Modelos Marginais para Respostas Binárias com Estruturas Hierárquicas de Agrupamento

André Gabriel Ferreira Calaça da Costa

DISSERTAÇÃO APRESENTADA PARA A OBTENÇÃO
DO GRAU DE MESTRE EM ESTATÍSTICA PELA
UNIVERSIDADE FEDERAL DE MINAS GERAIS

Programa: Programa de Pós-graduação em Estatística
Orientador: Enrico Antonio Colosimo
Co-Orientador: Leila Amorim

Belo Horizonte, maio de 2013

Modelos Marginais para Respostas Binárias com Estruturas Hierárquicas de Agrupamento

André Gabriel Ferreira Calaça da Costa

Esta versão da dissertação contém as correções e alterações sugeridas pela banca durante a defesa da versão original do trabalho, realizada em 6 de maio de 2013.

Comissão Julgadora:

- Prof. Dr. Enrico Antonio Colosimo (Orientador)
- Prof. Dr. Julio da Motta Singer
- Prof. Dra. Denise Duarte
- Prof. Dr. Leo Heller

Agradecimentos

Gostaria de aproveitar essas poucas linhas para manifestar minha gratidão à Deus por ter me concedido a maior das oportunidades, à ter vindo a esse mundo. A minha esposa por ter sempre me estimulado a vencer todos os desafios. Ao meu filho, que recém chegado já é uma fonte inesgotável de estímulos para o meu desenvolvimento. Aos meus pais e irmãos, que sempre me cercaram com todo amor e afeto. Ao meu orientador Dr. Enrico Antônio Colosimo pela motivação, atenção e conselhos. A minha co-orientadora Dra. Leila Amorim por todo auxílio durante o trabalho. Aos amigos e professores do mestrado pelos conhecimentos transmitidos. Aos meus sócios pela grande amizade. Aos membros da banca, professores Julio da Motta Singer, Denise Duarte e Leo Heller pelas importantes contribuições dadas ao texto. Ao criador da Logosofia, Carlos Bernardo Gonzáles Pecotche que ampliou minha vida em seus múltiplos aspectos, tornando a experiência vivida no mestrado, ainda mais saborosa.

Para triunfar é necessário vencer, para vencer é necessário lutar, para lutar é necessário estar preparado, para estar preparado é necessário prover-se de uma grande inteireza de ânimo e de uma paciência a toda a prova. Isto requer, por sua vez, levar constantemente ao íntimo da vida o incentivo da suprema esperança de alcançar aquilo que se anela como culminação feliz da existência. (Carlos Bernardo Gonzáles Pecotche, Logosofia Ciência e Método)

Resumo

Estudos que envolvem respostas binárias com estruturas hierárquicas de agrupamento são frequentemente encontrados nas diversas áreas do conhecimento. Em modelos de regressão essas estruturas de agrupamento devem ser devidamente tratadas, uma vez que violam o pressuposto básico de independência das observações. Quando o foco principal do estudo está na estrutura da média, os Modelos Marginais de primeira ordem, ajustados pelo método GEE1, propostos por Liang e Zeger (1986), oferecem uma solução elegante por sua facilidade na interpretação e ausência de suposições distribucionais. Entretanto com o método GEE1, não é possível acomodar satisfatoriamente estruturas hierárquicas ou múltiplas estruturas de agrupamento, podendo ocasionar perda de eficiência na estrutura da média. Para acomodar de forma satisfatória estruturas hierárquicas de agrupamento, o GEE1 foi estendido para o GEE2 (Prentice, 1988) com a introdução de uma segunda equação de estimação, o que possibilitou estimar e realizar inferências para os parâmetros de associação a partir de covariáveis. Quando o foco principal do estudo está na estrutura da associação e se tem estruturas hierárquicas de agrupamento, é muito comum à presença de um grande número de observações nos grupos. Os métodos numéricos para GEE2 podem ser inviáveis computacionalmente se a quantidade de observações dentro dos grupos for grande, além de não permitir utilizar a razão de chances para interpretação da estrutura de associação, sendo possível utilizar somente o coeficiente de correlação. Com adequadas modificações na segunda equação de estimação, Carey, Zeger e Diggle (1993) desenvolveram o ALR e Zink (2003) o ORTH, que além de permitir utilizar a razão de chances para interpretação da estrutura de associação, exigem um menor esforço computacional quando comparados aos métodos já propostos. Neste trabalho apresentamos e comparamos os modelos marginais (GEE2, ALR e ORTH) em simulações e aplicações reais com estruturas hierárquicas de agrupamento em respostas binárias, com o foco da pesquisa não somente nos coeficientes da média, mas também nas medidas de associação. Nossos resultados indicaram que os métodos ALR e ORTH se mostraram eficazes para aplicações com respostas binárias na presença de múltiplas estruturas hierárquicas, especialmente nos casos com um grande número de observações nos grupos. Nossos resultados também indicaram que quando o objetivo principal da pesquisa estiver na estrutura de associação, deve-se ter um cuidado especial na modelagem da estrutura da média, pois sua mal especificação pode induzir uma falta de consistência nas medidas de associação.

Palavras-Chaves: respostas binárias; estruturas hierárquicas de agrupamento; modelos marginais; equações de estimação generalizada; regressão logística alternada; resíduos ortogonalizados.

Abstract

Studies involving binary responses with hierarchical clustering are often found in different areas of knowledge. In regression models clustering structures should be properly handled, once they violate basic assumption of independence of observations. When the main focus of the study lies in the mean structure, first order Marginal Models , known as GEE1 proposed by Liang and Zeger (1986), has no distributional assumptions and offers an ease interpretation solution. However with GEE1 method, it is not possible satisfactorily accommodate hierarchical structures or multiple clustering structures, which may cause loss of efficiency in the mean structure. In order to satisfactorily accommodate hierarchical clustering, GEE1 been extended to the GEE2 (Prentice, 1988) by introducing a second estimation equation, allowing to estimate and make inferences for association parameters starting from covariates. When the main focus of the study lies in association structure and there is a hierarchical clustering structures, it is very common the presence of a large number of observations in the clusters. Numerical methods for GEE2 can be computationally infeasible if the number of observations within the clusters is large, and do not allow use of odds ratio for interpretation of association structure, being possible only the of correlation coefficient. With appropriate modification in the second equation estimation, Carey, Zeger and Diggle (1993) developed the ALR and Zink (2003) the ORTH, that requires less computational effort when compared with existing methods allows the use of odds ratio for interpretation of association structure. Here we report the comparison of marginal models (GEE2, ALR and ORTH) in simulations and real applications cases with hierarchical clustering in binary responses, with the research focus not only on the coefficients of the mean, but also in the association measures. Our results indicate that the methods ALR and ORTH proved to be effective for studies with binary responses in the presence of multiple hierarchical structures, especially in cases with a large number of observations in the clusters. Our results also indicate that when the primary purpose of the research is association structure, one should take special care in modeling the structure of the mean, because its misspecification may inconsistency of the association measures.

Keywords: correlated binary observations; hierarchical clustering structures; marginal models; generalized estimating equations; alternating logistic regressions; orthogonalized residuals.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Estruturas de agrupamento em respostas binárias são frequentemente encontradas em estudos epidemiológicos e em outras áreas do conhecimento, exigindo uma abordagem mais sofisticada para modelagem estatística. Em modelos de regressão essas estruturas de agrupamento devem ser devidamente tratadas, uma vez que violam o pressuposto básico de independência das observações. Na presença de dados agrupados, pressupõe-se que existe correlação entre as observações do mesmo grupo, enquanto que não existe nenhuma correlação entre as observações de grupos distintos. Tais estruturas podem ser induzidas pelo próprio desenho da pesquisa, como, estudos longitudinais, familiares ou com componentes espaciais. As estruturas hierárquicas ou múltiplas estruturas de agrupamento surgem com a presença de diferentes níveis dentro do mesmo grupo, em que se espera que as observações dentro dos níveis sejam correlacionadas, porém diferentemente do caso de um única estrutura de agrupamento, espera-se que também exista correlação entre as observações de níveis diferentes, uma vez que estas observações pertencem todas ao mesmo grupo.

Este trabalho foi motivado por dois estudos envolvendo estruturas hierárquicas com respostas binárias, com o foco da pesquisa não somente nos coeficientes da média, mas também nas medidas de correlação ou associação de cada nível hierárquico. O primeiro estudo apresenta dois níveis hierárquicos, sendo o segundo nível hierárquico, longitudinal. O segundo estudo apresenta uma complexa estrutura hierárquica de 4 níveis. Vamos descrever a seguir as duas pesquisas.

A primeira pesquisa trata de um estudo epidemiológico com 631 crianças de 548 domicílios que foram acompanhadas ao longo de três medidas no tempo, durante um período de um ano, com o objetivo de comparar a ocorrência de pelo menos uma infecção parasitológica em dois grupos de interesse, controlando por possíveis fatores de confundimento. Nesse problema há duas estruturas hierárquicas de agrupamento, crianças dentro de domicílios e medidas ao longo do tempo para a mesma criança, o que caracteriza esse segundo nível do agrupamento como longitudinal. Nesse estudo, as crianças encontradas com alguma infecção parasitológica eram tratadas por medicamentos específicos, o que poderia gerar interesse nas medidas de associação intradomicílio e intracriança, pois se o tratamento foi eficaz, espera-se que a criança que apresentou a infecção parasitológica em alguma etapa não apresente mais essa infecção na próxima etapa, o que pode induzir uma ausência de associação intracriança. Ao tratar uma criança que tinha alguma infecção parasitólgica em um domicílio em determinado tempo, espera-se que no próximo tempo ela não apresente mais essa infecção, porém se existirem outras crianças no mesmo domicílio, essas devem apresentar uma maior chance de infecção quando comparadas às crianças de domicílios que ainda não apresentaram nenhum caso de infecção. Dessa forma, espera-se uma associação intradomicílio positiva.

A segunda pesquisa trata de um estudo da ecologia de microorganismo, em que a presença ou ausência de um grupo de fungos foi medida em 5 diferentes locais de coleta, dois no Brasil e três na Argentina. Em cada local era realizado um transecto e vinte plantas eram selecionadas a cada 5 metros aproximadamente. Para cada planta foram selecionadas 5 folhas, sendo que a coleta dos fungos foi realizada em 6 diferentes fragmentos da folha, cada fragmento com sua propriedade biológica específica. Com 600 medidas para cada local, em todo o estudo foram coletadas 3000 medidas de ausência ou presença do grupo de fungos. Nesse estudo o objetivo principal da pesquisa é obter medidas de associação do grupo de fungos, intralocal, intraplanta, intrafolha e intrafragmento, sendo que a associação intralocal está condicionada à distância entre plantas, uma vez que é razoável pensar que à medida que se aumenta essa distância em um mesmo local, diminui-se a associação do grupo de fungos. O objetivo secundário é verificar, através da estrutura da média, se a prevalência do grupo de fungos no Brasil é diferente daquela obtida na Argentina.

Quando o foco principal da pesquisa está na estrutura da média, um caminho para contabilizar a correlação existente em um mesmo grupo dentre as medidas repetidas é introduzir um efeito aleatório, porém esse procedimento em respostas não normais ou em modelos não lineares nos parâmetros implica que os valores esperados para a média sejam condicionais ao efeito aleatório, dificultando a interpretação do modelo. Os Modelos Marginais de primeira ordem, estimados pelo método GEE1, propostos por Liang e Zeger (1986), oferecem uma solução elegante por sua facilidade na interpretação e ausência de suposições distribucionais. Entretanto a estrutura de agrupamento em GEE1 é considerada como um fator de perturbação, não

possibilitando acomodar adequadamente estruturas hierárquicas ou múltiplas estruturas de agrupamento. Quando se tem estruturas hierárquicas de agrupamento, geralmente as medidas de correlação ou associação intragrupo e intergrupo também são de interesse científico, não devendo ser tratadas simplesmente como fatores de perturbação.

Para acomodar mais de uma estrutura de agrupamento o GEE1 (Liang e Zeger, 1986) foi estendido para o GEE2 com a introdução de uma segunda equação de estimação, o que possibilitou estimar e realizar inferências para os parâmetros de associação a partir de covariáveis. Atualmente a definição de GEE2 não é única nem claramente estabelecida porque muitos procedimentos são entendidos por esse termo. Dependendo da construção da segunda equação de estimação, as estimações das médias e das associações podem ser ortogonais ou não ortogonais (Prentice,1988, Prentice e Zhao, 1991 e Liang, Zeger e Qaqish, 1992). A desvantagem da estimação não ortogonal é que os parâmetros estimados da média podem não ser consistentes se a estrutura de associação estiver mal especificada (Liang, Zeger e Qaqish, 1992). Para manter a consistência dos parâmetros da média as equações de estimação devem ser construídas cuidadosamente, para que mesmo com a mal especificação da estrutura de dependência não se corrompa a estrutura da média.

