As Equações de Estimação Generalizadas e Aplicações

Este exemplar corresponde a redação final da dissertação devidamente corrigida e defendida pela **Sra. Lusane Leão Baia** e aprovada pela Comissão Julgadora.

Campinas, 11 de novembro de 1997

Profa. Dra. Eliana H.de Freitas Marques
(Orientadora)

Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica, UNICAMP, como requisito parcial para obtenção do Título de MESTRE em Estatística.

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO IMECC DA UNICAMP

Baia, Lusane Leão

B149e As equações de estimação generalizadas e aplicações / Lusane Leão Baia -- Campinas, [S.P. :s.n.], 1997.

Orientador : Eliana Heiser de Freitas Marques

Dissertação (mestrado) - Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Estudos longitudinais. 2. Correlação (Estatística). I. Marques, Eliana Heiser de Freitas. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Universidade Estadual de Campinas - UNICAMP
Instituto de Matemática Estatística e Computação Científica - IMECC

As Equações de Estimação Generalizadas e Aplicações

Lusane Leão Baia

Profa. Dra. Eliana Heiser de Freitas Marques
Orientadora

Nov/97

Dissertação de Mestrado defendida e aprovada em 11 de novembro de 1997

pela Banca Examinadora composta pelos Profs. Drs.

Prof (a). Dr (a). ELIANA HEISER DE FREITAS MARQUES

Prof (a). Dr (a). RINALDO ARTES

Prof (a). Dr (a). MÁRIO ANTONIO GNERI

Aos meus amados pais, incansáveis incentivadores, que têm sido exemplo de amor, compreensão, força e sabedoria, dedico este trabalho.

AGRADECIMENTOS

A Deus, criador dos céus e da terra, que me conduziu durante estes anos e me concedeu esta vitória.

Aos meus amados pais, Nisan e Isis, que sempre me apoiaram e me estimularam com palavras sábias, em todos os momentos.

Aos meus queridos irmãos, Nisan Jr. e Ana Maria, pela compreensão e apoio.

À minha orientadora, Profa. Eliana Marques, pela força, dedicação, paciência e competência em me orientar nesta jornada.

À Mary e à Renatinha, amigas verdadeiras de todas as horas.

Aos amigos sempre presentes nos momentos de alegrias e tristezas.

Aos colegas do mestrado em Estatística, especificamente à Lucila e à Rosemeire pelo companheirismo e disposição em me ajudar.

À Dra.Ilka Boin que colocou à minha disposição os dois conjuntos de dados usados nesta dissertação e sempre, com tanta amabilidade, me orientou esclarecendo todos os termos e técnicas utilizadas na área médica.

Aos colegas de trabalho, por toda compreensão e companheirismo.

À Isabel, que com muita presteza e competência, usou do seu precioso tempo para fazer a revisão deste trabalho.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES, orgão financiador dos meus estudos nestes anos de pesquisa.

Ó profundidade da riqueza, tanto da sabedoria,
como do conhecimento de Deus!
Quão insondáveis são os seus juízos e quão
inescrutáveis os seus caminhos !
Porque Dele e por meio Dele e para Ele são
todas as coisas. A Ele, pois, a glória eternamente.
Amém.

(Romanos 11:33,36)

SUMÁRIO

Capítulo I. Introdução 8

Capítulo II. As Equações de Estimação Generalizadas 16

2.1. Introdução ---------- 16

2.2. O Método das Equações de Estimação Generalizadas ---------- 20

2.2.1. A Função de Quase Verossimilhança ---------- 20

2.2.2. As Equações de Estimação Generalizadas ---------- 21

2.2.3. As Equações de Estimação Generalizadas de Segunda Ordem ---------- 27

2.3. Métodos Clássicos de Estimação e as Equações de Estimação Generalizadas ---------- 29

2.3.1. O Método dos Quadrados Mínimos e as Equações de Estimação Generalizadas ---------- 29

2.3.1.1. Estimação do Modelo dos Quadrados Mínimos em Dois Estágios ---------- 33

2.3.1.2. Exemplo ---------- 34

2.3.1.3. Estudos de Simulação---------- 37

2.3.2. Os Quadrados Mínimos Ponderados (QMP) e as Equações de Estimação Generalizadas ---------- 40

2.3.2.1. Outros Estudos de Simulações ---------- 45

Capítulo III. Aplicações Práticas 49

3.1. Programas Computacionais ---------- 49

3.1.1. RMGEE (Repeated Measures using Generalized Estimating Equations) - Descrições Gerais ---------- 50

3.1.1.1. Estrutura do Arquivo de Dados ---------- 51

3.1.2. GEE : A Macro do SAS ---------- 52

3.1.2.1. Estrutura do Arquivo de Dados ---------- 53

3.1.3. Outros Programas---------- 54

3.2. Exemplos Ilustrativos ---------- 55

3.2.1. Dados Utilizados no Trabalho ---------- 57

3.2.2. Comentários Finais ---------- 71

Referências Bibliográficas 72

Apêndices

A Funções de Estimação ---------- **A1**
B Matrizes de Correlação ---------- **B1**
C Listagem dos Programas ---------- **C1**

RESUMO

A realização deste trabalho tem por finalidade apresentar aplicações práticas do método das equações de estimação generalizadas (EEG), como uma nova alternativa para análise de dados. A proposta inclui breve resumo da teoria, descreve programas computacionais existentes e apresenta análise de dois conjuntos de dados reais. A intenção é colocar ao alcance do profissional de estatística mais uma ferramenta para análise de dados complexos, aplicando a metodologia em dois conjuntos de dados do Hospital de Clínicas (HC) da Universidade Estadual de Campinas - Unicamp. Por se tratar de um assunto que envolve uma teoria mais complexa, as EEG têm sido mais usadas e descritas em revistas científicas teóricas, dificultando o uso das mesmas por pesquisadores de outras áreas nos seus dados de pesquisa. A motivação deste trabalho foi estudar esta técnica e fazer aplicações que respondessem questões resultantes de dados levantados por profissionais de saúde brasileiros.

ABSTRACT

The purpose of dissertation is to present some practical applications using the Generalized Estimating Equations (GEE) method as a new alternative to data analysis. This proposal includes a brief summary of the theory, a description of computer programs available and an analysis of two sets of data based on actual findings colleted medical School Hospital at the Campinas State University- Unicamp. It also provides one more tool to be used by professionals in Statistics in their analysis of complex data as well as by professionals from other areas who, because the GEE consists of a complex theory and is mostly described scientific journals, may have difficulty in using it. It result of an attempt to answer questions presented by the Brazilian health professionals.

Capítulo I

Introdução

Diferentemente do avanço obtido pela metodologia estatística para dados contínuos, ocorrido entre o início e o meio deste século, somente nos últimos 30 anos tal avanço se verificou na metodologia para dados discretos. A forte influência de Fisher, Yule e de outros estatísticos, em experimentos de agricultura e ciências biomédicas, assegurou uma adoção difundida da técnica de regressão e modelos de análise de variância, em meados do século XX . Por outro lado, apesar dos importantes artigos na passagem do século de Karl Pearson e Yule, sobre a associação entre variáveis categóricas, houve pouco trabalho subseqüente em modelos de resposta categorizada.

Estudiosos renomados na história da estatística, tais como Fisher, Neyman, Cochran e Bartlett, deixaram grandes contribuições para a literatura de dados categorizados No entanto,

modelos para respostas categorizadas, análogos aos de regressão, receberam pouca atenção até o início da década de 70. Os desenvolvimentos mais recentes de métodos para dados categorizados foram obtidos, em grande parte, pela sofisticação das metodologias no campo das ciências sociais e biomédicas, por meio da análise destes dados para variáveis, tais como: atitudes, opiniões, características demográficas, estágio de uma doença e outras. Tal foi a importância e estreita relação das ciências sociais e de saúde no desenvolvimento dos modelos para dados categorizados, que grandes profissionais da área de estatística, como Leo Goodman, Shelby Haberman, Frederick Berkson, Jerome Cornfield e Gary Kock, utilizaram-se dos mesmos para maior aprofundamento dos seus estudos.

Nos dias atuais, observa-se que conjuntos de dados com respostas categorizadas e estruturas complexas vêm se tornando cada vez mais freqüentes em diversas áreas do conhecimento (Sociologia, Biologia, Economia, Psicologia, Medicina e Epidemiologia). Nas áreas de saúde e de ciências sociais, critérios de respostas em pesquisas têm, em geral, natureza categórica e são mais complexos que um simples resultado binário. Mais especificamente, esses dados apresentam uma estrutura vinda de respostas politômicas repetidas com as observações, ocorrendo de forma agregada e induzindo uma possível estrutura de correlação.

Novas técnicas para análise de dados categorizados com medidas repetidas começam a aparecer e, é por esta razão, que propomos neste trabalho o estudo das equações de estimação generalizadas (EEG) (LIANG & ZEGER, 1986), por ser uma técnica mais avançada para analisar dados tanto categorizados como contínuos, embora tratemos aqui apenas da situação de respostas categorizadas.

Os estudos com medidas repetidas definidos como aqueles nos quais a resposta de cada indivíduo é observada sob duas ou mais condições, hoje em dia já competem com estudos que possuem uma única medida e têm como objetivos principais a caracterização de modelos de

resposta individual, mudança no tempo e investigação dos efeitos das covariáveis. Nesses estudos, chamados longitudinais, os indivíduos são avaliados ao longo de uma dimensão específica, em geral o tempo, distância de uma certa origem ou dosagem de uma substância. Em um sentido mais amplo, a dimensão pode ser os componentes de um conceito ou processo (KOCH et al., 1989).

WARE (1985) argumentou sobre a superioridade dos estudos longitudinais em relação aos de corte ou transversais, mencionando que aqueles estudos oferecem ao pesquisador oportunidade para controlar e uniformizar medidas de exposição histórica e outros fatores relacionados ao resultado. Mesmo em estudos que não são intencionalmente longitudinais, uma vez os indivíduos selecionados para a amostra e avaliados, muitas vezes é mais fácil e mais eficiente observá-los repetidamente do que descartá-los depois de uma medição e recomeçar com uma nova amostra mais tarde .

Por outro lado, segundo DAVIS (1993), os estudos longitudinais trazem duas dificuldades principais. Primeiro, a análise é complexa por causa da dependência entre as observações repetidas, feitas numa mesma unidade experimental. Em segundo lugar, já que o investigador normalmente não pode controlar por completo as circunstâncias para obter as medidas, os dados podem não estar balanceados ou serem parcialmente incompletos.

Entre as várias alternativas de tratamentos para dados longitudinais, duas são mais relevantes: a primeira prevê a modelagem probabilística do problema através da determinação de uma distribuição multivariada de probabilidades, supostamente adequada aos dados e assim sugere-se um estimador para os parâmetros com base nessa distribuição. Essa abordagem pode apresentar dificuldades com respeito à definição do modelo probabilístico, ou seja, em como gerar um modelo multivariado que se ajuste aos dados e tenha parâmetros facilmente estimáveis. A segunda alternativa é baseada no uso de funções de estimação (ver ARTES,1997, por exemplo) para a obtenção da estimativa de parâmetros de um modelo multivariado que não é, necessariamente,

completamente conhecido.

A dependência entre as observações repetidas num mesmo indivíduo vem sendo alvo de publicações mais recentes sobre dados categorizados (ROSNER,1989; NEUHAUS & JEWELL, 1990; RAO & SCOTT, 1992; DUNLOP, 1994). Na área de saúde, conjuntos de dados apresentando uma estrutura de grupamento ou conglomerado, em termos de repetição das medidas sobre um mesmo indivíduo, estão se tornando regra e não exceção! Nestas situações, métodos multivariados tradicionais, como por exemplo, a regressão logística, não devem ser utilizados por causa da falta de independência entre as respostas do indivíduo no conglomerado.

Dados categorizados com estruturas complexas, resultantes de esquemas amostrais envolvendo conglomerados, têm sido freqüentes na literatura e têm gerado preocupação por parte dos pesquisadores, no que diz respeito aos métodos de estimação dos parâmetros de interesse. Os métodos tradicionais de análises que ignoram a estrutura de conglomerado tendem a subestimar o erro-padrão verdadeiro das estimativas dos efeitos do tratamento. Similarmente, os testes qui-quadrado-padrões podem aumentar significativamente o erro Tipo I (RAO & SCOTT, 1992).

Vários métodos e modelos de análises têm sido sugeridos e/ou estudados para a estrutura de conglomerados. Alguns desses métodos são extensões do modelo de regressão logística para dados binários; outros trabalham com modelos de variáveis latentes para conglomerados com dados ordinais, ou ainda, modelos mistos não lineares em termos das variáveis latentes normais (CATALANO & RYAN, 1992; QU et al., 1992; QU, PIEDMONTE, MEDENDORP, 1995). A questão das estruturas complexas, dos ajustes dos modelos apropriados e dos métodos de estimação, será tratada aqui apenas para dados de resposta binária onde o indivíduo com suas medidas repetidas forma o conglomerado.

Um exemplo de estrutura de dados envolvendo conglomerados é o seguinte: deseja-se analisar um determinado tipo de material dentário para ser usado em restaurações. Existem vários

fatores que podem influenciar este material, tais como: natureza do material, habilidade do operador, higiene bucal do paciente, tipo e posição dos dentes, o tamanho das restaurações, etc. Um aspecto importante a ser considerado neste tipo de dado tem a ver com a higiene bucal de cada paciente e a estrutura de correlação que pode existir envolvendo dentes e/ou posição com respeito à boca de cada paciente. Ao levar em consideração a questão de uma possível estrutura de conglomerados para esses tipos de dados, a análise produz resultados mais sensíveis a essa estrutura, possibilitando, assim, uma melhor avaliação dos dados obtidos e, conseqüentemente, uma evolução desses estudos clínicos (MARQUES, 1987; LANDIS et al., 1988).

Um aspecto importante desses estudos é que a resposta de interesse observada para os elementos do conglomerado não é necessariamente uma variável aleatória univariada, ela pode ser bivariada ou multivariada. Um exemplo de conglomerado com resposta binária pode acontecer da seguinte maneira: deseja-se saber o efeito, no tempo, de uma certa droga analgésica em pacientes com dores reumáticas. Após a droga ter sido ingerida pelo paciente, anotam-se os resultados de alteração ou não das dores reumáticas após 10, 20 e 30 minutos, respectivamente. A variável resposta, neste caso, é binária, *sim* ou *não*, e cada paciente define um conglomerado com medidas observadas nos três tempos. A estrutura de conglomerado aparece também em estudos toxicológicos onde o conglomerado é a ninhada, e os récem-nascidos são os indivíduos dentro do conglomerado. Em estudos de doenças visuais, o conglomerado é o indivíduo, e os dois olhos são as medidas observadas no conglomerado. Então, o vetor de resposta para o conglomerado é um vetor de medidas repetidas com respostas multinomiais (LIPSITZ, KIM, ZHAO, 1994).

Métodos para analisar dados com estrutura de conglomerado e variáveis de respostas contínuas com distribuição normal multivariada têm sido amplamente usados, mas, métodos para esses tipos de estruturas e variáveis com respostas categorizadas, somente nos últimos tempos

vêm sendo apresentados na literatura. Em termos práticos, métodos mais sofisticados só agora começam a ser utilizados com o surgimento de programas computacionais.

Na tentativa de complementar e produzir extensões de técnicas já existentes, que cada vez mais atendam às necessidades de conjuntos de dados complexos com respostas categorizadas, os quais aparecem como resultado de estudos com amostras feitas ao longo do tempo, técnicas inferenciais de máxima verossimilhança e quadrados mínimos ponderados têm sido revisadas, aperfeiçoadas e estendidas. O resultado tem sido uma evolução para métodos mais sofisticados, envolvendo, por exemplo, modelos lineares generalizados e a função de quase verossimilhança. Dada a ausência de metodologias capazes de incorporar todas essas questões complexas, resultantes de tratamento abrangente dos levantamentos executados pelos pesquisadores nos dias de hoje, estes métodos, que até meados de 1996 não eram sequer abordados em pacotes mais completos, surgem na literatura estatística com aplicações.

