizações Monte Carlo, obtidas por meio da decomposição de valores singulares, aplicada nas tabelas geradas com frequência, em diferentes graus de correlação em relação aos níveis organizados em “linhas”, seguindo o layout da Tabela 1 para cada bloco.

Tabela 2 - Cenários de simulação e média das estimativas das componentes A+B e A-B obtidas pela decomposição de valores singulares.

Modelo GEE-Logit			
n	Grau de correlação (ρ)	A+B	A-B
50	0,2	0,0271	0,0055
	0,5	0,0178	0,0037
	0,8	0,0073	0,0023
100	0,2	0,0151	0,0024
	0,5	0,0086	0,0018
	0,8	0,0036	0,0013
150	0,2	0,0094	0,0017
	0,5	0,0059	0,0012
	0,8	0,0022	8e-04
Modelo GEE-Cloglog			
n	Grau de correlação (ρ)	A+B	A-B
50	0,2	0,0287	0,0053
	0,5	0,0174	0,0034
	0,8	0,0072	0,0024
100	0,2	0,0141	0,0027
	0,5	0,0089	0,0021
	0,8	0,0033	0,0012
150	0,2	0,0091	0,0018
	0,5	0,0060	0,0013
	0,8	0,0024	7e-04

Fonte: Do autor (2019).

Em se tratando do modelo GEE-logit, notou-se que o aumento do grau de correlação ρ, resultou em uma redução na estimativa da componente A+B, para todos os tamanhos amostrais avaliados. No tocante a componente A-B, este comportamento foi menos expressivo, o que conduz a afirmar que a decomposição da variância amostral, pelos dois primeiros componentes principais extraídos do efeito de igualdade entre os blocos (A-B) é mais robusto em relação ao aumento do grau de correlação, independente do tamanho amostral considerado.

Especificamente ao modelo GEE-cloglog, evidenciou-se que o comportamento das estimativas é análogo ao do modelo logit.

4. Conclusões

Portanto, o aumento do grau de correlação resultou na redução das somas de quadrados das componentes A+B e A-B, em todos os tamanhos amostrais para os modelos logit e cloglog. Como esperado o valor das somas de quadrados foi baixo em ambas as componentes. A citar que na diferença entre os blocos ele foi menor que na soma.

5. Agradecimentos

Agradecimentos à Cnpq e Fapemig - Edital Universal APQ -0242-16.

6. Referências bibliográficas

CIRILLO, M.A.; RAMOS, P. de S. Goodness-of-fit tests for modified multinomial logit model. Chilean Journal of Statistics, v.5,n.1,p.73-85,2014.

R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria, 2016. ISBN 3-900051-07-0. Disponível em: <http://www.R-project. org>.

STONER, J. A.; LEROUX, B. G. Analysis of clustered data: A combined estimating equations approach. Biometrika, v.89, n.3, p.567-578, 2002.