Os ajustes dos modelos baseados no GEE2 podem ser inviáveis computacionalmente se a quantidade de medidas dentro do grupo for grande, fato muito comum em problemas que envolvem estruturas hierárquicas de agrupamento. Carey, Zeger e Diggle (1993) propõe o método ALR (Alternating Logistic Regression) como solução para os problemas computacionais dos métodos já propostos. A ALR é estruturalmente diferente das outras abordagens, uma vez que para evitar o esforço computacional dos métodos para GEE2, define a segunda equação de estimação sobre resíduos condicionais (diferença entre a resposta e uma esperança condicional). Além de evitar o esforço computacional, Carey, Zeger e Diggle (1993) mostraram que as estimativas para a estrutura de associação são tão eficientes quanto as do método GEE2 proposto por Liang, Zeger e Qaqish(1992). Cabe destacar que o método GEE2 (Liang, Zeger e Qaqish, 1992) possui a estimação da média e das associações não ortogonais, o que implica na troca da consistência da estrutura da média por estimativas altamente eficientes para estrutura de associação.

A estratégia adotada por Carey, Zeger e Diggle(1993) para a ALR apresenta um problema teórico, uma vez que os resíduos condicionais induzem uma matriz de covariância para a segunda equação de estimação com uma natureza estocástica, não coerente com a teoria padrão de Equações de Estimação. Além do problema teórico, a estratégia dos resíduos condicionais também apresenta um problema prático, pois o estimador da variância robusta não é invariante a permutações da resposta (Carey, 1992). Zink(2003) propõe uma nova abordagem para a segunda equação de estimação, substituindo a estratégia dos resíduos condicionais pela dos resíduos ortogonalizados (ORTH), criando assim uma nova representação para o método ALR, solucionando seu problema prático e teórico.

O objetivo desse trabalho é comparar as metodologias GEE2(Prentice, 1988), ALR(Carey, Zeger e Diggle, 1993) e ORTH(Zink, 2003) utilizando respectivamente as funções `geese()`, `ordgee()` e `orth()`, dos pacotes `geepack` e `orth` do software R (R Development Core Team, 2012), em simulações e aplicações reais com estruturas hierárquicas de agrupamento, com o foco da pesquisa não somente nos coeficientes da média, mas também nas medidas de associação.

Esse trabalho foi organizado da seguinte forma. Na Seção 2 nós apresentamos os métodos GEE1(Liang e Zeger, 1986), GEE2(Prentice, 1988), ALR(Carey, Zeger e Diggle, 1993) e ORTH(Zink, 2003). Na Seção 3 nós realizamos simulações, a fim de uma melhor comparação entre os métodos. Na Seção 4 nós apresentamos as duas aplicações, em que acomodamos estruturas hierárquicas de agrupamento. Na Seção 5 nós comentamos sobre os resultados da simulação e aplicação dos métodos utilizados.

2 Metodologia

Inicialmente, vamos apresentar a estrutura do modelo marginal considerando uma única estrutura de agrupamento, possibilitando posteriormente a apresentação do modelo marginal para os métodos que permitem acomodar múltiplas estruturas de agrupamento.

2.1 Uma Única Estrutura de Agrupamento

Suponha dados em que são avaliados N grupos ou sujeitos independentes, cada um com n_i observações. Considere o índice i identificando o i-ésimo grupo, com $i = 1, ..., N$; considere ainda j e k identificando duas observações dentro do grupo, sendo que $1 \leq j < k \leq n_i$. Então, o índice ijk estará se referindo às observações j e k do i-ésimo grupo. Para o i-ésimo grupo o vetor resposta é dado por $Y_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})'$, sendo que cada Y_{ij} segue uma distribuição de Bernoulli com média $\mu_{ij} = pr(Y_{ij} = 1)$.

O modelo marginal proposto por Liang e Zeger (1986) pode ser apresentado através das seguintes especificações:

1. $\mathbb{E}(Y_{ij}|X_{ij}) = \mu_{ij}(\beta)$ é assumida depender de um vetor de p covariáveis X_{ij} através de uma função de ligação do tipo: $g(\mu_{ij}) = \eta_{ij} = X_{ij}\beta$;
2. $Var(Y_{ij}|X_{ij}) = \phi\upsilon(\mu_{ij})$;
3. A correlação intra-indivíduo $Corr(Y_{ij}, Y_{ik})$ é assumida ser função de um vetor adicional de parâmetros, denotado por α.

Para a resposta binária em geral assume-se que:

1. $g(\mu_{ij}) = log(\frac{\mu_{ij}}{1-\mu_{ij}})$ (Ligação logit).
2. $Var(Y_{ij}|X_{ij}) = \mu_{ij}(1 - \mu_{ij}),$ sendo $\phi = 1$ (Parâmetro de dispersão fixo).
3. $Corr(Y_{ij}, Y_{ik}) = \begin{cases} 1, & \text{se } j = k, \\ \alpha_{jk}, & \text{se } j \neq k. \end{cases}$

O estimador GEE para o modelo marginal é obtido através da minimização de:

$$\sum_{i=1}^{N} \{Y_i - \mu_i(\beta)\}' V_i^{-1} \{Y_i - \mu_i(\beta)\}, \tag{1}$$

em que $\mu_i = (\mu_{i1}, ..., \mu_{in_i})$ e $V_i = A_i^{\frac{1}{2}} R_i(\alpha) A_i^{\frac{1}{2}}$, sendo que $R_i(\alpha)$ é uma matriz $n_i x n_i$, denominada matriz de correlações de trabalho, e A_i é uma matriz diagonal com os elementos da diagonal dados por $Var(Y_{ij}) = \mu_{ij}(1 - \mu_{ij})$.

Apresentamos a seguir algumas possíveis escolhas para a matriz de correlações de trabalho $R_i(\alpha)$. Os estimadores de α, baseados nos resíduos padronizados $e_{ij} = (Y_{ij} - \hat{\mu_{ij}})/\sqrt{\hat{\mu}_{ij}(1 - \hat{\mu}_{ij})}$, também estão apresentados a seguir:

Independente:

$$Corr(Y_{ij}, Y_{ik}) = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}.$$

Simétrica Composta:

$$Corr(Y_{ij}, Y_{ik}) = \begin{bmatrix} 1 & \alpha & \alpha & \dots & \alpha \\ \alpha & 1 & \alpha \dots & \alpha & \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \alpha & \alpha & \alpha & \dots & 1 \end{bmatrix}, \text{ sendo } \hat{\alpha} = \frac{\sum_{i=1}^{N} \sum_{j \neq k} e_{ij} e_{ik}}{\sum_{i=1}^{N} n_i(n_i - 1) - p}.$$

Não-Estruturada:

$$Corr(Y_{ij}, Y_{ik}) = \begin{bmatrix} 1 & \alpha_{12} & \alpha_{13} & \dots & \alpha_{1n_i} \\ \alpha_{21} & 1 & \alpha_{23} & \dots & \alpha_{2n_i} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \alpha_{n_i1} & \alpha_{n_i2} & \alpha_{n_i3} & \dots & 1 \end{bmatrix}, \text{ sendo } \hat{\alpha}_{jk} = \frac{\sum_{i=1}^{N} e_{ij} e_{ik}}{N - p}.$$

AR-1:

$$Corr(Y_{ij}, Y_{ik}) = \begin{bmatrix} 1 & \alpha & \alpha^2 & \dots & \alpha^{n_i - 1} \\ \alpha & 1 & \alpha \dots & \alpha^{n_i - 2} & \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \alpha^{n_i - 1} & \alpha^{n_i - 2} & \alpha^{n_i - 3} & \dots & 1 \end{bmatrix}, \text{ sendo } \hat{\alpha} = \frac{\sum_{i=1}^{N} \sum_{j \leq n_i - 1} e_{ij} e_{i,j+1}}{\sum_{i=1}^{N} (n_i - 1) - p}.$$

Minimizando a função (1), temos a primeira Equação de Estimação para estimar β:

$$S(\beta) = \sum_{i=1}^{N} \frac{\partial \mu_i^{'}}{\partial \beta} V_i^{-1} (Y_i - \mu_i(\beta)) = 0 \tag{2}$$

O algoritmo para ajustar o modelo marginal de primeira ordem é dado por:

1. Considerar as estimativas iniciais do β, assumindo R_i na forma Independente;
2. Determinar o R_i a ser utilizado;
3. Estimar a matriz de trabalho R_i, baseada nos resíduos padronizados e_{ij};
4. Encontrar as estimativas das matrizes de covariâncias: $\hat{V}_i = \hat{A}_i^{\frac{1}{2}} R_i(\hat{\alpha}) \hat{A}_i^{\frac{1}{2}}$;
5. Atualizar β até a convergência:

$$\hat{\beta}_{(m+1)} = \hat{\beta}_{(m)} - \left[\sum_{i=1}^{N} \frac{\partial \hat{\mu}_i'}{\partial \hat{\beta}} \{\hat{V}_i\}^{-1} \frac{\partial \hat{\mu}_i}{\partial \hat{\beta}} \right]^{-1} \sum_{i=1}^{N} \frac{\partial \hat{\mu}_i'}{\partial \hat{\beta}} \{\hat{V}_i\}^{-1} \{Y_i - \mu_i(\hat{\beta})\}. \tag{3}$$

Com o algoritmo apresentado para ajustar o modelo marginal de primeira ordem, não é possível modelar a estrutura de correlação utilizando covariáveis, o que também não possibilita incluir mais de uma estrutura de agrupamento. Quando a média marginal μ_{ij} está corretamente especificada por $g(\mu_{ij}) = X_{ij}\beta$, o estimador de β é consistente e assintoticamente distribuído segundo uma normal com média β e matriz de variância-covariância dada por:

$$Var(\hat{\beta}) = I_0^{-1} \Lambda_{11} I_0^{-1}, \tag{4}$$

em que

$$I_0 = \sum_{i=1}^{N} \frac{\partial \mu_i^{'}}{\partial \beta} V_i^{-1} \frac{\partial \mu_i}{\partial \beta}, \tag{5}$$

$$\Lambda_{11} = \sum_{i=1}^{N} \frac{\partial \mu_i^{'}}{\partial \beta} V_i^{-1} Var(Y_i) V_i^{-1} \frac{\partial \mu_i}{\partial \beta}, \tag{6}$$

sendo $Var(Y_i) = (Y_i - \mu_i)(Y_i - \mu_i)'$.

Estimativa consistente de $Var(\hat{\beta})$ pode ser obtida substituindo-se todas as quantidades desconhecidas em (4) por estimativas consistentes, ou seja, $\hat{\beta}$ e $\hat{\alpha}$. Nota-se que quando R_i está corretamente especificado, $Var(Y_i) = V_i$ e então $\Lambda_{11} = I_0$, reduzindo (4) a I_0^{-1}, o que corresponde ao estimador obtido quando a função de verossimilhança é completamente especificada. A mal especificação de R_i acarreta em perda de eficiência, porém as estimativas pontuais para a média e para o erro padrão estarão sempre assintoticamente corretas (Molenberghs e Verbeke, 2005).

2.2 Múltiplas Estruturas de Agrupamento

Prentice (1988) estendeu a ideia proposta por Liang e Zeger (1986) e introduziu a segunda equação de estimação para os parâmetros de associação, com ambas as respostas marginais, as probabilidades μ_{ij} e os pares de correlação entre as observações do mesmo grupo. Especificadamente o estimador GEE do parâmetro α pode ser obtido utilizando uma correlação amostral:

$$Z_{ijk} = \frac{(Y_{ij} - \mu_{ij}(\beta))(Y_{ik} - \mu_{ik}(\beta))}{\sqrt{\mu_{ij}(\beta)(1 - \mu_{ij}(\beta))(\mu_{ik}(\beta)(1 - \mu_{ik}(\beta))}}, \tag{7}$$

e $\mathbb{E}(Z_{ijk}) = \rho_{ijk}(\alpha)$.