A realização deste trabalho tem por finalidade apresentar aplicações práticas do método das equações de estimação generalizadas (EEG), como uma alternativa para análise de dados complexos. A proposta inclui breve resumo da teoria, descreve alguns programas computacionais existentes e apresenta análise de dois conjuntos de dados reais. A intenção é colocar ao alcance do profissional de estatística mais uma ferramenta para análise de dados complexos, aplicando a metodologia em dois conjuntos de dados obtidos no Hospital de Clínicas (HC) da Unicamp. Por se tratar de um assunto que envolve uma teoria mais complexa, as EEG têm sido mais usadas e descritas em revistas científicas teóricas, dificultando o uso das mesmas por pesquisadores de outras áreas nos seus dados de pesquisa. A motivação deste trabalho foi estudar esta técnica e fazer aplicações que respondessem questões resultantes de dados levantados por profissionais de saúde brasileiros.

O método das EEG, proposto por LIANG & ZEGER (1986) e ZEGER & LIANG

(1986), é apropriado para analisar resultados categorizados e contínuos. As EEG são uma técnica de estimação que leva em consideração a correlação entre as variáveis. As mesmas produzem estimadores consistentes e assintoticamente normais dos parâmetros sob a especificação correta da função ligação *(link)* e da variância em função da média, evitando, assim, a necessidade de se conhecer totalmente a distribuição multivariada dos dados. Esta técnica é utilizada quando o interesse é modelar a estrutura marginal da média. Os parâmetros relativos à estrutura de correlação através do tempo são tratados como perturbação.

As EEG são uma extensão multivariada da função de quase verossimilhança, inicialmente apresentada por WEDDERBURN (1974), e mais tarde, por McCULLAGH & NELDER (1989). Esta função não exige conhecimento da distribuição paramétrica da variável resposta. Apenas, é necessário especificar a relação entre a média e a variância das observações, supondo alguma estrutura de correlação para os dados.

Desde a publicação de LIANG & ZEGER (1986), as EEG têm sido estudadas e extensões vêm aparecendo na literatura. PRENTICE (1988) estendeu o método das EEG para dados binários correlacionados, onde a especificação destas equações permite estimativas seqüenciais dos parâmetros associados. STRAM, WEI, WARE (1988) desenvolveram modelos marginais com respostas ordinais repetidas, ajustando as regressões separadamente em cada tempo e, mais tarde, aplicaram a teoria para amostras grandes com a finalidade de obter a distribuição conjunta desses parâmetros de estimação. Suas técnicas podem ser consideradas como métodos semiparamétricos para o modelo do logito cumulativo de respostas longitudinais ordinais, e como um caso especial de independência das EEG de LIANG & ZEGER (1986). Posteriormente, ZHAO & PRENTICE (1990) identificaram a classe dos modelos exponenciais quadráticos para dados binários correlacionados, nos quais a função escore das equações de estimação é a máxima verossimilhança, introduzindo a extensão das EEG de segunda ordem. LIPSITZ, LAIRD e

HARRINGTON (1991) modificaram as equações de estimação de PRENTICE (1988) para permitir modelos de associação entre medidas repetidas, via uso da razão de chance ("*odds ratio*"). Em 1992, LIANG, ZEGER, QAQISH denominaram as EEG de LIANG & ZEGER (1986) de EEG1 e a extensão apresentada por ZHAO & PRENTICE (1990) de EEG2. As EEG2 são uma extensão das EEG1 quando não se deseja tratar a correlação como parâmetro de perturbação entre as medidas repetidas, e, dependendo do conjunto de dados, elas podem produzir estimativas mais eficientes que as EEG1. Alguns artigos recentes como LIANG et al., (1992) e QAQISH & LIANG (1992) trazem aplicações usando as EEG2, mostrando vantagens e desvantagens no uso dessas equações.

O Capítulo II traz um resumo da teoria das EEG1 e EEG2. A segunda parte do capítulo apresenta exemplos comparativos discutidos na literatura dos métodos clássicos de estimação, quadrados mínimos, quadrados mínimos ponderados e as EEG (MILLER, DAVIS, LANDIS, 1993; PARK, 1994), e que incluem estudos de simulações.

O Capítulo III mostra aplicações práticas das EEG em dados reais. A primeira parte deste capítulo descreve e informa sobre os programas computacionais existentes para o método das EEG. Na segunda parte, são mostrados e analisados exemplos ilustrativos de dois conjuntos de dados reais, onde a variável resposta é binária. Estes dados são analisados com o uso de programas computacionais particulares, cedidos pelos professores pesquisadores que os escreveram, e os resultados são apresentados e discutidos.

Posteriormente, seguem as referências bibliográficas, apêndices, complementando a teoria das EEG e também as listagens dos programas computacionais.

Capítulo II

As Equações de Estimação Generalizadas

2.1. Introdução

Este capítulo trata do método das equações de estimação generalizadas para a análise de dados categorizados, com medidas repetidas. A literatura descreve várias alternativas de modelos para dados categorizados com respostas binárias e politômicas. A seguir, são mencionadas algumas dessas metodologias e as respectivas referências bibliográficas.

A estimação pelo método de máxima verossimilhança tem sido usada para analisar dados categorizados com estrutura de conglomerado (ASHBY et al., 1992). Embora seja uma técnica antiga e bastante usada, a máxima verossimilhança apresenta algumas limitações em certos modelos, pois pode ser difícil fazer a identificação da distribuição paramétrica e, muitas vezes, a

técnica torna-se impraticável computacionalmente. Apesar das suas limitações, a máxima verossimilhança oferece estimativas e testes com propriedades assintóticas desejáveis.

Um outro método de estimação bastante utilizado é o de Quadrados Mínimos Ponderados (QMP) de GRIZZLE, STARMER, KOCH (GSK,1969) para dados discretos apresentados em categorias, que permite descrever a variação entre o conjunto de estimativas produzido para os dados, no contexto geral de modelo linear aplicado a dados categóricos. Estas estimativas incluem uma variedade de funções como proporções, médias, razões, etc. Por exemplo, em um estudo clínico de material dentário para retenção da cor de dentes restaurados com certo material X ou Z, a proporção de dentes que retém a cor para diferentes materiais pode ser uma medida de interesse. Para realizar as análises pelos QMP, é necessário ter um conjunto de estatísticas como um vetor e um estimador consistente da matriz covariância deste vetor. A técnica GSK envolve três estágios: 1) construção das funções das respostas; 2) ajuste de modelos de regressão para aquelas funções com estimação dos parâmetros por QMP; e 3) teste de hipóteses sobre as combinações lineares dos parâmetros do modelo. Esta metodologia foi também proposta por KOCH et al. (1977) e STANISH, GILLINGS, KOCH (1978) para análise de dados longitudinais com a presença ou não de dados faltantes.

Modelos log-lineares também podem ser usados para análise de dados categóricos multivariados. ROSNER (1984) apresentou um modelo de regressão logística politômica para controlar o efeito do conglomerado e de covariáveis individuais específicas quando existiu correlação entre as unidades dentro do conglomerado. Este modelo reduz-se a um modelo beta-binomial na ausência de covariáveis e a um modelo logístico com conglomerado de tamanho um para o caso de conglomerados maiores, quando a correlação não está presente. Este tipo de modelo é típico em dados oftalmológicos onde o indivíduo é um conglomerado e os olhos são as unidades dentro do conglomerado. ROSNER (1989) estendeu o modelo beta-binomial para dois

ou mais níveis hierárquicos ("*nesting*"). Uma outra técnica para modelar dados binários correlacionados envolve modelos logísticos normais mistos (PIERCE & SANDS, 1975).

Ainda, quando o número de observações para cada indivíduo é pequeno, ZHAO & PRENTICE (1990) sugeriram classes de modelos exponenciais quadráticos. Este modelo foi parametrizado em termos da média das marginais e correlação "dois a dois" para a análise de regressão de dados binários correlacionados.

A regressão logística condicional e não condicional também pode ser usada na análise de dados binários com estrutura de conglomerado. CONOWAY (1990) descreveu uma função baseada no modelo de RASCH (1960), onde é possível tratar o efeito latente como parâmetro de perturbação e obter estimadores gerais através da máxima verossimilhança condicional, sem ter que especificar a distribuição do efeito nos indivíduos. A análise condicional é relevante quando as estimativas dos parâmetros são de interesse primário e a informação sobre a distribuição dos efeitos latentes não é prioridade. Quando a distribuição dos efeitos latentes e as estimativas dos parâmetros são de interesse, uma alternativa para a análise condicional é baseada na suposição da distribuição específica para esses efeitos.

RAO & SCOTT (1992) apresentaram um método para comparar grupos independentes de conglomerados binários sujeitos à covariáveis específicas. O método é baseado nos conceitos de efeito de delineamento, "*deff*", usados em pesquisas amostrais (KISH, 1965), e assume modelos não específicos para as correlações dentro do conglomerado. É um método cujos resultados são assintoticamente corretos quando o número de conglomerados em cada grupo tende a ser infinito. Pode ser implementado usando-se algum programa computacional para análise de dados binários independentes. O método é aplicado para uma variedade de problemas da área biomédica envolvendo grupos independentes de dados binários em conglomerados. Em particular, testa a homogeneidade de proporções, estima modelos de dose-resposta, testa a

tendência nas proporções, calcula a estatística qui-quadrado de Mantel-Haenszel para independência em tabelas 2x2 e estima a razão de chance e suas variâncias quando a hipótese de independência é rejeitada. Este método não considera a suposição de estrutura de dependência entre as observações binárias dentro de cada conglomerado.

Os métodos de pseudo máxima verossimilhança também são utilizados para estimar assintoticamente a matriz de covariância dos parâmetros de regressão dentro da classe dos modelos lineares generalizados, para amostra de populações finitas com estrutura complexa (SNYDER, 1993). Na metodologia de pseudo verossimilhança encontra-se a quase verossimilhança (WEDDEBURN, 1979; McCULLAGH & NELDER, 1989).

Vários tipos de modelos específicos para dados categorizados, nos quais a correlação dentro do conglomerado é observada, são descritos na literatura (KUPPER & HOSEMAN, 1978; ANDERSEN, 1980; STIRATELLI, LAIRD, WARE, 1984; HAVE, LANDIS, HARTEZES, 1996). O método das EEG é recente e, por possuir, sob hipóteses gerais, propriedades relevantes do ponto de vista assintótico e estimativas consistentes dos parâmetros, começa a ser explorado tanto do ponto de vista teórico quanto prático. Uma das vantagens das EEG é que elas levam em consideração a questão da falta de independência entre as observações repetidas, e sua teoria mostra propriedades importantes de consistência dos estimadores. Métodos para análise de dados que ignoram a estrutura complexa tendem a subestimar os erros-padrões (WARE, 1985; ZEGER & KARIM, 1991; DUNLOP, 1994; DIGGLE, LIANG, ZEGER, 1994).

As próximas seções trazem, resumidamente, o método das equações de estimação generalizadas, e comparações entre os métodos de quadrados mínimos (QM), quadrados mínimos ponderados (QMP) e as EEG.

2.2. O Método das Equações de Estimação Generalizadas

2.2.1. A Função de Quase Verossimilhança

A função de quase verossimilhança é uma função de estimação que requer poucas suposições sobre a distribuição da variável dependente. Para definir a função de verossimilhança, especifica-se a distribuição das observações. Porém, para definir a função de quase verossimilhança, é necessário apenas especificar a relação entre a média e a variância. As EEG são baseadas na extensão da equação de quase verossimilhança. Esta extensão é importante porque, exceto para resultados aproximadamente gaussianos, existem poucas alternativas para a distribuição conjunta de medidas repetidas. A função de quase verossimilhança para dados longitudinais pode ser descrita da seguinte maneira:

Considere (y_{ij}) a variável resposta e x_{ij} o vetor $p \times 1$ de covariáveis para os tempos t_{ij}, $j = 1, 2, \ldots, n_i$ e indivíduos $i = 1, 2, \ldots, k$. Então y_i é um vetor $n_i \times 1$ $(y_{i1}, \ldots, y_{in_i})^T$ e x_i é uma matriz $n_i \times p$ $(x_{i1}, \ldots, x_{in_i})^T$ para o i-ésimo indivíduo.

Define-se μ_i como sendo a esperança de y_i e supõe-se que:

$$\mu_i = h(x_i \beta) \qquad (1)$$

onde β é um vetor de parâmetros $p \times 1$. A inversa da função h é chamada de função de ligação (McCULLAGH, 1983).

Na função de quase verossimilhança, a variância, v_i, de y_i é expressa como uma função conhecida, g, da esperança, μ_i, isto é:

$$v_i = \frac{g(\mu_i)}{\phi} \tag{2}$$

onde ϕ é o parâmetro de escala. O objetivo da função de quase verossimilhança é fazer inferência sobre β, uma vez que ϕ é tratado como parâmetro de perturbação.

As estimativas de quase verossimilhança de β podem ser obtidas através da solução do seguinte sistema de equações quase-escore:

$$S_p(\beta) = \sum_{i=1}^{k} \frac{\partial \mu_i}{\partial \beta_p} v_i^{-1} (y_i - \mu_i) = 0, \quad p = 1,2,\ldots,P. \tag{3}$$

As equações em (3) são equações escore para β quando y_i tem distribuição na forma de família exponencial. Suas soluções podem ser obtidas iterativamente pelo método dos quadrados mínimos ponderados. Os resultados são assintoticamente gaussianos sob condições de regularidade das funções de estimação (McCULLAGH, 1983; McCULLAGH & NELDER, 1989; ARTES, 1997).

2.2.2. As Equações de Estimação Generalizadas

Esta seção sumariza os principais resultados de LIANG & ZEGER (1986) e ZEGER & LIANG (1986). Seja $y_i = (y_{i1},\ldots,y_{in_i})^T$, $i = 1,2,\ldots,k$ o vetor de respostas discretas ou contínuas para cada indivíduo i, e $x_i = (x_{i1},\ldots,x_{in_i})^T$, onde x_{ij} é um vetor $p \times 1$ de covariáveis associadas com y_{ij}. Assume-se a densidade marginal de y_{ij} como:

$$\pi(y;\theta_i,\phi_1) = \exp\{\phi_i[y\theta_i - a(\theta_i) + c(y,\phi_i)]\} \tag{4}$$

onde ϕ é o parâmetro de escala, $\theta_{ij} = h(\eta_{ij})$ sendo $\eta_{ij} = x_{ij}^T \beta$, e β é o vetor $p \times 1$ dos coeficientes da regressão. O parâmetro η_{ij} está relacionado à média μ_{ij} através da função ligação g tal que $\eta_{ij} = g(\mu_{ij})$. Utilizando a equação (4), tem-se que os primeiros dois momentos de y_{ij} são $E(y_{ij}) = \mu_{ij} = a'(\theta_{ij})$ e $var(y_{ij}) = v_{ij} = a''(\theta_{ij})/\phi$.

Quando $n_i = 1$, $i = 1, 2, \ldots, k$ utiliza-se o estimador da quase verossimilhança para β, que é a solução das seguintes equações de estimação

$$\sum_{i=1}^{k} \frac{\partial \mu_i}{\partial \beta_p} v_1^{-1} (y_i - \mu_i) = 0, \quad p = 1, \ldots, P. \tag{5}$$

No caso de se ter dados longitudinais, LIANG & ZEGER (1986) introduziram a idéia da "matriz correlação de trabalho", $\mathbb{R}_i(\alpha)$ simétrica $n_i \times n_i$ positiva definida e definiram

$$\sum_i = (A_i^{1/2} \mathbb{R}_i(\alpha) A_i^{1/2})/\phi \tag{6}$$

como sendo a matriz covariância de trabalho de $\mathbf{y}_i$, onde α é um vetor $s \times 1$ que caracteriza completamente a forma de $\mathbb{R}_i(\alpha)$ podendo ser conhecido ou desconhecido. O α quando é desconhecido pode ser estimado de várias maneiras como será apresentado nas páginas seguintes. $A_i = diag(a''(\theta_{i1}), \ldots, a''(\theta_{in_i}))$ é uma matriz diagonal com dimensão $n_i \times n_i$. Quando $\mathbb{R}_i(\alpha)$ é uma matriz correlação de trabalho que descreve bem a correlação entre os dados, então $\sum_i = cov(y_i)$. As equações de estimação generalizadas podem ser definidas da seguinte forma:

$$\sum_{i=1}^{k} D_i^T \sum_i^{-1} Q_i = 0 \tag{7}$$

onde $\mathbf{D}_i = \partial\mu_i / \partial\beta$, e $\mathbf{Q}_i = y_i - \mu_i$.

Considere agora, para cada i, $U_i(\beta,\alpha) = D_i^T \sum_i^{-1} Q_i$ é semelhante a função de quase verossimilhança apresentada por WEDDERBURN (1974) e McCULLAGH (1983) exceto que aqui $\sum_i^{-1}$ não é apenas função de β mas também de α. A equação (7) pode ser reescrita como função apenas de β, substituindo α em (6) e (7) por $\hat{\alpha}(y,\beta,\phi)$ onde $k^{(1,2)}$ - um estimador consistente de α quando β e ϕ são conhecidos, isto é, $\hat{\alpha}$ para o qual $k^{1/2}(\hat{\alpha}-\alpha) = O_p(1)$. Para completar o processo, chama-se ϕ de $\hat{\phi}(\beta)$, um estimador $k^{(1,2)}$ - consistente quando β é conhecido. Conseqüentemente, (7) tem a forma:

$$\sum_{i=1}^{k} U_i[\beta, \hat{\alpha}\{\beta, \hat{\phi}(\beta)\}] = 0 \qquad (8)$$

e $\hat{\beta}$ é a solução das equações em (8).

A obtenção das estimativas de β é feita de modo iterativo. À cada interação l, define-se a variável dependente z_i como:

$$z_i = D_i \beta^l + Q_i. \qquad (9)$$

através da resolução de uma regressão linear ponderada, onde a variável dependente é $\mathbf{D}_i$, obtém-se a nova estimativa. A solução de (7) é dada por $\hat{\beta} = \left(\sum_{i=1}^{k} [D_i^T \sum_i^{-1} D_i] \right)^{-1} \left(\sum_{i=1}^{k} [D_i^T \sum_i^{-1} z_i] \right)$. Os pesos dessas regressões são $\sum_i^{-1}$ e o processo é repetido até chegar à convergência.

LIANG & ZEGER (1986) também mostraram que, sob condições gerais de regularidade, a solução de $\hat{\beta}$ para a equação (7) é assintoticamente consistente e normalmente distribuída, com:

$$k^{1/2}(\hat{\beta} - \beta) \xrightarrow{D} N(0, V_{\hat{\beta}}) \qquad (10)$$

e

$$V_{\beta}^{A} = \lim_{k \to \infty} k \left(\sum_{i=1}^{k} D_i^T \sum_i^{-1} D_i \right)^{-1} \left[\sum_{i=1}^{k} D_i^T \sum_i^{-1} cov\, y_i) \sum_i^{-1} D_i \right] \cdot \left(\sum_{i=1}^{k} D_i^T \sum_i^{-1} D_i \right)^{-1}, \quad (11)$$

sob as seguintes hipóteses:

i) $\hat{\alpha}$ é $k^{1/2}$ - consistente dado β e ϕ;

ii) $\hat{\phi}$ é $k^{1/2}$ - consistente dado β; e

iii) $| \partial \hat{\alpha} (\beta, \phi) / \partial \phi | \leq H(y, \beta)$ que é $O_p(1)$.

Os resultados acima se verificam mesmo quando a matriz correlação de trabalho $\mathbb{R}_i(\alpha)$ não corresponde à verdadeira matriz de y_i. Pode-se obter uma estimativa assintoticamente consistente de V_{β}^{A} substituindo $cov\ (y_i)$ por $\mathbf{Q}_i\mathbf{Q}_i^{\mathrm{T}}$ e substituindo β por $\hat{\beta}$ na equação (11). Este é um estimador consistente de $cov\ (\beta)$, mesmo quando $\mathbb{R}_i(\alpha)$ não é especificada corretamente (CARR & CHI, 1992). Note que, quando $\mathbb{R}_i(\alpha)$ for corretamente especificada, isto é, $\sum_i = cov\ (y_i)$, tem-se $V_{\beta}^{A} = \lim_{k \to \infty} k \left(\sum_{i=1}^{k} (D_i^T \sum_i^{-1} D_i) \right)^{-1}$, e portanto o estimador pode ser mais eficiente que o obtido em (11) (ROTNITZKY & JEWELL, 1990; FIRTH, 1992; CARR & CHI, 1992). A técnica não supõe a necessidade das observações y_i terem a mesma estrutura de correlação. A propriedade de robustez só é garantida quando existe uma pequena fração de dados faltantes ou quando estes são completamente aleatórios (FITZMAURICE, LAIRD, ROTNITZKY, 1993).

O α e o parâmetro de escala ϕ podem ser estimados através dos resíduos de Pearson definidos por:

$$\hat{r}_{ij} = \{ y_{ij} - a'(\hat{\theta}_{ij}) \} / \{ a''(\hat{\theta}_{ij}) \}^{1/2} \quad (12)$$

onde $\hat{\theta}_{ij}$ depende do valor de β. Pode-se especificar ϕ como:

$$\hat{\phi}^{-1} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \hat{r}_{ij}^2/(N-p), \qquad (13)$$

onde $N=\sum n_i$.

O estimador específico de α depende da escolha de $\mathbb{R}_i(\alpha)$. De forma geral α pode ser estimado por uma função de

$$\hat{R}_{uv} = \sum_{i=1}^{k} \hat{r}_{iu}\hat{r}_{iv}/(N-p). \qquad (14)$$

Estimadores específicos são apresentados no Apêndice B.

As EEG permitem que a estrutura de correlação entre as observações num mesmo indivíduo seja especificada de várias maneiras, através de diferentes matrizes de correlação, como por exemplo:

[1]. $\mathbb{R}_i = I_{n_i}$, onde I é uma matriz identidade $n_i \times n_i$, isto é, as observações repetidas não são correlacionadas.

[2]. $[\mathbb{R}_i]_{jk} = \alpha$, $j \neq k$, esta estrutura de correlação é chamada correlação permutável. É obtida de modelos com efeito aleatório para cada indivíduo (LIANG & ZEGER, 1986). O estimador de α para esta matriz de correlação é apresentado no Apêndice B.

[3]. $\mathbb{R}_i(\alpha)$ pode ser tridiagonal com $\mathbb{R}_{n_i,n_i-1} = \alpha_{n_i}$. Esta estrutura de correlação é equivalente ao modelo 1-dependente, ou seja, as observações são correlacionadas apenas com a observação imediatamente anterior e posterior. Para $\hat{\alpha}$ ver Apêndice B.

[4]. $\mathbb{R}_i(\alpha)$ pode ainda ser tratada como possuindo uma correlação auto regressiva, ou seja,

$$[\mathbb{R}_i]_{jk} = \begin{cases} \alpha^{|t_{ij}-t_{ik}|} & , |t_{ij}-t_{ik}| \leq m \\ 0 & , |t_{ij}-t_{ik}| > m \end{cases}$$

onde t_{ij}, t_{ik} são as j-ésimas e k-ésimas observações no tempo para o i-ésimo indivíduo. Esta é uma estrutura de correlação para um processo m-estacionário.

[5]. Uma outra especificação, além de $\mathbb{R}_i = I_{n_i}$, é usada quando as observações no tempo são as mesmas para todos os indivíduos ($n_i = n$). Quando $\mathbb{R}(\alpha)$ é totalmente não especificada estima-se as $n(n-1)/2$ correlações. Esta estrutura de correlação pode ser chamada de não estruturada ou não especificada e pode ser estimada por:

$$\frac{\Phi}{k}\sum_{i=1}^{k} A_i^{-1/2} Q_i Q_i^T A_i^{-1/2} \qquad (15)$$

Este tipo de matriz só deve ser utilizada quando o número de observações no tempo (medidas repetidas) for pequeno (LIANG & ZEGER, 1986).

Outras estruturas de correlação podem ser consideradas, dependendo do conjunto de dados que se queira analisar. Desde que $\hat{\beta}_R$ e $\hat{V}_R$ sejam estimativas consistentes e assintoticamente normais para a escolha de $\mathbb{R}_i$, intervalos de confiança para β e outras inferências estatísticas serão assintoticamente corretas, mesmo quando $\mathbb{R}_i$ for especificada erradamente. Por outro lado, escolhendo a matriz de trabalho que seja próxima da realidade dos dados, aumenta-se a eficiência das estimativas (ROTNITZKY & JEWELL, 1990; LIANG et al., 1992; FITZMAURICE et al., 1993). A especificação de $\mathbb{R}_i(\alpha)$ pode ser expressa mais genericamente como $g(\mathbb{R}_i) = Z_i\alpha$, onde Z_i é o conjunto de covariáveis específicas do indivíduo e $g(\mathbb{R}_i)$ é alguma função ligação adequada. Alternativamente, Z_i pode representar uma matriz para a dependência do tempo (FITZMAURICE et al., 1993; ALBERT & McSHANE, 1995). Os estimadores são menos eficientes quando o tamanho amostral for pequeno ou moderado (DRUM et al., 1993). A

utilização desta forma para especificar $\mathbb{R}_i(\alpha)$ traz vantagens no sentido de aumentar as alternativas e opções nas escolhas dessas matrizes, porém pode trazer outras complicações na resolução dos modelos em face às restrições que provavelmente surgirão.

Algumas alternativas para especificações da dependência do tempo vêm sendo desenvolvidas. LIPSITZ et al. (1991) e LIANG et al. (1992) sugeriram modelar a associação por meio da razão de chance ("*odds ratio*") marginal emparelhada. Com respostas binárias, a razão de chance marginal é uma medida natural de associação e ln (γ_t) pode ser modelada como a função linear das covariáveis.

2.2.3. As EEG de Segunda Ordem (EEG2)

Recentemente, ZHAO & PRENTICE (1990) e PRENTICE & ZHAO (1991) apresentaram extensões das EEG que permitem a estimação-conjunta para a média e a covariância dos parâmetros. Para o caso especial de dados binários, PRENTICE (1988) formalizou esta extensão para estimar as equações dos parâmetros na matriz de correlação.

A extensão das EEG é dada pela seguinte expressão:

$$\sum_{i=1}^{k} \begin{pmatrix} \frac{\partial \mu_i}{\partial \beta} & 0 \\ \frac{\partial \delta_i}{\partial \beta} & \frac{\partial \delta_i}{\partial \alpha} \end{pmatrix}^T \begin{pmatrix} V_i & C_i \\ C_i^T & B_i \end{pmatrix}^{-1} \begin{pmatrix} y_i - \mu_i \\ s_i - \sigma_i \end{pmatrix} = 0 \qquad (16)$$

onde $S_{ist} = (Y_{is} - \mu_{is})(Y_{it} - \mu_{it})$, $\delta_{ist} = E(S_{ist})$, $C_i \approx cov(Y_i, S_i)$ e $B_i \approx cov(S_i)$ e $s_i = (s_{i11}, \ldots, s_{i23}, \ldots)$ sendo o vetor de covariâncias empiricas com cada $s_{ist} = 1$. As matrizes C_i e B_i são

matrizes covariância de trabalho, expressas como uma função dos dois primeiros momentos. Resultados assintoticamente normais foram obtidos, sob condições gerais de regularidade, para os estimadores-conjuntos de β e α. LIANG et al. (1992) especificaram o tempo de dependência em termos da razão de chance (*"odds ratio"*) marginal, descrevendo um conjunto equivalente de equações de estimação para estimativas-conjuntas da média e da associação dos parâmetros marginais.

A equação (16) é denominada na literatura como EEG2 porque os momentos empíricos de segunda ordem são utilizados para a estimação dos parâmetros β e α. As EEG propostas por LIANG & ZEGER (1986) são denominadas na literatura de EEG1. Embora estas nomenclaturas para as EEG estejam sendo utilizadas na literatura, existem algumas críticas feitas ao uso das mesmas (ZEGER, 1988).

Uma vantagem das EEG2 é que elas produzem estimativas mais eficientes para β e α, desde que o modelo para a média e a associação marginal sejam corretamente especificados. Entretanto, uma séria desvantagem é que $\hat{\beta}$ pode não ser consistente se for utilizada uma estrutura de correlação incorreta, mesmo quando o modelo para a média é especificado corretamente. Já as EEG1 necessitam apenas que o modelo para estimar a média esteja correto e, assim, as estimativas de β são consistentes. Além disso, dado (β, α), a especificação de (C_i, B_i) requer suposições adicionais sobre o terceiro e quarto momentos. As EEG1 e EEG2 podem ser menos eficientes quando os conglomerados não possuem o mesmo tamanho ou quando existem diferentes padrões de associação dentro do conglomerado. LIANG et al. (1992) e CAREY, ZEGER, DIGGLE (1993) recomendaram o seguinte: quando α é considerado como uma perturbação, ou seja, estimar α não é o interesse principal, e o número de conglomerados k é grande em relação ao tamanho de cada conglomerado, n_i, as EEG1 produzem estimativas eficientes dos coeficientes de regressão, mesmo quando as estimativas de associação entre os resultados são ineficientes.

Quando o número de conglomerados é pequeno e o objetivo não é tratar a correlação como parâmetro de perturbação, então as EEG2 produzem estimativas mais eficientes para os parâmetros de associação.

2.3. Métodos Clássicos de Estimação e as Equações de Estimação Generalizadas

2.3.1. O Método dos Quadrados Mínimos (QM) e as EEG

Na literatura existem poucos artigos que fazem comparação entre os métodos tradicionais de estimação e as EEG. Nesta seção serão comparadas as equações de estimação generalizadas com o método de quadrados mínimos em dois estágios através de exemplos aplicados. O método dos QM em dois estágios não requer suposições a respeito da distribuição da variável resposta e é uma extensão do modelo proposto por PARK & WOOLSON (1992). Chama-se dois estágios por ser uma técnica que possui duas etapas: no primeiro estágio obtém-se a matriz de covariâncias e um parâmetro de escala ϕ; no segundo estágio usam-se estas estimativas para se obter o estimador de β.

Supondo-se a distribuição marginal conhecida, o método dos QM em dois estágios tem propriedades assintóticas semelhantes às EEG de LIANG & ZEGER (1986). A principal diferença entre o procedimento de QM em dois estágios e as EEG é que o primeiro é uma extensão direta dos QM, enquanto as EEG são uma extensão da função de quase verossimilhança de McCULLAGH & NELDER (1989).

Antes de iniciar as comparações, PARK & WOOLSON (1992) e PARK (1994), apresentaram alguns conceitos para melhor compreensão do método dos QM em dois estágios.