Mantendo a mesma equação de estimação em (2) para estimar β, Prentice (1988) formalizou a segunda classe de equações de estimação para α da seguinte forma:

$$S(\alpha) = \sum_{i=1}^{N} \frac{\partial \rho_i'}{\partial \alpha} W_i^{-1}(Z_i - \rho_i(\alpha)) = 0, \tag{8}$$

em que $Z_i = \{Z_{ijk}\}$, $\rho_i = \{\rho_{ijk}\}$ são vetores com dimensões m_i, sendo $m_i = n_i(n_i - 1)/2$. W_i é a matriz de trabalho de Z_i, sendo comum adotar $W_i = diag(w_{i12}, ..., w_{i1n_i}, w_{i23}, ...)$, em que:

$$w_{ijk} = 1 + (1 - 2\mu_{ij})(1 - 2\mu_{ik})[(\mu_{ij}(1 - \mu_{ij}))(\mu_{ik}(1 - \mu_{ik}))]^{\frac{-1}{2}} \rho_{ijk} - \rho_{ijk}^2.$$

Tomando X_{ijk} uma matriz m_ixq em que q é a quantidade de covariáveis utilizadas para modelar a estrutura de correlação, $\mathbb{E}(Z_{ijk}) = Corr(Y_{ij}, Y_{ik}|X_{ijk}) = \rho_{ijk}(\alpha)$ pode depender de covariáveis através de uma função de ligação do tipo: $g(\rho_{ijk}) = \eta_{ijk} = X_{ijk}\alpha$. Como a correlação está restrita à (-1,1), opta-se por utilizar como função de ligação a transformação z de Fisher para o coeficiente de correlação, dado por:

$$g(\rho_{ijk}) = log\frac{1 + \rho_{ijk}}{1 - \rho_{ijk}}, \tag{9}$$

sendo que se pode obter uma expressão para ρ_{ijk}, utilizando a função inversa de (9), dada por:

$$\rho_{ijk} = corr(Y_{ij}, Y_{ik}|X_{ijk}) = \frac{exp(X_{ijk}\alpha) - 1}{exp(X_{ijk}\alpha) + 1}. \tag{10}$$

É importante destacar que a estrutura de covariância V_i da primeira equação de estimação (2) não é mais uma matriz de trabalho desde que o segundo momento é especificado por (8). Em contraste, W_i não contém as suposições de uma matriz de trabalho porque as correlações de terceira e quarta ordem são definidas iguais a zero. Uma grande contribuição de Prentice (1988) foi possibilitar a realização de inferências formais sobre os parâmetros de correlação utilizando uma estrutura de regressão. Ele provou que a distribuição conjunta de $\sqrt{N}(\hat{\beta} - \beta)$ e $\sqrt{N}(\hat{\alpha} - \alpha)$ é assintoticamente normal com média zero e matriz de variâncias-covariâncias consistentemente estimada por:

$$N \begin{pmatrix} I_0 & 0 \\ I_1 & I_2 \end{pmatrix} \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix} \begin{pmatrix} I_0 & I_1' \\ 0 & I_2 \end{pmatrix}, \tag{11}$$

em que

$$I_1 = \left(\sum_{i=1}^{N} \frac{\partial \rho_i'}{\partial \alpha} W_i^{-1} \frac{\partial \rho_i}{\partial \alpha} \right)^{-1} \left(\sum_{i=1}^{N} \frac{\partial \rho_i'}{\partial \alpha} W_i^{-1} \frac{\partial Z_i}{\partial \beta} \right) \left(\sum_{i=1}^{N} \frac{\partial \mu_i'}{\partial \beta} V_i^{-1} \frac{\partial \mu_i}{\partial \beta} \right)^{-1}, \tag{12}$$

$$I_2 = \left(\sum_{i=1}^{N} \frac{\partial \rho_i'}{\partial \alpha} W_i^{-1} \frac{\partial \rho_i}{\partial \alpha} \right)^{-1}, \tag{13}$$

$$\Lambda_{12} = \left(\sum_{i=1}^{N} \frac{\partial \mu_i'}{\partial \beta} V_i^{-1} Cov(Y_i, Z_i) W_i^{-1} \frac{\partial \rho_i}{\partial \alpha} \right), \tag{14}$$

$$\Lambda_{21} = \Lambda_{12}', \tag{15}$$

$$\Lambda_{22} = \left(\sum_{i=1}^{N} \frac{\partial \rho_i'}{\partial \alpha} W_i^{-1} Var(Z_i) W_i^{-1} \frac{\partial \rho_i}{\partial \alpha} \right), \tag{16}$$

sendo I_0 e Λ_{11} definidos em (5) e (6) respectivamente. Tem-se ainda que $Var(Y_i)$, $Cov(Y_i, Z_i)$ e $Var(Z_i)$ são estimadas respectivamente por $(Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)'$, $(Y_i - \hat{\mu}_i)(Z_i - \hat{\rho}_i)'$, $(Z_i - \hat{\rho}_i)(Z_i - \hat{\rho}_i)'$.

O algoritmo para o método GEE2 (Prentice, 1998) poderia começar com (β_0, α_0) de um modelo supondo independência entre as observações de um mesmo grupo e atualizar as estimativas de (β_{m+1}, α_{m+1}) a partir de (β_m, α_m) utilizando o algoritimo Escore de Fisher:

$$\hat{\beta}_{(m+1)} = \hat{\beta}_{(m)} - \left[\sum_{i=1}^{N} \frac{\partial \hat{\mu}_i'}{\partial \hat{\beta}} \{\hat{V}_i\}^{-1} \frac{\partial \hat{\mu}_i}{\partial \hat{\beta}} \right]^{-1} \sum_{i=1}^{N} \frac{\partial \hat{\mu}_i'}{\partial \hat{\beta}} \{\hat{V}_i\}^{-1} \{Y_i - \mu_i(\hat{\beta})\}. \tag{17}$$

$$\hat{\alpha}_{(m+1)} = \hat{\alpha}_{(m)} - \left[\sum_{i=1}^{N} \frac{\partial \hat{\rho}_i'}{\partial \hat{\alpha}} \{\hat{W}_i\}^{-1} \frac{\partial \hat{\rho}_i}{\partial \hat{\alpha}} \right]^{-1} \sum_{i=1}^{N} \frac{\partial \hat{\rho}_i'}{\partial \hat{\alpha}} \{\hat{W}_i\}^{-1} \{Z_i - \rho_i(\hat{\alpha})\}. \tag{18}$$

Para respostas binárias o coeficiente de correlação não é muito utilizado como medida de associação, principalmente pela dificuldade na interpretação. No caso de respostas binárias, uma medida mais conveniente é a razão de chances. Lipsitz (1991) propôs uma modificação na segunda equação de estimação proposta por Prentice (1988) utilizando a razão de chances para contabilizar a associação intragrupo. Para isso definiu o logaritmo das razões de chances $log\, \tau_{ijk}$, como sendo:

$$logOR(Y_{ij}, Y_{ik}) = log\, \tau_{ijk}(\alpha) = log \left(\frac{\mu_{ijk}(1 - \mu_{ij} - \mu_{ik} + \mu_{ijk})}{(\mu_{ij} - \mu_{ijk})(\mu_{ik} - \mu_{ijk})} \right) = X_{ijk}\alpha, \quad 1 \leq j \leq k \leq n_i. \tag{19}$$

Deve-se notar que a expressão para $\tau_{ijk}(\alpha)$ é obtida através de uma tabela de contingência para as respostas das medidas repetidas, como pode-se observar na Tabela 1.

Tabela 1. Tabela de contingência para respostas com medidas repetidas

Associação entre respostas com medidas repetidas		Observação k		Total
		1	0	
Observação j	1	μ_{ijk}	$\mu_{ij} - \mu_{ijk}$	μ_{ij}
	0	$\mu_{ik} - \mu_{ijk}$	$1 - \mu_{ij} - \mu_{ik} + \mu_{ijk}$	$1 - \mu_{ij}$
Total		μ_{ik}	$1 - \mu_{ik}$	1.0

Para construção da segunda equação estimação, Lipsitz (1991) substituiu Z_i por U_i, sendo que $U_i = U_{ijk} = (Y_{i1}Y_{i2}, ..., Y_{in_i-1}Y_{in_i})$. Definiu $E(U_{ijk}) = \mu_{ijk} = Pr(Y_{ij} = 1, Y_{ik} = 1)$ e mostrou que se pode estimar μ_{ijk} a partir das razões de chances $\tau_{ijk}(\alpha)$ e das propabilidades marginais $\mu_{ij}(\beta)$ e $\mu_{ik}(\beta)$ utilizando uma solução (Mardia, 1967) de uma equação quadrática dada por:

$$\mu_{ijk} = \begin{cases} \frac{f_{ijk} - [f_{ijk}^2 - 4\tau_{ijk}(\tau_{ijk}-1)\mu_{ij}\mu_{ik}]^{\frac{1}{2}}}{2(\tau_{ijk}-1)}, & \text{se} \quad \tau_{ijk} \neq 1, \\ \tau_{ijk}\mu_{ij}\mu_{ik}, & \text{se} \quad \tau_{ijk} = 1, \end{cases} \tag{20}$$

em que $f_{ijk} = 1 - (1 - \tau_{ijk})(\mu_{ij} + \mu_{ik})$. Nota-se que $\mu_{ijk} = \mu_{ijk}(\beta, \alpha)$ é função dos β's através de μ_{ij} e μ_{ik}, e de α através de τ_{ijk}. Definindo por conveniência computacional, $G_i = Var(U_{ijk}) = diag(\mu_{ijk}(1 - \mu_{ijk}))$, Lipsitz (1991) propôs a segunda equação de estimação como:

$$S(\alpha) = \sum_{i=1}^{N} \frac{\partial \mu_{ijk}^{'}}{\partial \alpha} G_i^{-1}(U_i - \mu_{ijk}(\alpha, \beta)) = 0. \tag{21}$$

Para obter $\hat{\beta}$ e $\hat{\alpha}$, Lipsitz (1991) utiliza o mesmo algoritmo apresentado por Prentice (1988), porém substitui na equação (18) Z_i, $\hat{\rho}_i$ e $\hat{W}_i$, respectivamente por U_i, $\hat{\mu}_{ijk}$ e $\hat{G}_i$.

As equações de estimação propostas por Prentice (1988) e Lipsitz (1991) consideram β e α como sendo ortogonais, garantindo assim a consistência dos estimadores. Liang, Zeger e Qaqish (1992) propuseram as equações de estimação não considerando mais β e α ortogonais, ganhando-se em eficiência quando a estrutura de associação está bem especificada, porém a consistência de β também depende de sua correta especificação.

Os métodos numéricos para GEE2 (Prentice, 1988, Lipsitz, 1991 e Liang, Zeger e Qaqish, 1992) podem ser inviáveis computacionalmente se a quantidade de medidas dentro do grupo for grande, fato muito comum em problemas que envolvem estruturas hierárquicas de agrupamento, como em nosso segundo problema apresentado na aplicação, seção 4. Carey, Zeger e Diggle(1993) apresentaram a Regressão Logística Alternada (ALR) como solução para os problemas computacionais dos métodos já propostos. A ALR mantém a primeira ordem das equações de estimação para estimar os β, pois garante a robustez e uma razoável eficiência quando assumimos uma forma para $Var(Y_i)$ próxima da verdadeira matriz de covariância, porém é estruturalmente diferente das outras abordagens, uma vez que para evitar o esforço computacional dos métodos para GEE2, define a segunda equação de estimação sobre resíduos condicionais, estimando α usando m_i eventos condicionais de Y_{ij} dado $Y_{ik} = y_{ik}$. Considerando $\gamma_{ijk} = \log \tau_{ijk}(\alpha) = X_{ijk}\alpha$, Carey, Zeger e Diggle(1993) definiram $\xi_{ijk} = E(Y_{ij}|Y_{ik} = y_{ik})$ como:

$$\xi_{ijk} = logit^{-1}\{\gamma_{ijk}y_{ik} + log\left(\frac{\mu_{ij} - \mu_{ijk}}{1 - \mu_{ij} - \mu_{ik} + \mu_{ijk}}\right)\}. \tag{22}$$

Considerando $\gamma_{ijk} = \alpha$, é importante observar que α é o coeficiente da regressão logística de Y_{ij} sobre Y_{ik}, desde que o segundo termo da equação (22) é usado como "offset". Note que o "offset" depende dos valores atuais de β e α, de modo que é necessário o processo de iteração.

Denotando o vetor dos resíduos condicionais de dimensão m_i por $C_i = Y_{ij} - \xi_{ijk}$ e H_i a matriz diagonal com os elementos $\xi_{ijk}(1 - \xi_{ijk})$, tem-se que o estimador ALR para $\theta = (\beta, \alpha)$ é a solução simultânea das seguintes equações de estimação:

$$S(\beta) = \sum_{i=1}^{N} \frac{\partial \mu_i^{'}}{\partial \beta} V_i^{-1}(Y_i - \mu_i(\beta)) = 0 \tag{23}$$

$$S_{\alpha,ALR} = \sum_{i=1}^{N} \frac{\partial \xi_i^{'}}{\partial \alpha} H_i^{-1} C_i = 0 \tag{24}$$

É importante notar que embora os elementos em C_i não sejam independentes, Carey, Zeger e Diggle(1993) argumentam que as correlações entre os resíduos condicionais deveriam ser substancialmente menores que as correlações entre os resíduos não condicionais, estimados pelos métodos GEE2. Consequentemente a fixação de $H_i = diag(\xi_{ijk}(1-\xi_{ijk}))$ em (24) deve proporcionar uma melhor aproximação da matriz de pesos quando comparada com W em (8), possibilitando ganhos em eficiência. Também deve-se notar que C_i em (24) é uma função linear de Y_i, enquanto $Z_{ijk} - \rho_{ijk}$ em (8) e $U_{ijk} - \mu_{ijk}$ em (21) são funções quadráticas no sentido de envolver combinações lineares do produto $Y_{ij}Y_{ik}$, $j < k$.