Especificação de um modelo para medidas repetidas

Suponha que existem k indivíduos e t pontos no tempo para os dados coletados. Seja y_{ij} a resposta do i-ésimo indivíduo no j-ésimo ($j=1,2, \ldots, t$) tempo e:

$$\delta_{ij} = \begin{cases} 1, & se\ y_{ij}\ é\ observado \\ 0, & se\ y_{ij}\ não\ é\ observado \end{cases}$$

Considere $t_i = \sum_{j=1}^{t} \delta_{ij}$ e $n_j = \sum_{i=1}^{k} \delta_{ij}$ como sendo o número total de observações no tempo j. Seja x_{ij} um vetor de covariáveis para o i-ésimo indivíduo no tempo j e seja $x_i = (x_{i1}, \ldots, x_{it_i})^T$ uma matriz $t_i \times p$ dos valores das covariáveis para o i-ésimo indivíduo ($i = 1, 2, \ldots, k$). Os valores de t_i e n_j acima só serão aplicados quando y_{ij} e x_{ij} forem observados. Para funções V e a conhecidas, suponha que os primeiros dois momentos de y_{ij} são:

$$E(y_{ij}) = \mu_{ij} \quad e \quad Var(y_{ij}) = V(\mu_{ij})\, a(\phi). \tag{17}$$

onde $V(\mu_{ij})$ é a função da variância e ϕ é um possivel parâmetro de escala conhecido.

Observa-se que se a resposta for binária, então $V(\mu_{ij}) = \mu_{ij}(1-\mu_{ij})$ e $a(\phi) = 1$; se for uma variável resposta de Poisson, têm-se $V(\mu_{ij}) = \mu_{ij}$ e $a(\phi) = 1$. Para um resultado normal, $V(\mu_{ij}) = 1$ e $a(\phi) = \phi$.

O modelo descrito pode ser parametrizado por cada tempo ou por indivíduos, como pode ser visto a seguir. PARK (1994) trouxe em seu artigo a teoria de construção de modelos para variáveis com respostas normais e com qualquer outro tipo de resposta.

Parametrização do modelo por ponto no tempo

Seja $\eta_j^* = (\eta_{1j}, \ldots, \eta_{n_j j})^T$ um vetor $n_j \times 1$, onde $\eta_{ij} = g(\mu_{ij})$ e g é a função ligação. Seja $X_j = (x_{1j}, \ldots, x_{n_j j})^T$ uma matriz $n_j \times p$ de covariáveis no tempo j. O modelo marginal no

tempo j, descrito por PARK (1994), que permite diferentes parâmetros de regressão através do tempo é:

$$\eta_j^{\cdot} = X_j \beta_j \tag{18}$$

onde $\beta_j = (\beta_{1j}, \ldots, \beta_{pj})^T$ é um vetor $p \times 1$ de parâmetros desconhecidos a serem estimados.

O modelo em (18) assume que os efeitos de todas as covariáveis mudam no tempo. Na maioria dos casos, entretanto, nem todas as variáveis variam com o tempo. Para simplificar, distinguem-se dois tipos de variáveis: uma que não muda com o tempo (por exemplo, sexo) e a outra que varia à medida que o tempo passa. Para melhor definição do modelo anterior, levando em consideração estes tipos de variáveis, algumas considerações são necessárias:

Sejam $x_{ij}=(x_{ij1}, x_{ij2}, \ldots, x_{ijq_1}, x_{ijq_1+1}, \ldots, x_{ijq_1+q_2})$ as covariáveis do modelo tal que $q_1+q_2=p$. Agora, seja $x_{ij1}, x_{ij2}, \ldots, x_{ijq_1}$ as covariáveis que não mudam no tempo (fixas) e $x_{ijq_1+1}, \ldots, x_{ijq_1+q_2}$ as que variam com o tempo. Chamando as variáveis fixas de f_{ij} e as variáveis que variam com o tempo de w_{ij}, tem-se $F_j = (f_{1j}^T, \ldots, f_{n_j j}^T)^T$ e $W_j = (w_{ij}^T, \ldots, w_{n_j j}^T)^T$, com $j = 1, 2, \ldots, t$. Portanto X_j pode ser escrito da seguinte forma:

$$X_j = [F_j \; W_j]. \tag{19}$$

Tomando $\beta_j = (\xi, \gamma_j)^T$ pode-se reescrever o modelo (18) como segue :

$$\begin{aligned} \eta_j^{\cdot} &= X_j \beta_j \\ &= F_j W_j \begin{bmatrix} \xi \\ \gamma_j \end{bmatrix} \\ &= F_j \xi + W_j \gamma_j, \qquad j = 1, 2, \ldots, t. \end{aligned} \tag{20}$$

Agora, define-se $\eta^* = [\eta_1^*,\ldots,\eta_t^*]^T$, logo

$$\eta^* = \begin{bmatrix} \eta_1^* \\ \vdots \\ \eta_t^* \end{bmatrix} = \begin{bmatrix} F_1\xi + W_1\gamma_1 \\ \vdots \\ F_t\xi + W_t\gamma_1 \end{bmatrix} = X\beta, \tag{21}$$

sendo $X = \begin{bmatrix} F_1 & W_1 & 0 & \ldots & 0 \\ F_2 & 0 & W_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ F_t & 0 & \ldots & \ldots & W_t \end{bmatrix}$ e $\beta = [\xi,\gamma_1,\ldots,\gamma_t]^T$ onde η^* é um vetor com dimensão

$\sum_{j=1}^{t} n_j \times 1$ e β um vetor de parâmetros com dimensão $q_1 + tq_2 = p^*$.

Parametrização do Modelo por Indivíduo

Considere agora a parametrização de (18) com respeito a cada indivíduo. Seja $y_i = (y_{i1},\ldots,y_{it_i})^T$ um vetor de $t_i \times 1$ de respostas e $\eta_i = (\eta_{i1},\ldots,\eta_{it_i})^T$ um vetor $t_i \times 1$, para o i-ésimo indivíduo. Seja $A_i = diag\ \{a''(\theta_{ij})\}$ e $\mathbb{R}$ a matriz correlação definida anteriormente (2.2.2). Portanto, a matriz covariância de trabalho é dada por:

$$var\ (y_i) = \sum_i = (A_i^{1/2}\,\mathbb{R}_i A_i^{1/2})/\phi. \tag{22}$$

para $i=1,2,\ldots,k$.

Segundo PARK & WOOLSON (1992) o modelo (18) pode ser escrito equivalentemente da seguinte forma:

$$\eta_i = x_i\beta, \qquad i = 1,2,\ldots,k, \tag{23}$$

onde x_i é uma matriz $t_i \times p^*$ para o i-ésimo indivíduo dada por:

$$x_i = \begin{bmatrix} f_{i1} & w_{i1} & 0 & \cdots & 0 \\ f_{i2} & 0 & w_{i2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{it_i} & 0 & \cdots & \cdots & w_{it_i} \end{bmatrix}. \tag{24}$$

Esta parametrização em termos de cada indivíduo tem sido usada na análise de dados com medidas repetidas. Esta técnica deve ser utilizada quando os focos da análise da inferência forem os indivíduos. Este modelo inclui o modelo de LIANG & ZEGER (1986) como um caso especial.

2.3.1.1. Estimação do Modelo dos Quadrados Mínimos em Dois Estágios

A estimação em dois estágios, para o modelo por indivíduo (23) consiste em primeiro obter as estimativas de ϕ e $\mathbb{R}$ e no segundo estágio obter o estimador dos QM de β a partir dessas estimativas. É uma metodologia que torna-se similar as EEG embora seja uma extensão do método dos QM. PARK (1994) partiu da soma de quadrados dos resíduos dada por:

$$\sum_{i=1}^{k} (y_i - \mu_i)^T \sum_i^{-1} (y_i - \mu_i) = 0. \tag{25}$$

onde $\sum_i$ foi definida como em (2.2.2) e mostrou que o estimador $\hat{\beta}$ de β é solução das seguintes equações:

$$\sum_{i=1}^{k} D_i^T \sum_i^{-1} Q_i = 0. \qquad (26)$$

onde D_i, Q_i e V_i^{-1} foram definidas nas seções 2.2.2 e 2.3.1, respectivamente e que a solução da equação anterior fornece o seguinte resultado:

$$\hat{\beta} = \left(\sum_{i=1}^{k} D_i^T \sum_i^{-1} D_i \right)^{-1} \left(\sum_{i=1}^{k} D_i^T \sum_i^{-1} z_i \right), \qquad (27)$$

onde $z_i = D_i \hat{\beta}_I + Q_i$ e $\hat{\beta}_I$ é o estimador de QM para β sob a suposição de independência. A distribuição assintótica de $\hat{\beta}$ pode ser obtida do Teorema 2 de LIANG & ZEGER (1986).

A diferença entre a estimação em dois estágios proposta e o método de Liang e Zeger é na estimação de ϕ e $\mathbb{R}$. O método proposto por Liang e Zeger usa diferentes estimativas consistentes de ϕ e $\mathbb{R}$ a *cada iteração*, enquanto que o método em dois estágios usa as estimativas obtidas no primeiro passo.

A seguir compara as estimativas utilizando os QM em dois estágios e as EEG para uma variável resposta de Poisson através do modelo parametrizado por indivíduo. No item 2.3.1.2 serão apresentados os resultados obtidos através da utilização de dados reais e no item 2.3.1.3 a mesma comparação será feita usando-se dados simulados (PARK, 1994).

2.3.1.2. Exemplo

Um estudo foi conduzido com 73 crianças num período de um ano (KARIM, 1989; PARK, 1994). Em cada trimestre, a resposta de interesse foi o número de visitas ao hospital que

apresenta uma distribuição de Poisson. As variáveis estudadas foram:

$$sexo = \begin{cases} 0, \ se \ feminino \\ 1, \ se \ masculino \end{cases}$$

$$\begin{matrix} situação \ de \ fumo \\ da \ mãe \end{matrix} = \begin{cases} 0, \ se \ a \ mãe \ é \ não \ fumante \\ 1, \ se \ a \ mãe \ é \ fumante \end{cases}$$

e a idade (em meses) em que a criança iniciou os estudos. As variáveis serão denotadas como SEXO, FUMO e IDADE respectivamente, sendo cada covariável independente no tempo. Além das variáveis descritas acima, foram criadas três variáveis indicadoras:

$$Q_2 = \begin{cases} 1, \ para \ o \ 2° \ trimestre \\ 0, \ caso \ contrário \end{cases}$$

$$Q_3 = \begin{cases} 1, \ para \ o \ 3° \ trimestre \\ 0, \ caso \ contrário \end{cases}$$

$$Q_4 = \begin{cases} 1, \ para \ o \ 4° \ trimestre \\ 0, \ caso \ contrário \end{cases}$$

No modelo de regressão incluindo as covariáveis descritas acima foram utilizadas três estruturas de correlação: a permutável, modelo auto-regressivo de ordem 1 (AR-1) e não

estruturada. A sigla EEG representa a estimativa para β usando as equações de estimação generalizadas. QM indica a estimativa quando se utilizaram os quadrados mínimos em dois estágios. A *Tabela 1* apresenta as estimativas e os erros-padrões para as duas técnicas.

Tabela 1: Estimativas em dois estágios e estimativas das EEG para variável resposta de Poisson

Covariável	Estrutura de Correlação					
	Permutável		AR -1		Não Estruturada	
	EEG	MQ	EEG	MQ	EEG	MQ
Intercepto	.2460 (.3835)	.1470 (.3834)	.1493 (.4027)	.1672 (.4019)	.1945 (.3016)	.2015 (.3012)
Sexo	-.1942 (.2080)	-.1937 (.2080)	-.2420 (.2152)	-.2301 (.2157)	-.2279 (.1940)	-.2308 (.1943)
Fumo	.1688 (.2441)	.1686 (.2441)	.1754 (.2507)	.1711 (.2505)	.1934 (.2140)	.1950 (.2140)
Idade	.0030 (.0066)	.0029 (.0067)	.0065 (.0070)	.0059 (.0070)	.0044 (.0061)	.0046 (.0061)
Q2	-.4353 (.1956)	-.4353 (.1956)	-.4354 (.1956)	-.4353 (.1956)	-.4353 (.1956)	-.4353 (.1956)
Q3	-.3075 (.2028)	-.3075 (.2028)	-.3075 (.2028)	-.3075 (.2028)	-.3075 (.2028)	-.3075 (.2028)
Q4	-1.1285 (.2218)	-1.1285 (.2218)	-1.1285 (.2218)	-1.1285 (.2218)	-1.1285 (.2218)	-1.1285 (.2218)

Fonte: PARK, 1994

Essas técnicas (QM e EEG) oferecem estimativas similares dos parâmetros do modelo. As estimativas pelos QM para o INTERCEPTO e FUMO têm erros-padrões menores que as

estimativas das EEG. A variável SEXO possui erro-padrão maior nos QM para a estrutura de correlação AR-1. Já para a variável IDADE, observam-se praticamente os mesmos erros-padrões para o método QM. É interessante observar que as estimativas nos dois métodos, para as variáveis Q2, Q3 e Q4, são as mesmas, independentemente da estrutura de correlação utilizada. Em resumo, os autores observaram que as estimativas com os QM para covariáveis discretas, independentes no tempo, tendem a ter erros-padrões menores que as estimativas usando as EEG, as quais obtêm maiores erros-padrões para covariáveis contínuas e independentes do tempo. Para as covariáveis discretas dependentes do tempo, os dois procedimentos oferecem estimativas semelhantes. Estudos de simulação feitos por PARK (1994) confirmaram estes resultados (*Tabela 2*).

2.3.1.3. Estudos de Simulações

Foram realizados estudos de simulação pelo método de Monte Carlo com o objetivo de comparar propriedades dos métodos dos QM e das EEG para amostras pequenas. As EEG têm sido aplicadas em medidas repetidas, mas poucos estudos têm utilizado esta técnica com variáveis respostas de Poisson (PARK, 1994). As propriedades para amostras pequenas usando as EEG foram pesquisadas para variáveis Gama correlacionadas (PARK, 1988) e nas variáveis de respostas binárias correlacionadas (PRENTICE, 1988; ZHAO & PRENTICE, 1990; LIPSITZ et al., 1991; LIPSITZ et al., 1994).

Os dados gerados, consideram apenas dois grupos (Grupo A e B), e possuem respostas correlacionadas com distribuição de Poisson para $t = 2$, onde t é a quantidade de medidas repetidas. O modelo marginal definido por PARK (1994) é $\log(\mu_{ij}) = \beta_1 x_i + \beta_2 \frac{j}{2}$, onde $x_i = 0$ para o Grupo A e $x_i = 1$ para o Grupo B, e β_2 representa o efeito do tempo para $j = 1, 2$. O

vetor de parâmetros é $(\beta_1, \beta_2) = (1, 0.5)$, $\phi = 1$ e $\mathbb{R}_i$ é o modelo AR-1 com $\alpha = 0.1; 0.3; 0.5$ *e* 0.7. Usando estas especificações, foram realizadas 500 simulações. Os valores referentes aos vícios e aos erros quadráticos médios estão na *Tabela 2*.

Tabela 2: Vício relativo (%) e erro quadrático médios das estimativas de β para variável resposta de Poisson. O modelo marginal é $\log(\mu_{ij}) = \beta_1 x_i + \beta_2 (j/2)$, onde x_i é 0 para o Grupo A e 1 para o Grupo B. Aqui, $\beta = (1, 0.5)^T$ e $\mathbb{R}$ é AR-1.