Lipsitz e Fitzmaurice (1996), após definir $S_{\alpha,ALR}$ utilizando a correlação como medida de associação, mostraram que $S_{\alpha,ALR}$ é mais eficiente que os métodos de GEE2(Prentice, 1988), especialmente quando os pares de correlação são muitos ou quando o tamanho do grupo é uma variável de interesse sobre as medidas de associação.

A matriz de variâncias-covariâncias do estimador $\hat{\theta}$ via ALR é consistentemente estimada por:

$$\{F^*(\hat{\theta})\}^{-1}\{\sum F_i(\hat{\theta})F_i(\hat{\theta})^{'}\}\{F^*(\hat{\theta})^{'}\}^{-1}, \tag{25}$$

em que

$$F(\hat{\theta}) = \sum_i F_i(\hat{\theta}) = \begin{pmatrix} \sum_i \frac{\partial \mu_i^{'}}{\partial \beta} V_i^{-1}(Y_i - \mu_i) \\ \sum_i \frac{\partial \xi_i^{'}}{\partial \alpha} H_i^{-1} C_i \end{pmatrix}, \tag{26}$$

$$F^*(\hat{\theta}) = \begin{pmatrix} \sum_i \frac{\partial \mu_i^{'}}{\partial \beta} V_i^{-1} \frac{\partial \mu_i}{\partial \beta} & 0 \\ \sum_i \frac{\partial \xi_i^{'}}{\partial \alpha} H_i^{-1} \frac{\partial \xi_i}{\partial \beta} & \sum_i \frac{\partial \xi_i^{'}}{\partial \alpha} H_i^{-1} \frac{\partial \xi_i}{\partial \alpha} \end{pmatrix}. \tag{27}$$

Contudo a matriz H_i possui uma natureza estocástica e não consiste nos elementos da diagonal de alguma genuína matriz de covariancia, ou seja, $Var(C_i) \neq \xi_{ijk}(1-\xi_{ijk})$. A natureza estocástica de H_i e $\partial \xi_i / \alpha$ fazem com que a investigação teórica de (24) através da teoria padrão de Equações de Estimação não seja possível. Outro ponto é que $S_{\alpha,ALR}$ é invariante a permutações do vetor Y_i (Carey, 1992; Kuk, 2004) enquanto que o estimador da variância robusta não é.

Zink(2003) apresentou ORTH (Resíduos Ortogonalizados) como solução do problema prático e teórico do ALR. A abordagem dos resíduos ortogonalizados (ORTH), mantém novamente a mesma equação de estimação para a estrutura da média $S_{\beta,ORTH} = S_{\beta,GEE1}$, sendo que para a construção da segunda equação de estimação, utiliza-se como princípio duas idéias: Os pares dos resíduos são desenvolvidos através de um argumento de projeção, e a combinação ponderada desses resíduos é feita usando uma aproximação da matriz de covariância que é computacionalmente característica de grandes grupos (Qaqish, Zink e Preisser, 2012).

Sejam $R_{iYU} = Cov(Y_i, U_i)$ e $R_{iUU} = Var(U_i)$. A matriz R_{iYU} é computacionalmente característica de grandes grupos, pois tem elementos na forma $Cov(Y_{ij^*}, U_{ijk})$, uma vez que é natural esperar que $j^* = j$ ou $j^* = k$ para altos valores de j e k. Para eliminar essas correlações, a abordagem dos resíduos ortogonalizados utiliza regressões lineares de U_{ijk} sobre Y_{ij} e Y_{ik} especificando:

$$Q_{ijk} = U_{ijk} - [\mu_{ijk} + b_{ijk:j}(Y_{ij} - \mu_{ij}) + b_{ijk:k}(Y_{ik} - \mu_{ik})], \tag{28}$$

em que:
$b_{ijk:j} = \mu_{ijk}(1 - \mu_{ik})(\mu_{ik} - \mu_{ijk})/d_{ijk}$,
$b_{ijk:k} = \mu_{ijk}(1 - \mu_{ij})(\mu_{ij} - \mu_{ijk})/d_{ijk}$,
$d_{ijk} = \sigma_{ijj}\sigma_{ikk} - \sigma_{ijk}^2$,
$\sigma_{ijk} = \mu_{ijk} - \mu_{ij}\mu_{ik}$,
$\sigma_{ijj} = \mu_{ij}(1 - \mu_{ij})$,
$\sigma_{ikk} = \mu_{ik}(1 - \mu_{ik})$.

Como resultado, tem-se que $Cov(Y_{ij}, Q_{ijk}) = Cov(Y_{ik}, Q_{ijk}) = 0$. Esta definição de Q_{ijk} introduz $n_i - 1$ zeros dentro de cada linha de $R_{iYQ} = Cov(Y_i, Q_i)$, onde Q_i é um vetor de tamanho m_i com os elementos Q_{ijk}. Além do mais, a magnitude das outras entrada de R_{iYQ} tende a diminuir quando comparado com R_{iYU}, assim como os elementos fora da diagnonal de $R_{iQQ} = Var(Q_i)$, quando comparados com R_{iUU}. Logo, o ORTH tende a ter mais eficiência que GEE2(Liang, Zeger e Qaqish, 1992).

Após a definição dos resíduos ortogonalizados Q_{ijk}, a segunda equação de estimação é dada por:

$$S_{\alpha,ORTH} = \sum_{i=1}^{N} \frac{\partial \mu_{ijk}^{'}}{\partial \alpha} P_i^{-1} Q_i, \tag{29}$$

em que P é matriz diagonal com elementos $\nu_{ijk} = Var(Q_{ijk}) =$

$$\frac{\mu_{ijk}(\mu_{ij} - \mu_{ijk})(\mu_{ik} - \mu_{ijk})(1 - \mu_{ij} - \mu_{ik} + \mu_{ijk})}{\mu_{ij}\mu_{ik}(1 - \mu_{ij} - \mu_{ik} + \mu_{ijk}) - \mu_{ijk}^2}.$$

Para possibilitar ganhos em eficiência, R_{iQQ} pode ser aproximado por uma matriz de trabalho simétrica composta:

$$R_{iQQ}^{*} = \lambda \mathbf{1}\mathbf{1}^{'} + (1 + \lambda)I,$$

em que I é uma matriz identidade m_ixm_i, 1 é um vetor m_ix1 com valores 1 e λ é o parâmetro de correlação a ser estimado.

Denotando ν_{ijk} a variância de Q_{ijk}, pode-se aproximar $Var(Q_i)$ por:

$$P_i = diag(\sqrt{\nu_i}) R_{iQQ}^{*}(\lambda) diag(\sqrt{\nu_i}),$$

possibilitando obter $S_{\alpha,ORTH}$ como em (29).

O parâmetro λ pode ser estimado pelo método dos momentos:

$$\hat{\lambda(\theta)} = \frac{1}{M} \left[\sum_{j<k} \frac{Q_{ijk}}{\sqrt{\nu_i}} - \sum_{j<k} \frac{Q_{ijk}^2}{\nu_i} \right],$$

com $M = \sum_{i=1}^{N} m_i(m_i - 1)$.

Zink(2003) mostrou que $S_{\alpha,ORTH} = S_{\alpha,ALR}$ quando λ não é estimado, mas sim fixado igual a zero. Isto significa que usando (29) é possivel escrever a ALR de forma consistente com a teoria de Equação de Estimação. Se com a incorporação de λ, R_{iQQ}^{*} se aproximar da verdadeira matriz de correlação de Q_{ijk}, deve-se ter possíveis ganhos de eficiência.

Com resultados assintóticos similares aos de Prentice(1988) e Liang e Zeger(1986), Zink(2003) mostrou que a distribuição assintótica de $\sqrt{N}(\hat{\theta} - \theta)$ é normal multivariada com vetor de média zero e matriz de covariância consistentemente estimada por $NL^{-1}\Upsilon L^{-1\prime}$ em que L e Υ consistem nos seguintes blocos:

$$L_{11} = I_0, \tag{30}$$

$$L_{12} = 0, \tag{31}$$

$$L_{21} = -\left(\sum_{i=1}^{N} \frac{\partial \mu_{ijk}^{'}}{\partial \alpha} P_i^{-1} \frac{\partial \mu_{ijk}}{\partial \beta}\right), \tag{32}$$

$$L_{22} = \left(\sum_{i=1}^{N} \frac{\partial \mu_{ijk}^{'}}{\partial \alpha} P_i^{-1} \frac{\partial \mu_{ijk}}{\partial \alpha}\right), \tag{33}$$

$$\Upsilon_{11} = \Lambda_{11} \tag{34}$$

$$\Upsilon_{12} = \left(\frac{\partial \mu_i^{'}}{\partial \beta} V_i^{-1} Cov(Y_i, Q_i) P_i^{-1} \frac{\partial \mu_{ijk}}{\partial \alpha}\right), \tag{35}$$

$$\Upsilon_{21} = \Upsilon_{12}^{'}, \tag{36}$$

$$\Upsilon_{22} = \left(\frac{\partial \mu_{ijk}^{'}}{\partial \alpha} P_i^{-1} Var(Q_i) P_i^{-1} \frac{\partial \mu_{ijk}}{\partial \alpha}\right), \tag{37}$$

e $Var(Y_i)$, $Cov(Y_i, Q_i)$ e $Var(Q_i)$ são estimadas respectivamente pelas quantidades $(Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)^{'}$, $(Y_i - \hat{\mu}_i)Q_i^{'}$, $Q_i Q_i^{'}$.

Devido à definição dos resíduos ortogonalizados, tem-se como propriedade que $Q_{ijk} = Q_{ikj}$. Com essa propriedade é possível mostrar que $S_{\alpha,ORTH}$ e o associado estimador da variância robusta $NL^{-1}\Upsilon L^{-1\prime}$ são invariantes à permutação dos dados Y_i.

Os diferentes métodos para ajustar os modelos marginais apresentados na seção 2, estão disponíveis no software R (R Development Core Team, 2012). A função `geese()` do pacote `geepack` ajusta a segunda classe de equações de estimação, considerando os coeficientes de regressão e as medidas de associações ortogonais (Prentice, 1988). A função `ordgee()` também do pacote `geepack` possibilita ajustar o modelo marginal utilizando o método das Regressões Logísticas Alternadas, ALR (Carey, 1993), enquanto a função `orth` do pacote `orth` utiliza o método dos Resíduos Ortogonalizados, ORTH (Zink, 2003).

3 Simulação de Monte Carlo

Nessa seção iremos apresentar alguns resultados de estudos de simulação para comparar o desempenho dos métodos GEE2, ALR e ORTH. As simulações buscam explorar as propriedades dos estimadores da média e associação, assim como de seus respectivos estimadores da variância robusta.

Qaqish (2003) introduziu uma família de distribuições binárias multivariadas que possibilita, de forma simples, simular variáveis binárias correlacionadas dado um vetor de médias e uma matriz de correlação. As respostas binárias correlacionadas desse estudo de simulação foram geradas utilizando a metodologia proposta por Qaqish (2003) implementada no pacote `binarySimCLF` do software R.

Para realizar as comparações entre os modelos de interesse foram criados 6 diferentes cenários, sendo que para cada um foram utilizadas 1000 simulações.

1. Cenário-I:

   Foram considerados dois tamanhos de amostra, o primeiro subcenário com 800 e o segundo com 200 observações, isso a partir de respectivamente 200 e 50 conglomerados, ambos com 4 medidas repetidas. Em cada subcenário, metade das observações pertencem ao grupo A e a outra metade ao grupo B. Para gerar as respostas binárias correlacionadas e ajustar os modelos propostos foram consideradas as seguintes estruturas:

   $$logitPr(Y = 1) = -1.70 + 0.70I(Grupo = A),$$

   $$g(Corr(Y_j, Y_k)) = \begin{cases} 0.62\text{I(Mesmo Conglomerado=1)} \\ \text{se j e k forem observações diferentes no mesmo conglomerado,} \end{cases}$$

   $$LogOR(Y_j, Y_k)) = \begin{cases} 1.54\text{I(Mesmo Conglomerado=1)} \\ \text{se j e k forem observações diferentes no mesmo conglomerado,} \end{cases}$$

   sendo que a função de ligação g é a inversa da transformação z de Fisher, utilizada para o método GEE2.