		Amostra = 30				Amostra = 50			
	Parâmetro	Viés		EQM		Viés		EQM	
		QM	EEG	QM	EEG	QM	EEG	QM	EEG
α=0.1	β_1	-.6084	-.6252	.0267	.0268	-.5677	-.5705	.0172	.0172
	β_2	-.0380	.0150	.0283	.0284	.1358	.1444	.0172	.0173
α=0.3	β_1	-.1441	-.1580	.0271	.0271	-.2799	-.2977	.0170	.0170
	β_2	1.2783	1.2319	.0270	.0270	.3798	.3283	.1697	.1698
α=0.5	β_1	-.4077	-.4429	.0292	.0291	-.3490	-.3645	.0182	.0182
	β_2	-1.2148	-1.100	.0248	.0245	-.1966	-.1465	.0148	.0148
α=0.7	β_1	-.3908	-.4569	.0281	.0279	-.0373	-.0505	.0119	.0118
	β_2	.8275	.5966	.0184	.0179	-.0047	-.0041	.0119	.0118

Fonte: PARK, 1994

Analisando os vícios das estimativas dos QM para β_1, percebem-se valores absolutos menores comparados com os obtidos pelas EEG. Já, para β_2, os vícios são maiores quando as estimativas são obtidas pelo método das QM. Quando o tamanho da amostra é 50 percebe-se que praticamente não há diferença nos vícios para β_1 e β_2 entre os dois métodos. Os erros quadráticos médios (EQM) são praticamente idênticos para os dois métodos e para os parâmetros β_1 e β_2.

A *Tabela 3* sumariza a probabilidade de vezes que converge os intervalos de confiança

de 95% para β.

Tabela 3: Probabilidade de vezes de convergência para intervalos de confiança de 95% para β com variável resposta de Poisson. O modelo marginal é $\log(\mu_{ij}) = \beta_1 x_i + \beta_2 \frac{j}{2}$, *onde* x_i *é 0 para o Grupo A e 1 para o Grupo B.* $\beta = (1, 0.5)^T$ *e* $\mathbb{R}$ *é AR-1.*

	Parâmetro	Tamanho da Amostra = 30		Tamanho da Amostra = 50	
		QM	EEG	QM	EEG
α=0.1	β_1	94.0	94.0	94.8	94.8
	β_2	94.0	93.8	93.8	93.8
α=0.3	β_1	95.2	95.2	93.0	93.0
	β_2	95.2	95.0	94.2	94.0
α=0.5	β_1	93.6	93.6	93.0	93.0
	β_2	93.8	94.2	94.2	94.0
α=0.7	β_1	92.2	92.2	94.0	94.0
	β_2	95.6	95.6	93.6	93.6

Fonte: PARK 1994

Quando o tamanho da amostra é 30 e $\alpha = 0.3$, ambos estimadores demonstram probabilidade de vezes de convergência maior que o nivel nominal (95%). Porém, quando o tamanho total da amostra é 50, estes estimadores tendem a ter probabilidade de vezes de convergência um pouco menor que o nível nominal. Em geral, os dois estimadores produzem semelhantes probabilidades de vezes de convergência.

PARK (1994) argumentou que, para variáveis de Poisson, nos estudos de simulações pelo método de Monte Carlo, a utilização dos QM é preferível aos das EEG quando as covariáveis são discretas. E ainda: os estimadores pelo método dos QM em dois estágios possuem propriedades assintóticas semelhantes aos estimadores usando as EEG. O artigo traz também um

exemplo quando a variável resposta tem uma distribuição normal.

2.3.2. Os Quadrados Mínimos Ponderados (QMP) e as EEG

O método dos quadrados mínimos ponderados (QMP) desenvolvido por KOCH et al. (1977) foi uma das primeiras técnicas apresentada para dados longitudinais com respostas categorizadas. Uma das vantagens deste método é a flexibilidade que ele oferece para modelar as proporções marginais, logitos marginais, média dos escores e logitos cumulativos. Esta metodologia, contudo, pode ser ineficiente quando as freqüências dentro das categorias são pequenas e a mesma não pode ser usada com variáveis contínuas.

MILLER et al. (1993) compararam o método dos QMP com as EEG usando dados descritos por KOCH et al. (1989) de uma triagem clínica controlada e aleatorizada para um novo tratamento de doença respiratória.

O estudo foi realizado em 111 pacientes que foram aleatoriamente designados para um dos dois tratamentos (ativo, placebo). De cada quatro visitas durante o período em estudo, a resposta referente a situação respiratória de cada paciente foi classificada segundo uma escala ordinal de cinco pontos, sendo: 0 = terrível; 1 = ruim; 2 = regular; 3 = bom; e 4 = excelente. Para ilustração, os dados foram analisados apenas em três pontos da escala ordinal, ficando assim: (0-1) = ruim; (2-3) = bom e 4 = excelente. Os dados coletados estão na *Tabela 4*.

Tabela 4: Respostas dos 111 pacientes nas quatro visitas

Visita				Número de pacientes		Visita				Número de pacientes	
1	2	3	4	Ativo	Placebo	1	2	3	4	Ativo	Placebo
r	r	r	r	1	6	b	b	e	e	1	2
r	r	b	r	1	0	b	e	b	b	0	1
r	r	b	b	0	2	b	e	b	e	2	1
r	b	r	r	1	0	b	e	e	b	3	0
r	b	b	b	0	2	b	e	e	e	7	1
r	b	e	e	1	0	e	r	r	r	0	1
b	r	r	r	0	4	e	r	e	b	1	0
b	r	r	b	0	1	e	b	r	b	0	1
b	r	b	b	1	2	e	b	b	p	1	1
b	b	r	r	1	2	e	b	b	e	1	1
b	b	r	b	2	2	e	b	e	b	0	2
b	b	b	r	4	1	e	b	e	e	0	2
b	b	b	b	8	1	e	e	b	b	2	0
b	b	b	e	2	2	e	e	b	e	2	0
b	b	e	b	1	0	e	e	e	e	8	7

Fonte: MILLER et al., 1993

r = ruim ; b = bom; e = excelente

Para acomodar a natureza da variável resposta, a análise foi conduzida usando-se a função de ligação logito cumulativa. Esta transformação é a mais utilizada para dados politômicos e é baseada na soma dos logitos parciais das probabilidades multinomiais. Esta transformação pode ser apresentada como:

$$L_{itg} = logit\ [Pr(Y_{it} \leq g)] = \log\left[\frac{\pi_{it1} + \pi_{it2} + \ldots + \pi_{itg}}{\pi_{itg+1} + \ldots + \pi_{itr}}\right] = x_{itg}^{T}\beta, \qquad \text{(Modelo 1)}$$

onde $i = 1,2,...,k$ indivíduos, $g = 1,2,...,r$ categorias de cada $t = 1,2,...,d$ ocasiões.

Outras funções de ligação apropriadas para dados politômicos são discutidas por McCULLAGH & NELDER (1989) e AGRESTI (1989). Três procedimentos de estimação foram utilizados: QMP, EEG assumindo independência entre as medidas repetidas (EEG-IND) e EEG usando a suposição de correlação "saturada" (EEG-SAT). O procedimento EEG-SAT recebe este nome por se tratar do modelo saturado assumindo as mesmas correlações $[(r-1)\ d(d-1)] / 2$ dentro de S subgrupos. Cada subgrupo s ($s=1, 2, ..., S$) corresponde aos grupos formados com pessoas do mesmo perfil. Pode-se mostrar (MILLER, DAVIS, LANDIS, 1993) que o Modelo 1 pode ser reescrito como:

$$L_{stg} = \theta_{tg} + \tau_t, \qquad \text{(Modelo 2)}$$

onde L representa o g-ésimo logito cumulativo para o s-ésimo subgrupo na visita t. θ_{tg} são chamados de pontos de cortes de uma determinada visita, ou seja, θ_{tg} é o efeito na visita t de se ter a resposta (categoria) g, e τ_t é o efeito do tratamento nesta visita. Os resultados para este modelo são apresentados na *Tabela 5*, mostrando o efeito do tratamento em cada visita.

Tabela 5: Resultados assumindo o modelo de odds proporcional nas visitas - (Modelo2)

Parâmetro	EEG-IND		EEG-SAT		QMP	
	Estimativa	Estatística de Wald	Estimativa	Estatística de Wald	Estimativa	Estatística de Wald.
Visita 1:						
Ponto de corte (θ_{11})	-1.95	48.47	-1.95	47.20	-1.91	43.60
Ponto de corte (θ_{12})	.83	15.61	.85	17.63	.84	17.49
Tratamento (τ_1)	-.22	1.45	-.24	1.66	-.22	1.36
Visita 2:						
Ponto de corte (θ_{21})	-1.70	42.69	-1.72	43.21	-1.68	41.49
Ponto de corte (θ_{22})	.73	11.37	.71	11.29	.72	11.63
Tratamento (τ_2)	-.74	14.40	-.74	14.77	-.72	13.85
Visita 3:						
Ponto de corte (θ_{31})	-1.48	39.24	-1.45	36.94	-1.41	33.93
Ponto de corte (θ_{32})	.61	8.67	.61	9.09	.60	9.01
Tratamento (τ_3)	-.53	8.30	-.54	8.62	-.52	7.95
Visita 4:						
Ponto de corte (θ_{41})	-1.33	31.83	-1.35	34.08	-1.32	32.98
Ponto de corte (θ_{42})	.58	8.40	.59	8.86	.59	8.92
Tratamento (τ_4)	-.33	3.44	-.32	3.30	-.31	3.00
Falta de ajuste (gl=4)		2.33		2.36		2.31
Teste de hipótese (gl =3)						
$\theta_{11} = \theta_{21} = \theta_{31} = \theta_{41}$		4.69		4.65		4.48
$\theta_{12} = \theta_{22} = \theta_{32} = \theta_{42}$		1.54		1.79		1.73

Fonte: MILLER et al., 1993

Os autores observaram uma pequena variabilidade nos resultados obtidos para os três procedimentos de estimação. Os contrastes apresentados na *Tabela 5* foram utilizados como auxílio à obtenção dos pontos de corte dos parâmetros para cada visita. Os efeitos dos tratamentos utilizando os três métodos foram: $\hat{\tau}_1 = -0.22$, $\hat{\tau}_2 = -0.72$, $\hat{\tau}_3 = -0.52$ e $\hat{\tau}_4 = -0.31$ para os QMP; $\hat{\tau}_1 = -0.24$, $\hat{\tau}_2 = -0.74$, $\hat{\tau}_3 = -0.54$ e $\hat{\tau}_4 = -0.32$ para as EEG-SAT; e

$\hat{\tau}_1 = -0.22$, $\hat{\tau}_2 = -0.74$, $\hat{\tau}_3 = -0.53$ e $\hat{\tau}_4 = -0.33$ para as EEG-IND. De uma maneira geral, observou-se que os três métodos indicaram a não eficiência do tratamento na visita 1 e a existência de eficiência significativa do tratamento nas outras visitas, decrescendo a magnitude da estimativa da visita 2 para a visita 4.

O efeito do tratamento através das visitas foi testado utilizando-se diferentes estatísticas e suposições para as covariâncias entre as visitas. A *Tabela 6* contem o teste de Wald e testes escores que podem ser considerados apropriados para cada procedimento de estimação utilizado (ROTNITZKY & JEWELL, 1990; CARR & CHI, 1992). Os testes apresentados na *Tabela 6* verificam a mesma hipótese nula de igualdade do efeito do tratamento através das visitas, porém existem algumas diferenças nas magnitudes dessas estatísticas.

Testando as hipóteses nulas sob a suposição que $\sum_i = \sum_s$ ($\sum_i$ = matriz de covariância de Y), notam-se resultados similares para os QMP e o teste de escore generalizado usando as EEG-SAT. Em contraste, nas estatísticas calculadas sem levar em consideração esta suposição (isto é, $\sum_i \neq \sum_s$ para o mesmo i no perfil *s*), a amplitude da magnitude varia de 6.57 a 12.47, com as estatisticas calculadas pelo escore generalizado, sendo menores que as estatisticas de Wald. A diferença na magnitude entre os testes escores e Wald pode ser decorrente do fato que os parâmetros desconhecidos no teste escore são calculados sob as hipóteses nulas, as passo que, para os testes de Wald, estes parâmetros são calculados sob as hipóteses alternativas (ROTNITZKY & JAWELL, 1990).

Tabela 6: Testes para igualdade dos efeitos dos tratamentos

Técnica de estimação	Estatística do teste [a]	Suposição da variância	Estatística calculada	Nível descritivo
WLS	Wald	$\Sigma_1 = \Sigma_2$	8.96	.030
EEG-IND	G-Wald	$\Sigma_1 = \Sigma_2$	12.47	.006
	G-Score	$\Sigma_1 = \Sigma_2$	12.14	.007
EEG-SAT	G-Wald	$\Sigma_1 = \Sigma_2$	11.90	.008
	G-Wald	$\Sigma_1 = \Sigma_2$	10.91	.012
	W-Wald	$\Sigma_1 = \Sigma_2$	7.41[b]	.076[c]
	G-Score	$\Sigma_1 = \Sigma_2$	8.95	.030
	G-Score	$\Sigma_1 = \Sigma_2$	8.60	.035
	W-Score	$\Sigma_1 = \Sigma_2$	6.57[b]	.037[c]

Fonte: MILLER et al., 1993
[a] G indica estatística generalizada; W indica estatística de trabalho [ROTNITZKY & JAWELL,1990]
[b] Obtido usando correção de 1ª ordem de ROTNITZKY & JAWELL (1990)
[c] Obtido usando a correção de 1ª e 2ª ordem de ROTNITZKY & JAWELL (1990)

2.3.2.1. Outros Estudos de Simulações

MILLER et al. (1993) geraram dados de indivíduos os quais foram alocados em dois tratamentos (tratamentos A e B) em três tempos diferentes. A variável resposta foi medida em 3 pontos de uma escala ordinal. O modelo logito cumulativo, utilizado para as esperanças marginais, foi uma simplificação do Modelo 2. Assumiu-se um modelo de *odds* proporcional nos tempos e especificaram-se os parâmetros principais do modelo como: $\theta_1 = -1.68$, $\theta_2 = 0.64$, $\tau_1 = -0.75$, $\tau_2 = -0.50$ e $\tau_3 = -0.25$, onde o tratamento A foi codificado por 1 e o tratamento B por -1. Na *Tabela 7* encontram-se as esperanças marginais produzidas por este modelo. Os dados foram gerados sob diferentes suposições de matrizes de correlação com a finalidade de investigar o desempenho dos três métodos de estimação. As suposições de correlações geradas entre as visitas, para os dados, foram: "fraca" (I), "moderada" (II) e "não

correlacionada" (III).

Tabela 7: Proporção marginal para as simulações

Tratamento	Resposta Visita 1			Resposta Visita 2			Resposta Visita 3		
	1	2	3	1	2	3	1	2	3
A	.08	.39	.53	.10	.43	.47	.13	.47	.40
B	.28	.52	.20	.24	.52	.24	.19	.52	.29

Fonte: MILLER et al., 1993

Foram selecionadas para a simulação duas amostras: uma de 35 e outra de 70 observações por tratamento. Foram executadas 1000 replicações em cada uma das seis simulações (duas amostras diferentes × três suposições de correlação), e os parâmetros foram estimados para cada replicação em três métodos diferentes (QMP, EEG-IND e EEG-SAT). Em cada replicação, foram executados os mesmos testes da *Tabela 6,* para igualdade do efeito dos tratamentos através das visitas.

As estimativas resultantes das simulações realizadas pelo método das EEG-IND convergiram com sucesso em todas as replicações para cada combinação do tamanho de amostra e para diferentes modelos de correlação. Por outro lado, para os métodos QMP e EEG-SAT foram encontrados alguns problemas de estimação quando o tamanho da amostra era pequeno e/ou a correlação entre visitas era um pouco forte (correlação II). Pelos dois métodos (EEG-SAT e QMP) os resultados não foram obtidos quando a matriz de covariância para os dois grupos de tratamento foi próxima da singularidade ou quando esta matriz tornou-se singular com futuras iterações. Para QMP este problema aconteceu em menos de 2% de todas as replicações. Entretanto, para a correlação chamada II, as EEG1-SAT convergiram em apenas 90% das replicações para 35 observações por tratamento e 98% das replicações para 70 observações por

tratamento. Problemas similares têm sido discutidos por LIPSITZ et al. (1991) em suas simulações envolvendo dados com respostas binárias.