2. Cenário-II:

   Foram considerados dois tamanhos de amostra, o primeiro subcenário com 800 e o segundo com 200 observações, isso a partir de respectivamente 200 e 50 conglomerados, ambos com 4 medidas repetidas, sendo que 2 das 4 medidas repetidas pertencem ao segundo nível hierárquico. Em cada subcenário, metade das observações pertencem ao grupo A e a outra metade ao grupo B. Para gerar as respostas binárias correlacionadas e ajustar os modelos propostos foram consideradas as seguintes estruturas:

   $$logitPr(Y = 1) = -1.70 + 0.70I(Grupo = A),$$

   $$g(Corr(Y_j, Y_k)) = \begin{cases} 0.62\text{I(Mesmo Conglomerado=1)} \\ \text{se j e k forem observações diferentes no mesmo conglomerado,} \\ 0.51\text{I(Mesmo Subconglomerado=1)} \\ \text{se j e k forem observações diferentes no mesmo subconglomerado,} \end{cases}$$

   $$LogOR(Y_j, Y_k)) = \begin{cases} 1.54\text{I(Mesmo Conglomerado=1)} \\ \text{se j e k forem observações diferentes no mesmo conglomerado,} \\ 1.08\text{I(Mesmo Subconglomerado=1)} \\ \text{se j e k forem observações diferentes no mesmo subconglomerado.} \end{cases}$$

3. Cenário-III:

   Foram considerados dois tamanhos de amostra, o primeiro subcenário com 1600 e o segundo com 400 observações, isso a partir de respectivamente 200 e 50 conglomerados, ambos com 8 medidas repetidas, sendo que 4 das 8 medidas repetidas pertencem ao segundo nível hierárquico e 2 dessas 4 medidas repetidas pertencem ao terceiro nível hierárquico. Em cada subcenário, metade das observações pertencem ao grupo A e a outra metade ao grupo B. Para gerar as respostas binárias correlacionadas e ajustar os modelos propostos foram consideradas as seguintes estruturas:

$$logitPr(Y = 1) = -1.70 + 0.70I(Grupo = A),$$

$$g(Corr(Y_j, Y_k)) = \begin{cases} 0.62\text{I(Mesmo Conglomerado=1)} \\ \text{se j e k forem observações diferentes no mesmo conglomerado,} \\ 0.51\text{I(Mesmo Subconglomerado=1)} \\ \text{se j e k forem observações diferentes no mesmo subconglomerado,} \\ 0.43\text{I(Mesmo Sub-subconglomerado=1)} \\ \text{se j e k forem observações diferentes no mesmo sub-subconglomerado,} \end{cases}$$

$$LogOR(Y_j, Y_k)) = \begin{cases} 1.54\text{I(Mesmo Conglomerado=1)} \\ \text{se j e k forem observações diferentes no mesmo conglomerado,} \\ 1.08\text{I(Mesmo Subconglomerado=1)} \\ \text{se j e k forem observações diferentes no mesmo subconglomerado,} \\ 0.87\text{I(Mesmo Sub-subconglomerado=1)} \\ \text{se j e k forem observações diferentes no mesmo sub-subconglomerado.} \end{cases}$$

   Em estruturas hierárquicas de agrupamento como a apresentada acima, a definição da matriz de trabalho pode não ser uma tarefa trivial. Dessa forma, apenas para ilustrar, transformando o vetor $\alpha = (0.62, 0.51, 0.43)$ em correlações, teríamos a seguinte matriz de trabalho:

$$Corr(Y_j, Y_k) = \begin{bmatrix} 1 & 0.65 & 0.51 & 0.51 & 0.30 & 0.30 & 0.30 & 0.30 \\ 0.65 & 1 & 0.51 & 0.51 & 0.30 & 0.30 & 0.30 & 0.30 \\ 0.51 & 0.51 & 1 & 0.65 & 0.30 & 0.30 & 0.30 & 0.30 \\ 0.51 & 0.51 & 0.65 & 1 & 0.30 & 0.30 & 0.30 & 0.30 \\ 0.30 & 0.30 & 0.30 & 0.30 & 1 & 0.65 & 0.51 & 0.51 \\ 0.30 & 0.30 & 0.30 & 0.30 & 0.65 & 1 & 0.51 & 0.51 \\ 0.30 & 0.30 & 0.30 & 0.30 & 0.51 & 0.51 & 1 & 0.65 \\ 0.30 & 0.30 & 0.30 & 0.30 & 0.51 & 0.51 & 0.65 & 1 \end{bmatrix}.$$

   Note que $exp(0.62 + 0.51 + 0.43) - 1)/(exp(0.62 + 0.51 + 0.43) + 1) = 0.65$, o que corresponderia se j e k forem observações diferentes no mesmo sub-subconglomerado, já se j e k forem observações diferentes no mesmo subconglomerado, $exp(0.62 + 0.51) - 1)/(exp(0.62 + 0.51) + 1) = 0.51$, e por fim, $exp(0.62) - 1)/(exp(0.62) + 1) = 0.30$, se j e k forem observações diferentes no mesmo conglomerado.

4. Cenário-IV:

   Foram considerados dois tamanhos de amostra, o primeiro subcenário com 200 e o segundo com 100 observações, isso a partir de respectivamente 50 e 25 conglomerados, ambos com 4 medidas repetidas.

Em cada subcenário metade das observações pertencem ao grupo A e a outra metade ao grupo B. Para gerar as respostas binárias correlacionadas foi considerado para a estrutura da média, $logitPr(Y = 1) = -1.70+0.70I(Grupo = A)$, e para estrutura de associação um AR-1 com $Corr(Y_j, Y_k) = 0.5$. Para os ajustes dos modelos, a estrutura da associação foi mal especificada com uma simetria composta.

5. Cenário-V

   Foram considerados dois tamanhos de amostra, o primeiro subcenário com 200 e o segundo com 100 observações, isso a partir de respectivamente 50 e 25 conglomerados, ambos com 4 medidas repetidas. Em cada subcenário, as observações foram classificadas por uma variável discreta, denominada Var, sendo que $Var = -3, -2, -1, 0, 1, 2, 3$. Para gerar as respostas binárias correlacionadas foi considerado para a estrutura da média, $logitPr(Y = 1) = -0.80+0.20Var$, e para estrutura de associação um AR-1 com $Corr(Y_j, Y_k) = 0.5$. Para os ajustes dos modelos, a estrutura da associação foi mal especificada com uma simetria composta.

6. Cenário-VI:

   Foram considerados dois tamanhos de amostra, o primeiro subcenário com 400 e o segundo com 200 observações, isso a partir de respectivamente 50 e 25 conglomerados, ambos com 8 medidas repetidas, sendo que 4 das 8 medidas repetidas pertencem ao segundo nível hierárquico e 2 dessas 4 medidas repetidas pertencem ao terceiro nível hierárquico. Em cada subcenário, as observações foram classificadas por uma variável discreta, denominada Var, sendo que $Var = -3, -2, -1, 0, 1, 2, 3$. Para gerar as respostas binárias correlacionadas foram considerados, para a estrutura da média, $logitPr(Y = 1) = -0.90 + 0.10Var^2$ e para estrutura de associação:

$$g(Corr(Y_j, Y_k)) = \begin{cases} 0.62\text{I(Mesmo Conglomerado=1)} \\ \text{se j e k forem observações diferentes no mesmo conglomerado,} \\ 0.51\text{I(Mesmo Subconglomerado=1)} \\ \text{se j e k forem observações diferentes no mesmo subconglomerado,} \\ 0.43\text{I(Mesmo Sub-subconglomerado=1)} \\ \text{se j e k forem observações diferentes no mesmo sub-subconglomerado,} \end{cases}$$

$$LogOR(Y_j, Y_k)) = \begin{cases} 1.24\text{I(Mesmo Conglomerado=1)} \\ \text{se j e k forem observações diferentes no mesmo conglomerado,} \\ 1.01\text{I(Mesmo Subconglomerado=1)} \\ \text{se j e k forem observações diferentes no mesmo subconglomerado,} \\ 0.85\text{I(Mesmo Sub-subconglomerado=1)} \\ \text{se j e k forem observações diferentes no mesmo sub-subconglomerado.} \end{cases}$$

   Para o ajuste dos modelos foi considerada a verdadeira estrutura de associação, porém a estrutura da média foi mal especificada por uma tendência linear, uma vez que a verdadeira estrutura da média era quadrática, $logitPr(Y = 1) = -0.90 + 0.10Var^2$.

Na tabela 2 os resultados das simulações foram apresentados através do vício de $\hat{\theta} = (\hat{\beta}, \hat{\alpha})$ e da diferença entre a média dos erros padrões de $\hat{\theta}$ com o verdadeiro erro padrão, dado pelo desvio padrão de $\hat{\theta}$. Dessa forma, nota-se que independentemente do cenário criado não houve diferenças consideráveis entre os métodos GEE2, ALR e ORTH. As estimativas para média, associação e variância robusta foram altamente similares. Porém, dos resultados da simulação, pode-se destacar a já conhecida robustez dos modelos marginais, que

mesmo com a mal especificação da estrutura de associação, não se corrompe a consistência da estrutura da média, mesmo para os casos com menores tamanhos de amostra. Nos cenários IV e V em que foi mal especificada a estrutura de associação, os valores estimados para média e erro padrão foram bastantes próximos dos reais valores. No cenário VI em que foi mal especificada a estrutura da média, especialmente em α_0, houve uma maior diferença entre o valor estimado e o valor real, sugerindo uma falta de consistência para os tamanhos de amostra considerados.

Tabela 2. Resultados das simulações para comparação entre os métodos GEE2, ALR e ORTH.

Cenários	Nº de grupos	θ	$E[\hat{\theta}] - \theta$			$E[\widehat{E.P}(\hat{\theta})] - D.P(\hat{\theta})$		
			GEE2	ALR	ORTH	GEE2	ALR	ORTH
I	200	β0	-0,011	-0,011	-0,011	-0,003	-0,003	-0,003
		β1	0,007	0,007	0,007	-0,002	-0,002	-0,002
		α	-0,011	-0,028	-0,022	-0,005	-0,005	-0,007
	50	β0	-0,064	-0,064	-0,064	-0,036	-0,036	-0,036
		β1	0,051	0,051	0,051	-0,052	-0,052	-0,052
		α	-0,036	-0,064	-0,060	-0,029	-0,029	-0,029
II	200	β0	-0,019	-0,019	-0,019	0,011	0,011	0,011
		β1	0,014	0,014	0,014	0,008	0,008	0,008
		α0	-0,011	-0,026	-0,019	0,008	0,004	0,015
		α1	0,002	0,009	0,016	0,002	-0,004	0,004
	50	β0	-0,053	-0,053	-0,053	-0,021	-0,021	-0,021
		β1	0,020	0,020	0,020	-0,016	-0,016	-0,016
		α0	-0,034	-0,076	-0,072	-0,037	-0,004	-0,046
		α1	0,001	0,032	0,039	-0,049	-0,047	-0,083
III	200	β0	-0,003	-0,003	-0,003	-0,002	-0,002	-0,002
		β1	-0,010	-0,010	-0,010	-0,006	-0,006	-0,006
		α0	-0,009	-0,022	-0,014	-0,003	0,018	-0,004
		α1	0,000	0,004	0,010	-0,004	0,004	-0,004
		α2	-0,003	0,004	0,009	-0,007	-0,012	-0,016
	50	β0	-0,058	-0,058	-0,058	-0,015	-0,015	-0,015
		β1	0,023	0,023	0,023	-0,011	-0,011	-0,011
		α0	-0,054	-0,096	-0,092	-0,030	0,046	-0,023
		α1	0,016	0,048	0,056	-0,017	0,016	-0,013
		α2	0,002	0,027	0,029	-0,009	0,001	-0,009
IV	50	β0	-0,047	-0,047	-0,047	-0,020	-0,020	-0,020
		β1	0,006	0,006	0,006	-0,023	-0,023	-0,023
	25	β0	-0,072	-0,073	-0,074	-0,093	-0,091	-0,092
		β1	-0,003	0,000	0,002	-0,105	-0,102	-0,103
V	50	β0	-0,021	-0,021	-0,021	-0,004	-0,004	-0,004
		β1	0,010	0,009	0,009	0,000	0,001	0,001
	25	β0	-0,028	-0,028	-0,028	-0,014	-0,016	-0,016
		β1	0,010	0,007	0,008	-0,015	-0,012	-0,013
VI	50	α0	0,170	0,350	0,355	-0,002	-0,003	-0,004
		α1	-0,021	-0,040	-0,032	-0,004	-0,006	-0,008
		α2	-0,015	-0,031	-0,019	-0,006	-0,009	-0,011
	25	α0	0,167	0,345	0,352	-0,002	-0,001	-0,003
		α1	-0,023	-0,043	-0,035	-0,002	-0,002	-0,004
		α2	-0,009	-0,021	-0,008	-0,002	-0,001	-0,004

Na Tabela 3 pode-se observar o tempo computacional (em segundos) para os ajustes dos modelos apresentados em diferentes cenários de tamanho de amostra. Os modelos foram ajustados considerando na estrutura da média a comparação de dois grupos e para estrutura de associação $Cor(Y_j, Y_k) = \alpha$ para o método GEE2 e $LogOR(Y_j, Y_k) = \alpha$ para os métodos ALR e ORTH. Foi utilizado um computador com processador due core i5.

Observando os resultados na tabela 3, destaca-se que para o método GEE2, o tempo computacional aumenta exponencialmente a partir de situações com mais de 64 medidas repetidas no grupo, sendo que o ajuste se torna inviável computacionalmente no software R usando o pacote `geepack` para situações com mais de 256 medidas repetidas, sendo que para essas situações, o método ORTH apresentou o menor tempo computacional.

Tabela 3. Comparação do tempo computacional (em segundos) entre os métodos GEE2, ALR e ORTH.

Nº de grupos	Tamanho do grupo	Métodos		
		GEE2	ALR	ORTH
20	4	0,03	0,02	0,2
100		0,03	0,05	0,73
20	8	0,02	0,03	0,44
100		0,05	0,16	2,81
20	16	0,03	0,15	1,67
100		0,14	0,67	7,69
20	32	0,23	0,85	6,49
100		1,28	3,9	30,19
20	64	6,41	7,37	33,62
100		31,4	33,27	124,32
20	128	276,91	93,41	134,81
100		1118,57	405,66	519,53
20	256	*	1402,2	420,0
100		*	6373,8	2296,1

* Inviável computacionalmente no software R.

4 Aplicação

Nesta seção vamos ilustrar a metodologia apresentada na seção 2, utilizando dois estudos reais com estruturas hierárquicas de agrupamento em respostas binárias.

4.1 Infecção Parasitológica

Este primeiro exemplo trata de um estudo epidemiológico realizado nos municípios de Berilo e Chapada do Norte, ambos caracterizados pela sua semi-aridez e localizados no vale médio do Jequitinhonha, região nordeste de Minas Gerais, Brasil. Neste estudo, duas coortes de crianças foram investigadas para avaliar o efeito da cisterna na ocorrência de pelo menos uma infecção parasitológica. Estamos chamando de pelo menos uma infecção parasitológica, infecções causadas por algum dos 11 parasitas ou comensais detectados ao longo do estudo. As coortes, com duração de 12 meses, foram definidas nos seguintes grupos:

- Grupo1: Compostos por crianças de zero a sessenta meses de idade, residindo em área rural da região selecionada, e tendo acesso em suas próprias casas ou na de outras pessoas a uma cisterna para armazenamento de águas pluviais.
- Grupo2: Compostos por crianças de zero a sessenta meses de idade, residindo em área rural da região selecionada, mas sem acesso a uma cisterna.

As famílias selecionadas para participar do estudo responderam um questionário no primeiro contato, com o objetivo de obter informações mais precisas sobre suas características econômicas, pessoais, de higiene doméstica e das condições de saúde da criança. Quando encontrada uma infecção parasitológica na criança, essa era devidamente tratada por um profissional da saúde. As crianças foram avaliadas três vezes ao longo do tempo, sendo que no primeiro contato(Etapa=1) foram pesquisadas 572 crianças, no segundo contato(Etapa=2) 463 e no terceiro contato(Etapa=3) 461. O número médio de crianças por domicílio foi de 1.29 crianças, sendo que o valor máximo foi de 4 crianças. A idade média das crianças no início do estudo era de 28.7 meses.

Na Figura 1, pode-se visualizar a frequência observada de crianças com infecção parasitológica entre os dois grupos ao longo do tempo. Na primeira etapa, do grupo com cisterna(grupo1), 23.28% das crianças apresentavam pelo menos uma infecção parasitológica, enquanto que as que não tinham cisterna esse percentual foi de 26.76%. Na segunda etapa esses percentuais aumentam para 30.90% e 37.34%, respectivamente, nos grupos 1 e 2. Já na terceira etapa os percentuais são 25.73% e 24.03%, respectivamente, para os grupos 1 e 2.

Com o objetivo de comparar a ocorrência da infecção parasitológica entre os dois grupos de interesse, controlando pelos possíveis fatores de confusão, foram utilizadas as mesmas variáveis selecionadas por Fonseca(2012). Neste contexto, foi utilizado a seguinte estrutura para média:

$$logitPr(Y = 1) = \beta_0 + \beta_1 I(\text{Grupo} = 2) + \beta_2 I(\text{Etapa} = 2) + \beta_3 I(\text{Etapa} = 3) + \beta_4 \text{Idade}$$

$$+\beta_5 I(\text{Questão85} = \text{Pouco ou Não}) + \beta_6 I(\text{Questão82} = \text{Uma vez ao dia}),$$

sendo que a idade foi medida na i-ésima criança em meses (No início do estudo a idade mínima foi 0.06 e a máxima de 56.3), Questão85 se refere se a pessoa que cozinha para i-ésima criança lava as mãos (1=Pouco ou Não, 0=Sempre) e a Questão82 se refere à frequência de banho da i-ésima criança (1=Uma vez ao dia, 0= Mais de uma vez ao dia).

Para o método GEE2, foi proposta a seguinte estrutura de associação:

$$g(Corr(Y_j, Y_k)) = \begin{cases} \alpha_1 \text{I(Mesmo domicílio=1)} + \alpha_5 \text{(NCD - 2)}, \\ \text{se j e k forem crianças diferentes no mesmo domicílio,} \\ \alpha_1 \text{I(Mesmo domicílio=1)} + \alpha_2 \text{I(Etapas=1-2)} + \alpha_3 \text{I(Etapas=1-3)} \\ +\alpha_4 \text{I(Etapas=2-3)} + \alpha_5 \text{(NCD - 2)}, \\ \text{se j e k forem tempos diferentes na mesma criança,} \end{cases}$$

sendo que a função de ligação g é a inversa da transformação z de Fisher dada em (10) e NCD é o número de crianças no mesmo domicílio. Para obter interpretação para α_1 se j e k forem crianças diferentes no mesmo domicílio, centramos NCD em 2. Para os métodos ALR e ORTH, foi proposta a seguinte estrutura:

$$LogOR(Y_j, Y_k) = \begin{cases} \alpha_1 \text{I(Mesmo domicílio=1)} + \alpha_5 \text{(NCD - 2)}, \\ \text{se j e k forem crianças diferentes no mesmo domicílio,} \\ \alpha_1 \text{I(Mesmo domicílio=1)} + \alpha_2 \text{I(Etapas=1-2)} + \alpha_3 \text{I(Etapas=1-3)} \\ +\alpha_4 \text{I(Etapas=2-3)} + \alpha_5 \text{(NCD - 2)}, \\ \text{se j e k forem tempos diferentes na mesma criança.} \end{cases}$$

É importante observar que a estrutura de associação do GEE2 é idêntica à dos métodos ALR e ORTH, modificando-se somente a função de ligação, implicando em medidas de associação diferentes.

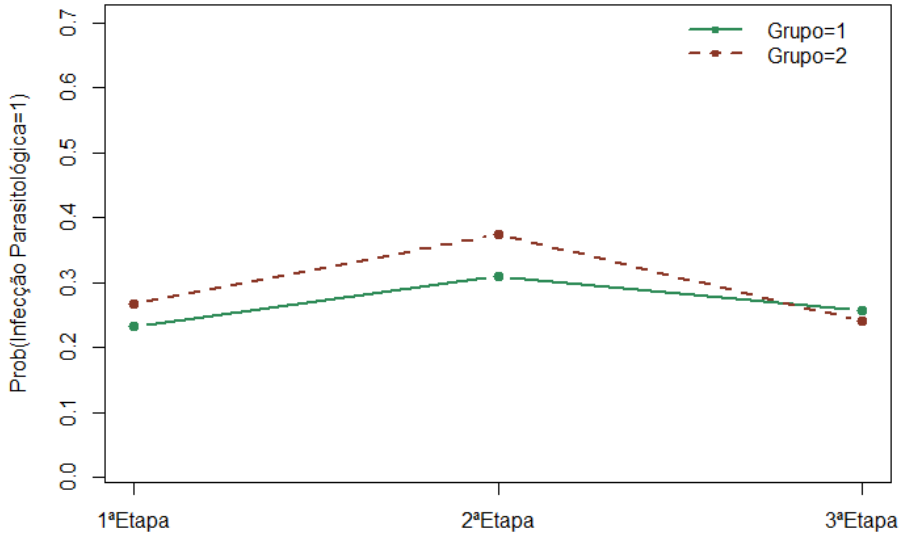

Figura 1. Perfil médio de infecção parasitológica entre as crianças dos dois grupos ao longo do tempo.

Para facilitar o entendimento de como os modelos apresentados acomodam os dados com níveis hierárquicos de agrupamento, vamos ilustrar a construção das matrizes de delineamento. A construção da matriz de delineamento para o modelo apresentado é ilustrada de acordo com os dados da tabela 4, que são do agrupamento de duas crianças residentes no mesmo domicílio medidas nas três etapas ao longo do tempo.

Tabela 4. Informações de duas crianças residentes no mesmo domicílio medidas nas três etapas ao longo do tempo.

Domicílio	Criança	Grupo	Etapa	Idade	Questão85	Questão82	Infecção(y_{ij})
1	1	2	1	41	Sempre	Uma vez ao dia	0
1	1	2	2	41	Sempre	Uma vez ao dia	1
1	1	2	3	41	Sempre	Uma vez ao dia	0
1	2	2	1	15	Sempre	Uma vez ao dia	1
1	2	2	2	15	Sempre	Uma vez ao dia	0
1	2	2	3	15	Sempre	Uma vez ao dia	0

Considerando o agrupamento das duas crianças no mesmo domicílio, a matriz de delineamento para os modelos apresentados seriam:

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 41 & 0 & 1 \\ 1 & 1 & 1 & 0 & 41 & 0 & 1 \\ 1 & 1 & 0 & 1 & 41 & 0 & 1 \\ 1 & 1 & 0 & 0 & 15 & 0 & 1 \\ 1 & 1 & 1 & 0 & 15 & 0 & 1 \\ 1 & 1 & 0 & 1 & 15 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{bmatrix} = X_{ij}\beta$$

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{bmatrix} = X_{ijk}\alpha$$

As colunas de X_{ij} são o intercepto, I(Grupo=2), I(Etapa=2), I(Etapa=3), Idade, I(Questão85=Pouco ou Não), I(Questão82=Uma vez ao dia). Já as colunas de X_{ijk} são I(Mesmo domicílio=1) que é o intercepto, I(Etapas=1-2), I(Etapas=1-3), I(Etapas=2-3) e NCD - 2.

Com os resultados apresentados na Tabela 5, pode-se concluir o seguinte:

- Praticamente não existe diferença dos valores $\hat{\beta}$ e $ep(\hat{\beta})$ entre os modelos ajustados, porém entre os modelos que as medidas de associação podem ser comparadas, ALR e ORTH, observa-se uma maior variação entre os valores de $\hat{\alpha}$ e $ep(\hat{\alpha})$.
- Os três modelos ajustados não apresentaram evidências significativas de diferenças entre os dois grupos.
- Independentemente do modelo ajustado, na etapa 2 a chance de infecção parasitológica é de aproximadamente 1.6 vezes a chance da etapa 1, sendo que não existe diferença significativa entre as etapas 1 e 3.

- Para os três modelos ajustados, a cada mês que se aumenta na idade da criança, espera-se um aumento médio de 1.02 vezes na chance de infecção parasitológica.
- A baixa frequência ou a não lavagem das mãos das pessoas que cozinham para as crianças aumentam a chance de infecção parasitológica quando comparada às pessoas que lavam sempre a mão, sendo os valores $\hat{\beta}$ e $ep(\hat{\beta})$ dos modelos ajustados bem similares.
- A frequência de banho não exerce influência significativa sobre a infecção parasitológica, embora os p-valores 0.056, 0.055 e 0.051 respectivamente dos métodos GEE2, ALR e ORTH estejam bem próximos da significância.
- Os resultados não significativos para as medidas de associação intra-indivíduo possivelmente estão refletindo o tratamento realizado nas crianças quando elas apresentavam alguma infecção parasitológica, enquanto que o resultado não significativo intradomicílio pode ser explicado pela baixa quantidade de domicílios com mais de uma criança.

Tabela 5. Ajuste dos modelos propostos para os dados sobre infecção parasitológica.