Foram calculadas três medidas: a média das estimativas do efeito do tratamento; o viés relativo destes e a probabilidade de convergência para efeitos do tratamento com intervalo de confiança de 95%. Na probabilidade de convergência, observaram-se tendências importantes na amostra de 35 observações por tratamento e em todas as categorias de correlação. Os vícios relativos dos QMP foram geralmente maiores que os observados para os outros dois métodos de estimação. As probabilidades de convergências dos QMP e EEG-SAT foram relativamente fracas em comparação às obtidas pelo método EEG-IND. Para a amostra de 70 observações por tratamento, os três métodos apresentaram similaridade, exceto para o viés positivo associado com os QMP

Na *Tabela 8* é apresentado o percentual de vezes que cada um desses métodos rejeitou a hipótese de igualdade do efeito dos tratamentos nas visitas, com dois graus de liberdade. Duas tendências foram observadas na tabela abaixo: *i)* houve um aumento do poder do teste com o aumento do tamanho da amostra e *ii)* um aumento do poder do teste com o aumento da magnitude da correlação entre as medidas repetidas. Também foi observado que os testes de Wald calculados sob a suposição que $\sum_i = \sum_s$ têm um maior poder dentre os testes considerados.

Tabela 8: Percentagem de vezes que as hipóteses de igualdade dos efeitos dos tratamentos foi rejeitada

Técnica de Estimação	Estatística do teste[a]	Suposição da Variância	35 obs. por tratamento			70 obs. por tratamento		
			Não Corr.	Corr. I	Corr. II	Não Corr.	Corr. I	Corr. II
WLS	Wald	$\Sigma_i = \Sigma_s$	26.9	41.4	50.3	48.3	60.3	78.4
GEE1-IND	G-Wald	$\Sigma_i = \Sigma_s$	26.4	35.0	41.8	47.9	59.0	73.8
	G-Score	$\Sigma_i = \Sigma_s$	19.1	22.7	25.8	37.7	44.8	52.8
GEE1-SAT	G-Wald	$\Sigma_i = \Sigma_s$	31.5	43.2	52.4	51.0	61.2	78.6
	G-Wald	$\Sigma_i = \Sigma_s$	30.8	38.4	44.5	50.2	59.9	76.9
	W-Wald[b]	$\Sigma_i = \Sigma_s$	30.3	37.1	43.8	50.4	59.7	74.7
	G-Score	$\Sigma_i = \Sigma_s$	17.9	24.3	35.2	38.7	49.1	69.1
	G-Score	$\Sigma_i = \Sigma_s$	17.2	19.9	30.9	37.4	43.9	67.7
	W-Score[b]	$\Sigma_i = \Sigma_s$	15.6	17.3	28.8	37.3	41.8	65.6

Fonte: MILLER et al., 1993

[a] G indica estatística generalizada; W indica estatística de "trabalho" (ROTNITZKY & JEWEL, 1990)
[b] Obtidas usando correções de 1ª e 2ª ordem de ROTNITZKY & JEWELL (1990)

Capítulo III

Aplicações Práticas

3.1. Programas Computacionais

Por ser uma técnica de estimação bastante recente, as EEG não possuíam até 1996 um programa computacional finalizado e comercializado. Alguns programas foram desenvolvidos particularmente por pesquisadores envolvidos com a teoria das EEG. Estes programas, ainda em fase de acabamento ou expansão, estão disponíveis a partir de contatos com seus autores e têm sido usados em artigos recém-publicados sobre o assunto. Dos cinco programas existentes, dois deles foram utilizados para analisar os dados apresentados a seguir. Os outros três programas foram estudados sem que os conjuntos de dados fossem analisados por eles e estão descritos na seção 3.1.3. Para a análise dos dados deste trabalho foram utilizados os programas RMGEE

(Repeated Measures using Generalized Estimating Equations), de Davis, e o programa em SAS, elaborado por Karim e Zeger (1ª versão, 1988 e a versão 2.03, 1993).

3.1.1. RMGEE - Descrições Gerais

O programa elaborado por DAVIS (1993) em linguagem Fortran 77 é aplicável para análise de dados de medidas repetidas com respostas categorizadas e continuas usando as EEG. Este programa pode ser utilizado quando as medidas são obtidas em vários tempos para cada indivíduo e também quando a unidade amostral básica é um grupo ou conglomerado de indivíduos com a resposta de interesse obtida em cada indivíduo dentro do conglomerado. Os resultados produzidos pelo RMGEE incluem os coeficientes da regressão, as estimativas de suas variâncias e covariâncias. É um programa construído em linguagem Fortran 77 e pode ser usado sem modificações em diversos microcomputadores, estações de trabalho e computadores de grande porte.

O RMGEE consiste de um programa principal de 21 subprogramas. Primeiramente, três sub-rotinas são chamadas para determinar o tipo de entrada e saída desejada dos dados, ler as opções de análise e checar os erros de inconsistências nos parâmetros de entrada. As opções de análises são gravadas para serem exibidas na tela ou em um arquivo de saída determinado.

O programa principal, então, chama cinco sub-rotinas para executar os procedimentos computacionais requeridos. Primeiro, estimativas iniciais dos parâmetros são calculadas usando quadrados mínimos ordinais (ignorando a dependência entre as observações repetidas). Logo após, vem o procedimento para estimar a matriz de correlação de trabalho, seguido da atualização dos vetores de parâmetros estimados, que é executada até a convergência ser obtida. Finalmente,

o estimador robusto da matriz de covariância do vetor de parâmetros estimados é calculado. Em cada uma dessas cinco sub-rotinas, os dados são lidos e as matrizes necessárias são acumuladas. Se solicitada, a hipótese relativa a um conjunto de parâmetros é então testada. Os subprogramas restantes manipulam operações de várias matrizes (inversão, multiplicação de duas matrizes, manipulação de vetores e matrizes, etc.) e calculam as probabilidades do qui-quadrado para o teste de Wald dos subconjuntos dos parâmetros.

Os métodos computacionais no RMGEE são os mesmos utilizados no programa do SAS, criado pelo Karim e Zeger, e pela implementação C para o sistema S, criada por Carey. O programa do SAS será descrito na seção seguinte. Todos os três programas apresentam resultados idênticos quando é escolhida a mesma função ligação, mesma função variância e a mesma matriz correlação de trabalho. A principal diferença é que o RMGEE não permite combinações arbitrárias da função ligação e da variância. Ele pode ser utilizado em variáveis respostas com distribuição Normal, Poisson ou Binomial, e as opções para a matriz correlação de trabalho podem ser: independente, permutável e não estruturada. A implementação C permite a ligação probito. Os outros dois programas também implementam matrizes de trabalho, além da identidade como a permutável (*exchangeable*) e a não estruturada.

3.1.1.1. Estrutura do Arquivo de Dados

O arquivo de entrada dos dados pode estar na forma texto-padrão em ASCII com números no formato de processador de texto. Este arquivo deve conter uma linha para cada medida no tempo por individuo avaliado, isto é, se os dados no tempo t são obtidos para cada um dos n indivíduos, o arquivo de entrada terá que conter $n\,t$ linhas. O programa aceita no máximo 5.000

indivíduos e 30 medidas repetidas por indivíduo; logo, o arquivo de dados pode conter no máximo 150.000 linhas. A estrutura do arquivo de entrada é mostrada na *Tabela 1* abaixo. O número das observações pode variar de indivíduo para indivíduo.

Tabela 1. Estrutura do arquivo de entrada dos dados

Número da linha	Indivíduo	Variáveis
1	1	Dado da 1ª obs. do indiv. 1
2	1	Dado da 2ª obs. do indiv. 1
⋮	⋮	
k	1	Última obs. do indiv. 1
k+1	2	Dado da 1ª obs. do indiv. 2
⋮	⋮	

Fonte: DAVIS, 1994

O programa oferece um pequeno manual mostrando algumas diretrizes dos procedimentos necessários para utilizá-lo, e também mostra alguns exemplos com seus arquivos de saída.

3.1.2. GEE: A Macro do SAS

Este programa foi elaborado por Karim e Zeger em 1988 (1ª versão). É uma macro do SAS para analisar dados longitudinais através das EEG, modelando estes dados para uma classe geral de variáveis respostas e incluindo respostas Gaussianas, Poisson, Binária e Gama. O programa usa um procedimento iterativo para estimar os coeficientes de regressão, tratando a correlação entre as observações, no mesmo indivíduo, como uma perturbação. O arquivo de saída do GEE inclui coeficientes de regressão, estimativa robusta da variância e a estatística z. São

fornecidas também opções de tipos de funções de ligação, tais como identidade, logarítmica, logito e recíproca.

Para aumentar a eficiência das estimativas, o usuário pode especificar a estrutura da matriz correlação de trabalho. Esta matriz se refere à correlação entre as observações dentro do conglomerado. As opções oferecidas são: a matriz de correlação identidade, estacionária m-dependente, não estacionária m-dependente, permutável, modelo auto-regressivo de ordem 1 (AR-1) e não especificada.

3.1.2.1. Estrutura do Arquivo de Dados

Os dados de entrada devem estar num arquivo do SAS contendo a variável resposta e covariáveis. Todas as informações correspondendo à variável resposta devem constituir um único registro no conjunto de dados. Registros correspondentes ao conglomerado devem ser colocados juntos. Se é desejado que o modelo tenha intercepto, então o conjunto de dados deve conter uma coluna com a covariável de valor 1 ou ser especificado no programa.

Esta versão 1 do GEE não pode trabalhar com intervalos de tempo diferentes, a menos que se assumam erros independentes ou correlação permutável. Esta macro não processa observações com dados faltantes. Se existirem alguns valores faltantes, devem ser removidos do conjunto de observações. Note que, removendo a observação faltante, pode gerar intervalos de tempo diferentes. Portanto, como esta macro não pode manusear conglomerados de tamanhos diferentes, então uma maneira de solucionar este problema seria desconsiderar todas as observações do indivíduo após o valor faltante ter sido encontrado ou fazer algum tipo de imputação. Outra maneira é utilizar a estrutura de correlação independente ou permutável.

A variável resposta (YVAR) pode ser categorizada ou contínua. Se alguma modificação da variável resposta é requerida, esta alteração deverá ser feita dentro do passo DATA antes de chamar a macro.

Alguns exemplos são executados e exibidos nos arquivos de saída. As versões mais recentes podem ser conseguidas via Internet* e oferecem os valores de p e também a razão de chance (*"odds ratio"*) e seus respectivos intervalos de confiança quando a variável resposta é binária.

3.1.3. Outros Programas

Além desses dois programas, existem alguns outros. Um deles, Qaqish, é um programa para ajustar regressão de dados multivariados binários com estrutura de conglomerado. O mesmo traz apenas a função ligação logito. Este programa foi desenvolvido por David Lean, em 1994, para um projeto de Conferência Populacional pela Universidade da Carolina do Norte e é baseado na tese de doutorado do Dr. Bahjat Qaqish. É um programa em linguagem Fortran compilado para rodar em sistemas de operação Microsoft Windows em microcomputadores.

Existe também um programa em linguagem Pascal elaborado pelo Prof. B. Qaqish (1989, 1990, 1991) para as EEG estendidas (EEG2), com dados binários correlacionados. Este programa ajusta modelos de regressão em dados binários multivariados que admitem mais de uma medida em cada conglomerado e oferece diferentes regressões para cada medida e para a dependência entre e dentro das observações. Permite também escolha entre as EEG1 e EEG2. A função de

* Endereço Eletrônico : www.statlab.uni-heidelberg.de/statlib/GEE/GEE1/

ligação utilizada é apenas a logito. O programa em linguagem S é uma outra alternativa para análise de dados binários longitudinais e que pode ser conseguido via Internet**.

A próxima seção menciona vários tipos de métodos para analisar dados com medidas repetidas e por fim, apresenta a análise de dois conjuntos de dados reais através das EEG.

3.2. Exemplos Ilustrativos

Artigos recentes com medidas repetidas, oriundos de pesquisas na área de saúde, têm apresentado as EEG como uma nova alternativa para analisar estes tipos de dados que, na sua maioria, são provenientes de amostras complexas (KEMPTHORNE & KOCH, 1983; ZEGER, LIANG, SELF, 1985; WARE, LIPSITZ, SPEIZER, 1988; WEI & STRAM, 1988; ZEGER, LIANG, ALBERT, 1988; LEWIS, 1993; DAVIS, 1994).

Uma metodologia para dados multivariados e respostas categorizadas é baseada na estimação de uma medida de razão que, no caso, foi a densidade de incidência*** de doenças respiratórias (DR), através de modelos log-lineares. Este método requer suposições mínimas a respeito da distribuição dos dados. LAVANGE et al. (1994) apresentaram a metodologia com uma aplicação a dados de um estudo em crianças com DR durante o primeiro ano de vida. Uma questão de interesse era verificar se as crianças com exposição passiva ao cigarro tendiam a ter maior razão de DR, em média, que aquelas crianças não expostas, levando-se em consideração a idade da criança e a estação do ano. Um modelo log-linear foi ajustado para as razões estimadas

**Endereço Eletrônico: statlib@lib.stat.cmu.edu ou netlib@research.att.com

***Densidade de incidência → é calculada como a razão do número de eventos pelo tempo em risco do indivíduo contrair a doença.

com a finalidade de testar o efeito da significância das covariáveis. Para comparar técnicas e verificar as limitações do método da razão com respeito ao número de covariáveis, foram utilizados métodos de regressão logística e Poisson, ajustados via método das EEG. A regressão logística foi utilizada quando a variável resposta representou a probabilidade de contrair DR, e regressão de Poisson quando a variável resposta foi o número de vezes qua a criança apresentou DR no primeiro ano de vida. Aplicadas as várias estratégias e feitas as análises, observou-se que o modelo de regressão logística, utilizando a metodologia das EEG com matriz correlação de trabalho independente, obteve resultados semelhantes a técnicas de estimação da razão.

Além de dados na área epidemiológica, as EEG têm sido aplicadas em dados de neurologia (ALBERT & McSHANE, 1995), análise de séries de tempo discretas (ZEGER, 1988), dados toxicológicos (LEFKOPOULOU, MOORE, RYAN, 1989), dados de doença peridontal (PACK, COXHEAD, McDONALD, 1990) e outros.

ALBERT & McSHANE (1995) utilizaram as EEG para análise de dados binários espacialmente correlacionados, quando existiu uma grande quantidade de observações correlacionadas espacialmente, em um número moderado de indivíduos. As EEG permitem tratar a correlação espacial como perturbação. A metodologia é ilustrada com dados de neuroimagem coletados no Instituto Nacional de Doenças Neurológicas e Enfarto (NINDS). O artigo também apresenta, graficamente, uma comparação das curvas estimadas pelos métodos das EEG e dos QMP e fala da importância de utilizar uma estrutura de correlação correta para que haja redução no vício das estimativas. Estudos de simulações foram realizados, mostrando a importância de se modelar a média marginal em estruturas de correlação espacial em dados com um grande número de observações espacialmente correlacionadas, como é o caso de estudos encontrados na neuroimagem. Em se tratando de doenças infectocontagiosas, FIELDING et al. (1995) analisaram a transmissão do HIV em heterossexuais pelas EEG.

Além da utilização das EEG1 para dados de saúde, as EEG2 começam a ser utilizadas em dados desta área. PODGOR & HELLER (1996) mediram o grau de associação de lentes opacas entre e dentro dos olhos dos indivíduos, utilizando as EEG de segunda ordem (EEG2).

3.2.1. Dados Utilizados no Trabalho

Dois conjuntos de dados do Hospital de Clínicas (HC) da Unicamp foram utilizados neste trabalho para ilustrar a técnica das EEG, sendo que, em um deles, são apresentadas medidas repetidas onde o tempo não influenciou a variável resposta e no outro, o tempo foi fator relevante na observação da resposta.