Modelos	GEE2				ALR				ORTH (λ=0,031)			
Estrutura da média	**β**	**exp(β)**	**ep(β)**	**P-valor**	**β**	**exp(β)**	**ep(β)**	**P-valor**	**β**	**exp(β)**	**ep(β)**	**P-valor**
Intercepto	-1,702	-	0,164	0,000	-1,708	-	0,165	0,000	-1,712	-	0,165	0,000
Grupo = Sem cisterna	0,096	1,101	0,121	0,426	0,097	1,102	0,121	0,422	0,099	1,104	0,121	0,415
Etapa = 2	0,486	1,625	0,138	0,000	0,486	1,626	0,138	0,000	0,485	1,624	0,138	0,000
Etapa = 3	0,019	1,019	0,154	0,901	0,021	1,021	0,154	0,892	0,021	1,021	0,154	0,892
Idade criança (meses)	0,020	1,020	0,004	0,000	0,020	1,020	0,004	0,000	0,020	1,020	0,004	0,000
Questão85 = Pouca ou não	1,196	3,306	0,253	0,000	1,195	3,304	0,253	0,000	1,200	3,322	0,254	0,000
Questão82 = Uma vez ao dia	-0,349	0,705	0,182	0,055	-0,347	0,707	0,182	0,056	-0,353	0,703	0,181	0,051
Estrutura da dependência	**α**	**$g^{-1}(\alpha)$**	**ep(α)**	**P-valor**	**α**	**exp(α)**	**ep(α)**	**P-valor**	**α**	**exp(α)**	**ep(α)**	**P-valor**
Intercepto (Mesmo Domicílio)	0,104	0,050	0,091	0,253	0,218	1,244	0,237	0,230	0,186	1,204	0,225	0,409
Mesmo Indivíduo = Etapas 1-2	-0,163	-0,081	0,135	0,228	-0,365	0,694	0,354	0,303	-0,314	0,730	0,343	0,359
Mesmo Indivíduo = Etapas 1-3	-0,207	-0,103	0,141	0,141	-0,446	0,640	0,383	0,244	-0,388	0,679	0,375	0,302
Mesmo Indivíduo = Etapas 2-3	-0,144	-0,072	0,134	0,283	-0,336	0,715	0,368	0,362	-0,281	0,755	0,370	0,448
NCD - 2	-0,068	-0,034	0,057	0,237	-0,178	0,837	0,155	0,251	-0,146	0,864	0,182	0,423

4.2 Ecologia de Microorganismos

Essa aplicação refere-se a um estudo da ecologia de microorganismos, em que a presença ou ausência de um grupo de fungos foi medida em 5 diferentes locais, dois no Brasil e três na Argentina. Em cada local era realizado um transecto e vinte plantas eram selecionadas a cada 5 metros aproximadamente. Para cada planta foram selecionadas 5 folhas, sendo que a coleta dos fungos foi realizada em 6 diferentes fragmentos da folha, cada fragmento com propriedade biológica específica. Dessa forma se tem 600 medidas dentro de cada local, totalizando em 3000 medidas ao longo do estudo. Durante todo o estudo a prevalência do grupo de fungos foi 6.8% na Argentina e 12.7% no Brasil.

Nesse estudo o objetivo principal é obter medidas de associação do grupo de fungos, intralocal, intraplanta, intrafolha e intrafragmento, sendo que ainda se deseja testar a hipótese de que à medida que se aumenta a distância entre as plantas de um mesmo local, diminui-se a associação do grupo de fungos. O objetivo secundário é verificar se a prevalência do grupo de fungos é diferente entre Brasil e Argentina.

Na Figura 2, com o objetivo de verificar de forma descritiva se os fungos tendem a aparecer em aglomerados entre plantas mais próximas ou intraplanta, intrafolha e intrafragmentos, todas as respostas de um mesmo local de coleta foram combinadas dois-a-dois, o que possibilitou estimar$Pr(Y_j = 1, Y_k = 1)$ e relacionar com as variáveis "Mesma Planta", "Mesma Folha", "Mesmo Fragmento" e a "Distância entre duas plantas". Cabe destacar que o fato de o grupo de fungos aparecer em aglomerados entre plantas mais próximas ou intraplanta, intrafolha e intrafragmentos é o que induz a estrutura de dependência entre as respostas de um mesmo local de coleta.

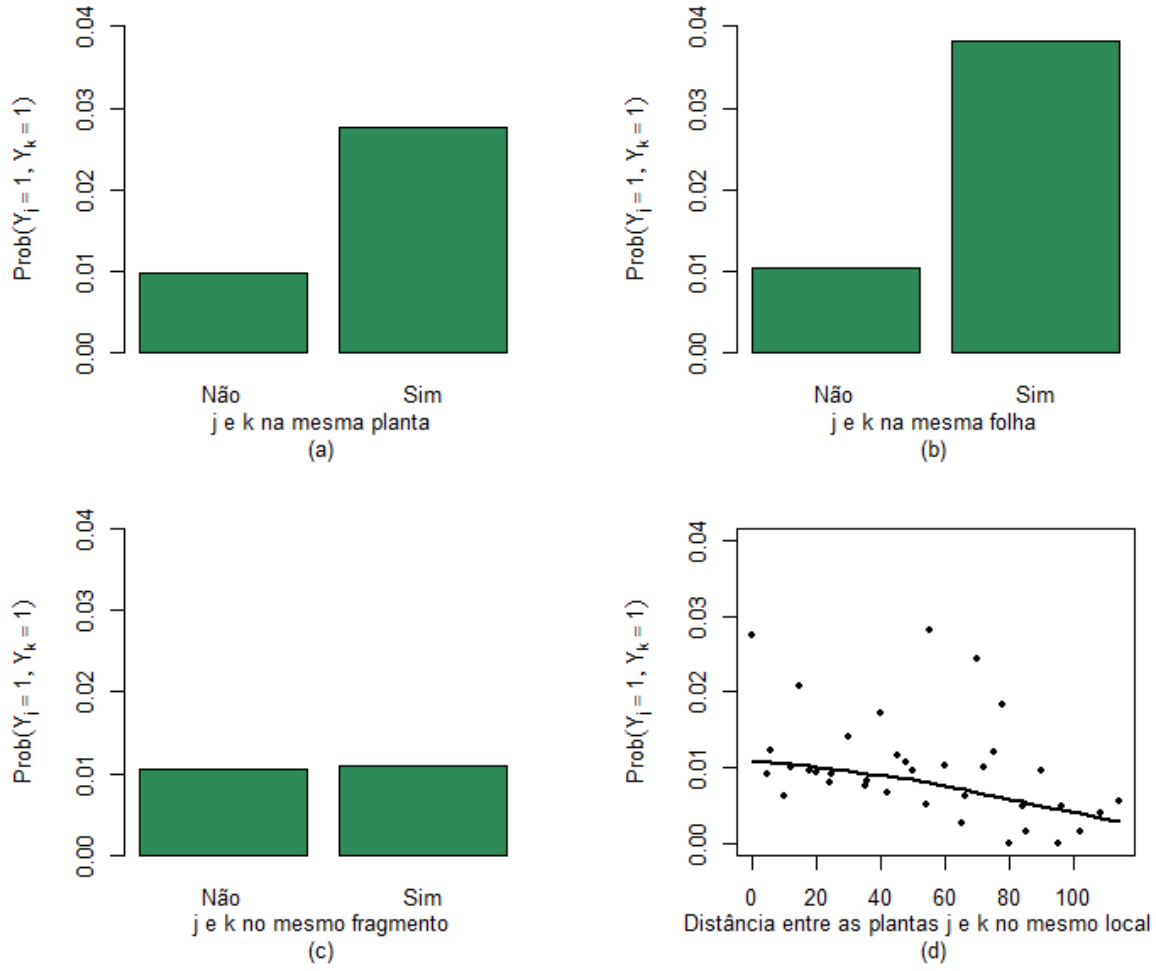

Figura 2. (a)Gráfico de barras para $Pr(Y_j = 1, Y_k = 1)$ onde j e k denotam dois indivíduos na mesma planta, (b) na mesma folha e (c) no mesmo fragmento. (d)diagrama de dispersão com o alisamento via função `lowess()` para $Pr(Y_j = 1, Y_k = 1)$ e a distância entre as plantas de um mesmo local de coleta.

Após uma análise da Figura 2, pode-se verificar que a probabilidade de se encontrar dois fungos na mesma planta é maior que a de encontrar dois fungos em plantas diferentes, sendo que à medida que se aumenta a distância entre duas plantas em um mesmo local de coleta a $Pr(Y_j = 1, Y_k = 1)$ diminui praticamente de forma linear. Comparando $Pr(Y_j = 1, Y_k = 1)$ entre observações da mesma folha e de folhas diferentes, a diferença é ainda maior que ao nível da planta, enquanto que ao nível do fragmento aparentemente não existe diferença. É interessante destacar que como a coleta dos fungos era realizada em 6 diferentes fragmentos da folha para medir a associação no mesmo fragmento, envolveu a comparação de diferentes folhas, podendo ser da mesma ou de diferentes plantas. Com essa análise descritiva, parece haver evidências de que os fungos se apresentam em aglomerados, indicando a presença de associação ao nível do local (através da distância entre as plantas), da planta e da folha. Um importante resultado dessa análise descritiva para a modelagem realizada posteriormente é a evidência que a influência da distância entre duas plantas sobre o parâmetro

de associação é linear.

Para medir a associação do grupo de fungos intralocal, de forma condicional à distância entre as plantas, intraplanta, intrafolha e intrafragmento, assim como para avaliar se a prevalência do grupo de fungos foi diferente entre Brasil e Argentina, para os métodos ALR e ORTH, foram utilizadas as seguintes estruturas para a média e associação:

$$logit Pr(Y = 1) = \beta_0 + \beta_1 I(\text{País} = \text{Brasil}),$$

$$LogOR(Y_j, Y_k) = \begin{cases} \alpha_1 \text{I(Mesmo Local=1)} + \alpha_5 \text{Distância}_{jk} + \alpha_4 \text{I(Mesmo Fragmento=1)}, \\ \text{se j e k forem plantas diferentes no mesmo local de coleta,} \\ \alpha_1 \text{I(Mesmo Local=1)} + \alpha_2 \text{I(Mesma Planta=1)} + \alpha_4 \text{I(Mesmo Fragmento=1)}, \\ \text{se j e k forem folhas diferentes na mesma planta,} \\ \alpha_1 \text{I(Mesmo Local=1)} + \alpha_2 \text{I(Mesma Planta=1)} + \alpha_3 \text{I(Mesma Folha=1)}, \\ \text{se j e k forem fragmentos diferentes na mesma folha,} \end{cases}$$

sendo que a variável distância está em decametros (Mínimo=0, Máximo=11.4).

Para essa aplicação utilizando a mesma estrutura apresentada para os métodos ALR e ORTH, apenas trocando a função de ligação da estrutura de associação, o método GEE2 não foi viável computacionalmente com o pacote `geepack` no software R. A inviabilidade computacional é devida ao esforço exigido pelas 600 medidas dentro de cada local. Na seção 3 abordamos sobre o esforço computacional dos métodos GEE2, ALR e ORTH.

Os modelos ajustados pelos métodos ALR e ORTH retratam muito bem os dados apresentados na Figura 2, sendo que com os resultados exibidos na Tabela 6, podemos concluir que:

- No mesmo local de coleta se, se observa um fungo em uma planta, a cada metro que se aumenta na distância entre a segunda planta, a chance de se observar outro fungo nessa diminui significativamente independente do modelo ajustado.
- Para os três modelos ajustados, no mesmo local de coleta se, se observa um fungo em uma planta, a chance de se observar outro fungo na mesma planta é de aproximadamente 3.6 vezes a chance de se observar outro fungo em outra planta.
- Para os três modelos ajustados, no mesmo local de coleta e na mesma planta se, se observa um fungo em uma folha, a chance de se observar outro fungo na mesma folha é de aproximadamente 2.5 vezes a chance de se observar outro fungo em outra folha.
- Não existe associação significativa intrafragmento para o método ALR, porém com o método ORTH tem-se uma grande proximidade do nível de 5% de significância, com o p-valor igual a 0.055.
- Para os três modelos ajustados, pode-se considerar que a chance de se encontrar esse tipo de fungo no Brasil é cerca de 4 vezes a chance de se encontrar esse tipo de fungo na Argentina.

Tabela 6. Ajuste dos modelos propostos para os dados sobre a ecologia de microorganismos.

Modelos	ALR				ORTH (λ=0,0002)			
Estrutura da média	**β**	**exp(β)**	**ep(β)**	**P-valor**	**β**	**exp(β)**	**ep(β)**	**P-valor**
Intercepto	-3,343	-	0,364	0,000	-3,345	-	0,3633	0,000
Pais=Brasil	1,461	4,310	0,367	0,000	1,464	4,324	0,366	0,000
Estrutura da dependência	**α2**	**exp(α2)**	**ep(α2)**	**P-valor**	**α2**	**exp(α2)**	**ep(α2)**	**P-valor**
Intercepto(Mesmo Local)	0,025	1,025	0,205	0,904	0,031	1,032	0,059	0,599
Mesma Planta	1,277	3,587	0,392	0,001	1,293	3,644	0,398	0,001
Mesma Folha	0,918	2,505	0,351	0,009	0,898	2,454	0,340	0,008
Mesmo Fragmento	0,075	1,078	0,048	0,117	0,077	1,080	0,040	0,055
Distância entre Plantas (10m)	-0,033	0,968	0,013	0,011	-0,035	0,966	0,007	0,000

5 Discussão

Nesse trabalho nós apresentamos modelos marginais e seus procedimentos de estimação para tratar dados correlacionados binários, especialmente quando temos estruturas hierárquicas, sendo que o foco dos modelos propostos é realizar inferência para a estrutura da média e para os pares de associação. Quando o problema prático envolve estruturas hierárquicas ou mais de uma classe de agrupamento o método proposto por Liang e Zeger (1986) torna-se inviável para acomodar as estruturas de dependência. O método proposto por Prentice(1988) já permite acomodar essas múltiplas estruturas de dependência e ainda realizar inferências para elas. Porém, quando se trata de respostas binárias, deseja-se medir as associações principalmente através de razões de chances, o que não é permitido por essa metodologia.