Primeiro Conjunto de Dados

Trata-se de medidas nutricionais coletadas em 55 pacientes operados no ano de 1987 no Hospital de Clínicas (HC) da Unicamp. Várias medidas foram feitas por três observadores (A, B e C) de maneira consecutiva nos pacientes, ou seja, cada observador mediu três vezes uma mesma variável no indivíduo. As medidas coletadas foram: idade em anos (IDADE), peso em gr. (PESO), altura em cm (ALTURA), circunferência do braço em cm (CB) e prega cutânea triciptal em mm (PCT). Estas medidas foram realizadas logo após o desjejum do paciente, para não haver alteração nas medições. Os pacientes não poderiam ser portadores de edemas no braço, cicatrizes ou possuir deficiência física. Realizou-se a medição no braço, considerado dominante. A PCT foi medida com um aparelho apropriado, chamado paquímetro, e obtida em milímetros; a CB com fita métrica tradicional e obtida em cm; o PESO com uma balança antropométrica e obtido em gramas; e a ALTURA com a mesma balança, obtida em cm (WAITZBERG, 1995; BOIN et al., 1988). A

partir dessas medidas, calcularam-se as seguintes variáveis de interesse:

circunferência muscular do braço (CMB)

$$CMB = CB - [\pi \ PCT(mm)];$$

índice de massa corporal (IMC)

$$IMC = PESO/(ALTURA)^2$$

superfície corpórea (SC)

$$SC=[(PESO)^{0.425} (ALTURA)^{0.725} 71.84]/10.000$$

A título de ilustração, no *Quadro 1* segue a estrutura deste conjunto de dados para a variável IMC.

Paciente	OBSERVADOR								
	A			B			C		
	medida 1	medida 2	medida 3	medida 1	medida 2	medida 3	medida 1	medida 2	medida 3
01	33.42	33.42	33.42	34.05	34.33	34.42	32.94	32.98	33.38
02	21.47	20.95	21.47	21.76	21.76	21.80	21.75	21.75	21.47
03	20.95	20.95	20.95	21.13	21.13	21.13	21.10	21.10	21.10
04	17.89	17.68	17.89	18.21	18.21	18.21	1821	18.21	18.43
05	33.83	33.79	33.78	31.75	32.12	32.03	33.43	33.83	33.75
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
55	25.54	25.54	25.54	26.14	26.18	26.18	26.09	26.09	26.09

Quadro 1. Conjunto de dados de 55 pacientes do HC da Unicamp para a variável IMC

A medida usada na análise foi a média de cada observador. Assim, os conglomerados que inicialmente tinham tamanho nove passaram a ter tamanho três (*Quadro* 2). Os dados analisados apresentaram a seguinte estrutura:

Paciente	OBSERVADOR		
	A	B	C
01	33.42	34.27	33.10
02	21.47	21.77	21.66
03	20.95	21.13	21.10
04	17.82	18.21	18.21
05	33.80	31.97	33.67
⋮	⋮	⋮	⋮
55	25.54	16.17	16.09

Quadro 2. Média dos observadores em 55 pacientes do HC da Unicamp para a variável IMC

Com o objetivo de verificar possíveis diferenças nas medições em cada par de observador, foi definida a variável resposta (y) da seguinte forma:

$$y = \begin{cases} 1, & \textit{se houve diferença maior que 3\% nas medidas realizadas entre cada par de observadores} \\ 0, & \textit{se não houve diferença maior que 3\% nas medidas realizadas entre cada par de observadores} \end{cases}$$

O modelo utilizado para este conjunto de dados foi:

$$\text{logito}(y_{ij}) = \text{intercepto} + b\ SEXO_{ij} + c\ IDADE_{ij} + d\ DIF2_{ij} + e\ DIF3_{ij} + E_{ij}\ ,$$

sendo y_{ij} a resposta do i-ésimo indivíduo no j-ésimo observador descrita anteriormente; SEXO uma covariável independente do tempo que assume valores 1, para o sexo feminino e 0 para o sexo masculino; IDADE uma covariável contínua também independente do tempo; DIF2 e DIF3 variáveis

indicadoras definidas da seguinte forma:

DIF2	DIF3	
0	0	se a diferença ocorreu entre o 1° e 2° observadores
1	0	se a diferença ocorreu entre o 1° e 3° observadores
0	1	se a diferença ocorreu entre o 2° e 3° observadores

e o E_{ij} a componente do erro aleatório do i-ésimo indivíduo no j-ésimo observador.

Para cada variável estudada empregaram-se três tipos de matrizes correlação de trabalho: independente, permutável e não estruturada, e as estimativas obtidas foram comparadas. Foram utilizados os programas RMGEE e a macro do SAS (versões 1 e 2.03). Os valores das estimativas e erros-padrões são mostrados nas *Tabelas 2, 3* e *4*:

Tabela 2. Estimativa e erro-padrão usando as EEG para dados do HC com três matrizes de correlação de trabalho distintas para a variável IMC

Covariável	Estrutura de Correlação		
	Independente	Permutável	Não Estruturada
Intercepto	-1.54.52	-1.5757	-1.5432
	(1.0636)	(1.0862)	(1.0262)
Sexo	0.6351	0.6046	0.5466
	(0.6808)	(0.6783)	(0.6811)
Idade	-0.0186	-0.0172	-0.0174
	(0.0245)	(0.0248)	(0.0242)
Dif2 (*)	-0.9379	-0.9359	-0.9398
	(0.4601)	(0.4593)	(0.4618)
Dif3 (**)	-0.3825	-0.3817	-0.3825
	(0.4713)	(0.4698)	(0.4688)

(*) Diferença entre o 1° e 3° observadores
(**) Diferença entre o 2° e 3° observadores

Tabela 3. Estimativa e erro-padrão usando as EEG para dados do HC com três matrizes de correlação de trabalho distintas para a variável SC

Covariável	Estrutura de Correlação		
	Independente	Permutável	Não Estruturada
Intercepto	-1.5723	-1.4976	-1.6039
	(0.6734)	(0.6947)	(0.6580)
Sexo	0.7654	0.6726	0.9008
	(0.4927)	(0.4943)	(0.4843)
Idade	0.0242	0.0232	0.0235
	(0.0159)	(0.0155)	(0.0159)
Dif2 (*)	-0.3477	-0.3285	-0.3422
	(0.2337)	(0.2312)	(0.2312)
Dif3 (**)	-1.8852	-1.8690	-1.8802
	(0.4596)	(0.4637)	(0.4580)

(*) Diferença entre o 1° e 3° observadores
(**) Diferença entre o 2° e 3° observadores

Tabela 4. Estimativa e erro- padrão usando as EEG para dados do HC com três matrizes de correlação de trabalho distintas para a variável CMB

Covariável	Estrutura de Correlação		
	Independente	Permutável	Não Estruturada
Intercepto	-1.6317	-1.6292	-1.6444
	(0.6101)	(0.6100)	(0.6100)
Sexo	-0.0319	-0.0259	-0.0457
	(0.4590)	(0.4600)	(0.4550)
Idade	0.0123	0.0122	0.0128
	(0.0140)	(0.0140)	(0.0140)
Dif2 (*)	0.1932	0.1931	0.1932
	(0.3861)	(0.3860)	(0.3860)
Dif3 (**)	0.0988	0.0988	0.0989
	(0.4080)	(0.4070)	(0.4080)

(*) Diferença entre o 1° e 3° observadores
(**) Diferença entre o 2° e 3° observadores

Observando os valores das estimativas dos parâmetros e seus erros-padrões para cada estrutura de correlação, verificou-se que tanto as estimativas quanto os erros não eram muito diferentes. Estes resultados estão de acordo com a literatura, a qual diz que, quando a correlação entre as observações não é muito forte, os resultados das estimativas dos coeficientes do modelo não são diferentes para as diferentes matrizes correlação de trabalho (QU et al., 1995).

Examinando a *Tabela 5*, verificou-se que a variável IMC possui diferença significativa na variável DIF2 entre os observadores, ou seja, os dados indicaram que o 1° e o 3° observadores tiveram diferença nas medições maior que 3%. A *Tabela 5* mostra os valores das estimativas e do teste de Wald.

Tabela 5. Estimativa, estatística do teste e nível descritivo para a variável IMC com diferentes matrizes de correlação de trabalho

Covariável	Estrutura de Correlação								
	Independente			Permutável			Não Especificada		
	estimativa	Wald	p	estimativa	Wald	p	estimativa	Wald	p
Intercepto	-1.5452	2.1106	0.1463	-1.5757	2.1042	0.1469	-1.5432	2.2614	0.1446
Idade	0.6351	0.8703	0.3509	0.6046	0.7948	0.3727	0.5466	0.6440	0.3610
Sexo	-0.0186	0.5764	0.4485	-0.0172	0.4809	0.4885	-0.0174	0.5170	0.4201
Dif2 (*)	-0.9379	4.1555	0.0415	-0.9359	4.1522	0.0416	-0.9398	4.1416	0.0411
Dif3 (**)	-0.3825	0.6587	0.4170	-0.3817	0.6601	0.4165	-0.3825	0.6657	0.4134

(*) Diferença entre o 1° e 3° observadores
(**) Diferença entre o 2° e 3° observadores

Para a variável SC, observou-se que a DIF3 foi significativa, ou seja, existe evidência de diferença maior que 3% nas medições entre os 2° e 3° observadores. A *Tabela 6* apresenta estes valores.

Tabela 6. Estimativa, estatística do teste e valor de p para a variável SC com diferentes matrizes de correlação de trabalho

Covariável	Estrutura de Correlação								
	Independente			Permutável			Não Especificada		
	estimativa	Wald	p	estimativa	Wald	p	estimativa	Wald	p
Intercepto	-1.5723	5.4516	0.0196	-1.4976	4.6470	0.0311	-1.6039	5.9414	0.0149
Idade	0.7654	2.4131	0.1203	0.6726	1.8515	0.1736	0.9008	3.4596	0.0752
Sexo	0.0242	2.3165	0.1288	0.0232	2.2404	0.1341	0.0235	2.1845	0.1606
Dif2(*)	-0.3477	2.1398	0.1368	-0.3285	2.0186	0.1554	-0.3422	2.1907	0.1375
Dif3(**)	-1.8852	16.824	0.0002	-1.8690	16.246	0.0001	-1.8802	16.853	0.0000

(*) Diferença entre o 1° e 2° observadores
(**)Diferença entre o 2° e 3° observadores

Para a variável CMB, observou-se que nenhuma das covariáveis estudadas mostrou influência significativa na variável resposta. A *Tabela 7* apresenta estes valores.

Tabela 7. Estimativa, estatística do teste e nível descritivo para a variável CMB com diferentes matrizes de correlação de trabalho

Covariável	Estrutura de Correlação								
	Independente			Permutável			Não Especificada		
	estimativa	Wald	p	estimativa	Wald	p	estimativa	Wald	p
Intercepto	-1.6317	7.1551	0.0075	-1.6292	7.1332	0.0076	-1.6437	7.2609	0.0070
Idade	-0.0319	0.0048	0.9446	-0.0257	0.0032	0.9550	-0.0457	0.0101	0.9199
Sexo	0.0123	0.7845	0.3790	0.0122	0.7593	0.3851	0.0128	0.8359	0.3561
Dif2 (*)	0.1932	0.2505	0.6169	0.1931	0.2503	0.6170	0.1932	0.2505	0.6170
Dif3 (**)	0.0988	0.0417	0.8084	0.0988	0.0597	0.8084	0.0989	0.0588	0.8084

(*) => Diferença entre o 1° e 3° observadores
(**) => Diferença entre o 2° e 3° observadores

As *Tabelas 2, 3 e 4* anteriores mostram estimativas e erros-padrões com valores semelhantes nos diferentes tipos de matrizes de correlação de trabalho. As diferenças nas medições entre os observadores mostram-se mais evidentes nas variáveis índice de massa corporal e superfície corpórea. A diferença foi significativa entre os 1° e 3° observadores para a variável índice de massa corporal, diferença esta maior que 3% entre uma medida e outra. Para a variável superfície corpórea, a diferença nas medições foi significativa entre os 2° e 3° observadores. Como de acordo com a literatura médica, diferenças nas medições ao nível de 5% já são consideradas despresíveis, foram realizadas análises também considerando diferenças entre as medições a esse nível. Observou-se que nenhuma variável apresentou significância estatística entre as medidas dos observadores. Portando, fazendo uma análise geral do ponto de vista médico, observou-se que os erros de medições entre os observadores nas três variáveis estudadas foi bem pequeno, isto é, apesar de ter havido duas estatísticas significativas, as diferenças entre as medições foram maiores que 3% e menores que 5%. Como estas diferenças são consideradas não relevantes segundo critérios de medidas nutricionais, concluiu-se portanto, que os observadores dessas enfermarias do HC têm coletado as medidas nutricionais sem maiores discrepâncias.

Segundo Conjunto de Dados

Pacientes portadores de esquistossomose e alteração da pressão sangüínea na veia que drena o sangue do intestino para o fígado (veia porta) possuem varizes no esôfago e/ou estômago, que os levam, em muitos casos, à hemorragia digestiva. Para eliminar estas varizes, usa-se um procedimento cirúrgico de desvascularização. Um grupo de 36 pacientes portadores dessas varizes foram operados na Disciplina de Moléstias do Aparelho Digestivo do Departamento de Cirurgia da Faculdade de Ciências Médicas da Unicamp, no período de janeiro de 1982 a fevereiro de 1989 (BOIN, 1991).

Todos os pacientes se encontravam em estado avançado da doença apresentando aumento no

tamanho do baço, fígado fibrótico e barriga d'água. Relataram, ainda, pelo menos um episódio de hemorragia digestiva caracterizada por vômito de sangue e/ou, por defecação de sangue, com necessidade de reposição sangüínea. Esta foi criteriosa, sendo o volume médio de 800ml de concentração de hemácias e 800ml de plasma fresco.

Os doentes foram submetidos aos mesmos cuidados pós-operatórios imediatos e receberam alta hospitalar em média no décimo dia de pós-operatório, retornando ao ambulatório para controle imediato no intervalo de sete a dez dias da alta hospitalar. Todos os pacientes foram orientados a retornar para revisão cirúrgica e controle de exames laboratoriais e endoscopias a partir do 30° dia pós-operatório e pelo menos a cada seis meses, durante os dois primeiros anos após a cirurgia.

Com a finalidade de avaliar a eficácia da cirurgia ao longo do tempo, foram coletadas amostras de sangue e avaliado os aspectos endoscópicos das varizes esofágicas nos indivíduos operados após um mês, seis meses e 12 meses ou mais da cirurgia. As variáveis foram: HEMOGLOBINA em g%; dosagens séricas de alanina aminotransferase [ALT (UI/l)]; SEXO; IDADE (em anos); bilirrubinas totais em mg% (BT); tamanho das varizes esofágicas (TAMANHO); presença ou ausência de "barriga d' água" (ascite) e confusão mental (encefalopatia), atividade de protrombina que mede a coagulação do sangue (AP) e de acordo com os parâmetros clínicos e laboratoriais utilizou-se a classificação de Child (CHILD).

A variável resposta foi denotada como o sangramento ocorrido após a cirurgia, chamada de recidiva hemorrágica (RH).

Portanto:

$$RH = \begin{cases} 1, & \textit{se houve sangramento pós-cirurgia} \\ 0, & \textit{se não houve sangramento pós-cirurgia} \end{cases}$$

Dos 36 pacientes tratados, seis não foram considerados na análise, por não ter sido possível coletar todos os dados sobre eles. A estrutura dos dados pode ser vista no *Quadro 3* a seguir.