Nossos resultados indicaram não haver diferença entre o desempenho dos métodos GEE2, ALR e ORTH nas estimativas da média, associação e de seus respectivos erros padrões. Porém, para um n_i grande, próximo de 250, o método proposto por Prentice(1988) se mostrou inviável computacionalmente, enquanto que os métodos ALR e ORTH se mostraram eficazes. Nossos resultados também indicaram que quando o objetivo principal da pesquisa estiver na estrutura de associação, deve-se ter um cuidado especial na modelagem da estrutura da média, pois sua mal especificação pode induzir uma falta de consistência nas medidas de associação.

A questão computacional continua sendo uma limitação para os métodos ALR e ORTH. Em nossa aplicação, devido à estrutura hierárquica de agrupamento, tinhamos 600 medidas repetidas em 5 diferentes locais, resultando em 898500 linhas na matriz de delineamento e horas de espera para obter os resultados. A busca por métodos ainda mais eficientes computacionalmente para tratar modelos com respostas binárias na presença de múltiplas estruturas hierárquicas deve continuar, uma vez que estudos com a presença de muitos grupos com elevado número de medidas repetidas são comuns nas diversas áreas do conhecimento.

Referências

[1] Carey, V. Regression Analysis of Large Binary Clusters. *PhD thesis dissertation, The Johns Hopkins University* , 1992.

[2] Carey, V., Zeger, S. L & Diggle P., Modelling Multivariate Binary Data with Alternating Logistic Regressions. *Biometrika*, **80**, 517−526, 1993.

[3] Fonseca, J. E. Implantação de cisternas para armazenamento de água de chuva e seus impactos na saúde infantil: um estudo de coorte em Berilo e Chapada do Norte, Minas Gerais. *Mestrado em Saneamento, Meio Ambiente e Recursos Hídricos − Escola de Engenharia, Universidade Federal de Minas Gerais, Belo Horizonte* , 2012.

[4] Qaqish B.F. A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika*, **90**, 455−463, 2003.

[5] Qaqish B.F., Zink, R. C. & Preisser J.S. Orthogonalized Residuals for Estimation of Marginally Specified Association Parameters in Multivariate Binary Data *Scandinavian Journal of Statistic*, **39**, 515−527, 2012.

[6] Kuk, A.Y.C. Permutation invariance of alternating logistic regression for multivariate binary data. *Biometrika* , **91**, 758−761, 2004.

[7] Liang, K. Y., Zeger, S. L. Longitudinal Data Analysis using generalized linear models. *Biometrika*, **73**, 13−22, 1986.

[8] Liang, K.Y., Zeger, S.L. & Qaqish, B., Multivariate regression analyses for categorical data. J. R. Statist. *Soc. B* **54**,–40, 1992.

[9] Lipsitz, S. R., Laird, N. M, Harrington, D. P. Generalized Estimating Equations for Correlated Binary Data: Using the Odds Ratio as a Measure os Association . *Biometrika*, **78**, 153−160, 1991.

[10] Mardia, K. V., Some contributions to the contingency-type bivariate distributions. *Biometrika*, **54**, 35−49, 1967.

[11] Lipsitz, S.R. & Fitzmaurice, G.M., Estimating equations for measures of association between repeated binary responses. *Biometrics*, **52**, 03−12, 1996.

[12] Molenberghs, G. & Verbeke, G., Models for Discrete Longitudinal Data. *Springer*, 2005.

[13] Prentice, R.L. Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 033−48, 1988.

[14] Prentice, R.L. & Zhao L.P. Estimating equations for parameters in mean and covariates of multivariate discrete and continuous responses. *Biometrics* **48**, 25−839, 1991.

[15] R Development Core Team, 2012. R: A Language and Environment for Statistical Computing. Viena, Austria: R Foundation for Statistical Computing. *http://www.R-project.org/.*

[16] Zink, R. C. Correlated Binary Regression Using Orthogonalized Residuals. *PhD thesis, University of North Carolina, Chapel Hill* , 2003.

Códigos usados no software R

No apêndice somente foi disponibilizado os códigos do R para as aplicações. Os códigos das simulações ficaram grandes demais para serem apresentados. Caso seja de interesse do leitor solicite os códigos das simulações ao autor.

```
#=====================================
# 1ª Aplicação:  Infecção Parasitológica
#=====================================

require(orth)
require(geepack)
banco <- read.csv2 ("C:/Users/André/Desktop/Mestrado/Dissertação/Aplicação/dados.csv")
attach(banco)

# Construindo a Matriz de Delineamento X_ijk

n <- as.vector(table(as.factor(banco$Número.do.domicílio)))
last <- cumsum(n)
first <- last - n + 1
Design <- NULL
for ( i in 1:length(n) )
{
n.i <- n[i]
id.i <- banco$Número.do.domicílio[first[i]]
id.id <- banco$Id[first[i]:last[i]]
age.i <- banco$Etapa[first[i]:last[i]]
PD.i <- banco$Pdomicílio[first[i]:last[i]]
l <- 1
if (n.i == 1) z.i <- cbind(NA,NA,NA,NA)
else
{

# note que:  ch2(m) = m(m - 1)/2

id.i <- rep(id.i, choose(n.i, 2))
z.i1 <- rep(NA, choose(n.i, 2) )
z.i2 <- rep(NA, choose(n.i, 2) )
z.i3 <- rep(NA, choose(n.i, 2) )
for( j in seq(1, n.i - 1) )
{
for( k in seq(j+1, n.i) )
{
z.i1[l] <- paste(age.i[j],-", age.i[k])
z.i2[l] <- ifelse(id.id[j]- id.id[k]==0,1,0)
z.i3[l] <- as.numeric(PD.i[1])
l <- l+1
}
}
z.i <- cbind(id.i, z.i2, z.i1, z.i3)
}
Design <- rbind(Design, z.i)
}
Design <- data.frame(Design)
```

```
colnames(Design) <- c("U", "ID", "Ctempo", "PD")
DesignF <- na.omit(Design)

# Criando a variável mesmo indivíduo ou não, e se mesmo indivíduo, entre quais tempos.

var1<- paste(DesignF$ID, DesignF$Ctempo)
MT <- factor(ifelse(var1=="1 1 - 2", "1 - 2", ifelse(var1=="1 1 - 3", "1 - 3",
ifelse(var1=="1 2 - 3", "2 - 3", "0"))))
DesignF$MT <- MT

# Centralizando o número de domicílio por 2.

DesignF$PD1 <- as.numeric(DesignF$PD)-2
z <- cbind(rep(1,dim(DesignF)[1]), ifelse(DesignF[,5]=="1 - 2",1,0),
ifelse(DesignF[,5]=="1 - 3",1,0),ifelse(DesignF[,5]=="2 - 3",1,0),
(as.numeric(DesignF[,6])))
colnames(z) <- c("Intercepto", "MIT12", "MIT13", "MIT23", "PD")

# MIT12 significa:  mesmo indivíduo entre os tempos 1 e 2, e assim sucessivamente.

z0 <- data.frame(z) # para usar orth a matriz design deve ser um data.frame

# ORTH

orth0 <- orth(Resposta ~ GrupoB + factor(Etapa) + Idade.criança + questão85c +
questão82b, data=banco, formula.z= ~ MIT12 + MIT13 + MIT23 + PD,
dataz=z0, id=Número.do.domicílio, estLam=T)
summary(orth0)

# ALR

ordgee0 <- ordgee(ordered(Resposta) ~ GrupoB + factor(Etapa) + Idade.criança +
questão85c + questão82b, id=Número.do.domicílio, mean.link="logit",
data=banco, corstr=("userdefined"), z = z )
summary(ordgee0)

# GEE2 (Prentice)

geeglm0 <- geese(Resposta ~ GrupoB + factor(Etapa) + Idade.criança + questão85c +
questão82b, id=Número.do.domicílio, family=binomial,cor.link = "fisherz",
data=banco, corstr=("userdefined"), z = z )
summary(geeglm0)

#=====================================
# 2ª Aplicação:  Ecologia de Microorganismo
#=====================================

dados<- read.csv2("C:/Users/André/Desktop/Aline Vaz/Mestrado/dados.csv")
attach(dados)

# Construindo a Matriz de Delineamento X_ijk

n <- as.vector(table(as.factor(dados$Pl)))
last <- cumsum(n)
first <- last - n + 1
Design <- NULL
for ( i in 1:length(n) )
{
n.i <- n[i]
```

```
id.i <- dados$Pl[first[i]]
ind.i <- dados$Ind[first[i]:last[i]]
fo.i <- dados$Fo[first[i]:last[i]]
fr.i <- dados$Fr[first[i]:last[i]]
dis.i <- dados$dis[first[i]:last[i]]
resp.i <- dados$Y[first[i]:last[i]]
l <- 1
if (n.i == 1) {z.i <- cbind(NA,NA,NA,NA,NA)}
else
{
# Note que:  ch2(m) = m(m - 1)/2
id.i <- rep(id.i, choose(n.i, 2))
z.i1 <- rep(NA, choose(n.i, 2))
z.i2 <- rep(NA, choose(n.i, 2))
z.i3 <- rep(NA, choose(n.i, 2))
z.i4 <- rep(NA, choose(n.i, 2))
z.i5 <- rep(NA, choose(n.i, 2))
for( j in seq(1, n.i - 1))
{
for( k in seq(j+1, n.i))
{
z.i1[l] <- ifelse(ind.i[j] - ind.i[k]==0,1,0)
z.i2[l] <- ifelse(fo.i[j] - fo.i[k]==0,1,0)
z.i3[l] <- ifelse(fr.i[j] == fr.i[k],1,0)
z.i4[l] <- abs(dis.i[j] - dis.i[k])
z.i5[l] <- ifelse(resp.i[j] & resp.i[k]==1,1,0)
l <- l+1
}
}
z.i <- cbind(id.i, z.i1, z.i2, z.i3, z.i4, z.i5)
}
Design <- rbind(Design, z.i)
}
Design<- data.frame(Design)
DesignF<- Design
colnames(DesignF)<- c("Local", "Ind", "Folhas", "Frag", "Dist", "Resp")
# Análise descritiva para estrutura de associação
par(mfrow=c(2,2))
par(mar=c(5,5,1.5,3))
barplot(tapply(DesignF$Resp, DesignF$Ind, mean), col="seagreen4",
xlab="j e k na mesma planta(a)",
ylab=expression(Prob(Y[j]==1,Y[k]==1)), ylim=c(0,0.04), names=c("Não","Sim"))
barplot(tapply(DesignF$Resp, DesignF$Folhas, mean), col="seagreen4",
xlab="j e k na mesma folha(b)", ylab=expression(Prob(Y[j]==1,Y[k]==1)),
ylim=c(0,0.04), names=c("Não","Sim"))
barplot(tapply(DesignF$Resp, DesignF$Frag, mean), col="seagreen4",
xlab="j e k no mesmo fragmento(c)", ylab=expression(Prob(Y[j]==1,Y[k]==1)),
ylim=c(0,0.04), names=c("Não","Sim"))
```

```
plot(sort(unique(DesignF$Dist)), tapply(DesignF$Resp,DesignF$Dist,mean),
pch=20, ylab=expression(Prob(Y[j]==1,Y[k]==1)), ylim=c(0,0.04),
xlab="Distância entre as plantas j e k no mesmo local(m)(d)")
lines(lowess(sort(unique(DesignF$Dist)), tapply(DesignF$Resp,DesignF$Dist,mean))

z0a<- cbind(rep(1,dim(DesignF)[1]), DesignF[,2], DesignF[,3], DesignF[,4], DesignF[,5])
Pais<- factor(ifelse(dis < 2320400, "Brasil", "Argentina"))
data1<- data.frame(Y, Pais, Pl)

# O método abaixo não roda devido ao esforço computacional.  O R irá travar!

# GEE (Prentice)

geeglm0 <- geese(Y ~ Pais , id=Pl, family="binomial",
corstr=("userdefined"), z = z0a , data=data1)
summary(ordgee0a)

# ALR

ordgee0 <- ordgee(ordered(Y) ~ Pais , id=Pl, mean.link="logit",
corstr=("userdefined"), z = z0a , data=data1)
summary(ordgee0)

# ORTH

orth0 <- orth(Y ~ Pais, data=data1, formula.z= ~ Ind + Folhas + Frag + Dist,
dataz=DesignF, id=Pl, maxiter=30, estLam=T)
summary(orth0)
```