Paciente	hemoglobina			ALT (u l)			AP (%)			BT			Tamanho [1]			Child [2]			Sexo [3]	Idade
	1m	6m	>12m	1m	6m	>12m	1m	6m	>12m	1m	6m	>12m	1m	6m	>12m	1m	6m	>12m		
01	11	10	11	7	7	7	86	82	86	1.1	1.1	1.1	1	1	0	0	0	0	1	34
02	9.5	9.5	10.5	12	13	12	54	70	90	0.8	0.8	0.6	1	0	1	1	1	0	0	29
03	9.5	10	10.4	15	14	15	93	90	94	1.0	1.0	0.8	1	0	1	0	0	0	0	39
04	9.9	11	12	11	12	12	90	100	100	4.0	9.2	2.8	1	0	0	1	1	0	0	19
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
30	10.7	10.5	10	11	13	14	70	86	86	0.7	0.2	0.2	0	0	0	1	1	1	0	38

Quadro 3. Estrutura dos dados para os 30 pacientes portadores de varizes esofágicas no HC da Unicamp

[1] valor 1 se não houve redução no tamanho da variz e 0 caso contrário.
[2] valor 1 se a classificação for B (soma entre sete e nove pontos) e 0 se a classificação for A (soma entre cinco e seis pontos).
[3] valor 1 se o indivíduo é do sexo feminino e 0 caso contrário.

Diante do número relativamente elevado de variáveis em relação ao número de pacientes observados, inicialmente foram realizadas análises univariadas para uma melhor compreensão do comportamento destas covariáveis em relação à variável resposta. A partir destas análises foi proposto o modelo multivariado descrito abaixo:

$$\text{logito}(RH_{ij}) = \text{intercepto} + a\ \text{HEMOGLOBINA}_{ij} + b\ \text{TAMANHO}_{ij} + E_{ij}, \quad \text{(Modelo A)}$$

onde RH_{ij} é a resposta do i-ésimo indivíduo no j-ésimo tempo; HEMOGLOBINA uma covariável contínua; TAMANHO uma covariável categorizada assumindo valor 1 quando a variz não foi reduzida e 0 caso contrário e E_{ij} a componente do erro aleatório do i-ésimo indivíduo no j-ésimo tempo.

Para cada variável estudada foram utilizados quatro tipos de matrizes de correlação de trabalho: independente, permutável, modelo AR-1 e a não estruturada e comparados os resultados

das estimativas para essas matrizes. Os valores das estimativas e erros-padrões são mostrados na *Tabela 8*.

Tabela 8. Estimativa e erro-padrão para os dados de varizes nas diferentes matrizes correlação de trabalho para o modelo log (RH_{ij}) ***= intercepto + a*** $HEMOGLOBINA_{ij}$ ***+ b*** $TAMANHO_{ij} + E_{ij}$

Covariável	Estrutura de Correlação			
	Independente	Permutável	AR - 1	Não Especificada
Intercepto	-1.8538	-2.7467	-2.0537	-2.7770
	(2.2950)	(2.0430)	(2.1890)	(2.0530)
Hemoglobina	-0.1702	-0.1229	-0.1637	-0.1170
	(0.1993)	(0.1584)	(0.1840)	(0.1655)
Tamanho	1.6593	2.2093	1.8208	2.3227
	(1.1070)	(1.3553)	(1.1550)	(1.1928)

Os resultados evidenciaram uma certa influência da variável TAMANHO na recidiva de hemorragia pós-cirurgia, o que não aconteceu para a variável HEMOGLOBINA. Diferentemente dos resultados no primeiro exemplo, as estimativas e erros-padrões variaram para as diferentes matrizes de correlação de trabalho. As estimativas e os resultados dos testes estatísticos são mostrados na *Tabela 9*.

Tabela 9. Estimativa, estatística do teste e nível descritivo do segundo conjunto de dados com diferentes matrizes correlação de trabalho para o modelo log (RH_{ij}) = intercepto + a $HEMOGLOBINA_{ij}$ + b $TAMANHO_{ij}$ + $erro_{ij}$

Covariável	Estrutura de Correlação											
	Independente			Permutável			AR - 1			Não Especificadea		
	Estimativa	Wald	p	Estimativa	Wald	p	Estimativa	Wald	p	Estimativa	Wald	p
Intercepto	-1.8538	0.6561	0.4197	-2.7467	1.7956	0.1788	-2.0537	0.8836	0.3470	-2.7770	1.8225	0.1587
Hemoglobina	-0.1704	0.7293	0.3924	-0.1229	0.6020	0.4400	-0.1637	0.7916	0.3745	-0.1170	0.4997	0.4500
Tamanho	1.6593	2.2467	0.1339	2.2093	2.6569	0.1031	1.8208	2.4850	0.1147	2.3227	3.7919	0.054

As diferenças obtidas nas estimativas e erros-pardões, para cada tipo de matriz de correlação de trabalho devem-se ao fato de haver, possivelmente, uma considerável estrutura de correlação entre as medidas repetidas. Provavelmente, o fator tempo, não incluído nesta análise por não terem sido levantados os tempos reais de retorno dos pacientes, está influindo nesse resultado. Observa-se também que a matriz de correlação identidade possui erros-padrões maiores na sua maioria, o que reforça o argumento de uma correlação considerável entre as medidas repetidas. Assim, será utilizada para as análises subseqüentes apenas a matriz de correlação Não Especificada por ser ela a que descreve melhor a estrutura de correlação para estes dados.

É observada, portanto, a influência da variável TAMANHO na recidiva de hemorragia pós-cirurgia (W= 3.7919). O valor estimado $\beta_2 = 2.3227$ indica que a medida que o tempo passa e o tamanho da variz não se reduz, há uma chance maior do individuo apresentar hemorragia digestiva. Avaliando a razão de chance (*odds ratio= OR*), observou-se um OR=10.21 com intervalo de confiança de [1.6207; 129.7681]. Assim, o indivíduo tem dez vezes mais chances de ter hemorragia digestiva quando o tamanho da variz não é reduzido.

Como a variável HEMOGLOBINA não foi significativa no modelo anterior, e segundo a

literatura médica a maior influência no sangramento pós-cirurgia é devido ao tamanho das varizes, foi construído o seguinte modelo com a finalidade exploratória:

$$\text{logito}\,(RH_{ij}) = \text{intercepto} + a\ TAMANHO_{ij} + E_{ij} \qquad \text{(Modelo B)}$$

Os valores dos testes estatísticos e das estimativas são mostrados na *Tabela 10* para diferentes matrizes de correlação de trabalho.

Tabela 10. Estimativa, estatística do teste e nível descritivo do segundo conjunto de dados com diferentes matrizes correlação de trabalho para o modelo $log\,(RH_{ij}) = intercepto + a\ TAMANHO_{ij} + E_{ij}$

Covariável	Estrutura de Correlação											
	Independente			Permutável			AR - 1			Não Especificada		
	estimativa	Wald	p	estimativa	Wald	p	estimativa	Wald	p	estimativa	Wald	p
Intercepto	-3.7136	13.1769	0.0002	-4.1042	9.4655	0,0011	-3.8337	12.7442	0.0003	-4.0918	10.4297	0.0003
Tamanho	1.7677	2.5434	0.1128	2.3167	2.8541	0.0916	1.9160	2.7806	0.1023	2.4388	3.8620	0.0499

Examinando-se o valor da estatística Wald (W= 3.8620) nota-se que realmente a variável TAMANHO tem influência na variável resposta. A razão de chance é *OR*=11.46 e intervalo de confiança de [10.12 ; 149.08], indicando que a não redução no tamanho da variz aumenta a chance em 11 vezes do indivíduo ter hemorragia digestiva.

Percebe-se, neste exemplo, que as estimativas e o teste de Wald apontam resultados diferentes para cada estrutura de correlação. Estas diferenças devem-se a existência de uma considerável estrutura de correlação entre as medidas repetidas. Comparando-se as estimativas e erros-padrões para os dois modelos (modelo A e B), nota-se que houve uma redução nos erros das estimativas para

o intercepto, enquanto que para a variável TAMANHO quase não houve alteração. A *Tabela 11* mostra estes valores.

Tabela 11. Estimativa e erro-padrão para os dados de varizes nas diferentes matrizes de correlação de trabalho para os modelos A e B

Covariável	Estrutura de Correlação							
	Indep(A)	Indep(B)	Permut(A)	Permut(B)	AR-1(A)	AR-1(B)	Não Esp(A)	Não Esp(B)
Intercepto	-1.8538	-3.7136	-2.7467	-4.1042	-2.0537	-3.8377	-2.7770	-4.0918
	(2.2950)	(1.0220)	(2.0430)	(1.3340)	(2.1890)	(1.0750)	(2.0530)	(1.2670)
Tamanho	1.6593	1.7677	2.2093	2.3168	1.8208	1.9160	2.3227	2.4388
	(1.1070)	(1.1080)	(1.3553)	(1.3713)	(1.1550)	(1.149)	(1.1928)	(1.2410)

(A) Modelo A
(B) Modelo B

Face aos resultados obtidos e as análises realizadas neste exemplo, concluiu-se que a variável TAMANHO influencia a resposta. Outro fator que, provavelmente, deveria ser incluído no modelo é o TEMPO, logo, é importante que seja feito o levantamento cuidadoso dos dados referentes ao tempo de retorno de cada paciente operado.

3.2.2. Comentários Finais

O método das EEG vem ganhando bastante espaço nas pesquisas científicas, para análise de dados com medidas repetidas, por ser uma técnica que tem um procedimento numérico relativamente simple e amplo, e por possuir estimativas consistentes mesmo quando se utiliza uma estrutura incorreta para a matriz de correlação. Por outro lado, apresenta algumas desvantagens pois a especificação do modelo é incompleta e as estimativas são ineficientes ao se utilizar uma estrutura de correlação incorreta (LIANG et al., 1992; FITZMAURICE et al., 1993).

As EEG ainda possuem várias limitações principalmente na parte computacional, por se tratar de uma metodologia recente. Os pacotes existentes possuem poucas opções para as funções de ligação e sobretudo para as estruturas de correlação, dificultando as análises de dados mais complexos. Outro problema encontrado é a existência de dados faltantes ou conglomerados não balanceados no conjunto de dados. Este problema é tratado em alguns pacotes quando os dados faltantes são completamente aleatórios. Já a teoria chama bastante a atenção para estes casos, pois bancos de dados com tais problemas podem apresentar estimadores viciados em experimentos não balanceados (FIRTH, 1992).

Há um grande espaço para novas pesquisas com esta metodologia quando se trata de conglomerados não balanceados, dados faltantes e nos casos de amostras pequenas.

*Referências Bibliográficas

1. AGRESTI, A. - A survey of models for repeated ordered categorical response data. **Statist. Med., 8:**1209-24, 1989

2. ALBERT, P.S. & McSHANE, L.M. - A generalized estimating equations approach for spatially correlated binary data: applications to the analysis of neuroimaging data. **Biometrics, 51:**627-38, 1995.

3. ANDERSEN, E.B. - **Discrete statistical models with social science applications.** New York: North - Holland, 1980.

4. ARTES, R. - **Extensões da teoria das equações de estimação generalizadas a dados circulares e modelos de dispersão**. São Paulo, 1997. [Tese Doutorado -Universidade de São Paulo]

5. ASHBY, M.; NEUHAUS, J.M.; HAUCK, W.W.; BACCHETTI, P; HEILBRON, D.C.; JEWELL, N.P.; SEGAL, M.R.; FUSARO, R.E. - An annotated bibliography of methods for analyzing correlated categorical data. **Statist. Med., 11:**67-99, 1992.

6. BOIN, I. F. S.F. - **Resultados da cirurgia de desvascularização esofagogástrica associada à esplenectomia e escleroterapia programada no pós-operatório em doentes esquistossomóticos hepatesplênicos.** Campinas, 1991. [Tese Mestrado - Universidade Estadual de Campinas]

---

* HERANI, M.L.G. - Normas para a apresentação de dissertações e teses. São Paulo, Bireme, 1990

7. BOIN, I.F.S.F.; RAMOS, M.L.C.; ANDRADE, RG.; BRAILLE, M.C.V.B.; LEONARDI, L.S.; - Avaliação dos erros de medidas na determinação da circunferência muscular do braço, circunferência braquial e da prega cutânea tricipital. **Nutrition, 4**:97, 1988

8. CAREY, V.; ZEGER, S.L.; DIGGLE, P. - Modelling multivariate binary data with alternating logistic regressions. **Biometrika, 80**:517-26, 1993.

9. CARR, G.J. & CHI, E.M. - Analysis of variance for repeated measures data: a generalized estimating equations approach. **Statist. Med., 11**: 1033-40, 1992.

10. CATALANO, P.J. & RYAN, L.M. - Bivariate latent variable models for clustered discrete and continuos outcomes. **J. Am. Statist. Assoc., 87**:651-8, 1992.

11. CONOWAY, M.R. - A random effects model for binary data. **Biometrica, 46**: 317-28, 1990.

12. CORDEIRO, G.M. - Modelos lineares generalizados. **In:VII Simpósio Nacional de Probabilidade e Estatística,** Anais. Campinas, 1986. 285p.

13. COX, D. R. & HIKLEY, D. V. - **Theoretical statistics.** London, Chapman and Hall, 1979.

14. DAVIS, C.S. - A computer program for regression analysis of repeated measures using generalized estimating equations. **Comput. Methods Progr. Biomed., 40**:15-31, 1993.

15. DAVIS, G.M. - **Applications of sample methodology to repeated measures Data structures in dentistry**. Chapel Hill, 1994. [Tese Doutorado - University of North Carolina]

16. DIGGLE, P.J.; LIANG, K.; ZEGER, S.L. - **Analysis of longitudinal data**. Oxford, Science Publication, 1994.

17. DOBSON, A.J. - **An introduction to generalized linear models.** London, Chapman and Hall, 1991.

18. DRUM, M. et al. - Comment. **Biometrika, 80:**300-9, 1993.

19. DUNLOP, D.D. - Regression for longitudinal data: a bridge from least squares regression. **Am. Statist., 48:**299-306, 1994.

20. FIELDING, K.L.; BRETTLE, R.P.; GORE, S.M.; O'BRIEN, F.; WYLD, R.; ROBERTSON, J.R.; WEIGHTMAN, R. - Heterosexual transmission of HIV analysed by generalized estimating equations. **Statist. Med., 14:**1365-78, 1995.

21. FIRTH, D. - Discussion of paper by K.Y. Liang, S.L. Zeger and B. Qaqish. **J. R. Statist. Soc., 45:**24-6, 1992.

22. FITZMAURICE, G.M.; LAIRD, N.M.; ROTNITZKY, A.G. - Regression models for discrete longitudinal responses. **Biometrika,** 80:284-309, 1993.

23. GRIZZLE, J.E.; STARMER, C.F.; KOCH, G.G. - Analysis of categorical data by linear models. **Biometrics, 25:**489-504, 1969.

24. HAVE, T.R.T.; LANDIS, R.; HARTZEL, J. - Population-averaged and cluster-specific models for clustered ordinal response data. **Statist. Med., 15:**2573-88, 1996.

25. KARIM, M.R. - GEE1 PC SAS. **Tecnical Report, 674.** Department of Biostatistics. The Johns Hopkins University, Baltimore, MD, 1989.

26. KEMPTHORNE, W.J. & KOCH, G.G. - **A general approach for the analysis of attribute data from a two-stage nested design: one and two treatments per cluster**. Contributions to Statistics: Essays in Honour of Norman L. Johnson, ed. P.K. Sen. New York, North Holland Publishing Company, 1983. p.259-280

27. KISH, L. - **Survey sampling.** New York : Wiley, 1965.

28. KOCH, G.G.; LANDIS, J.R.; FREEMAN, J.L.; FREEMAN, JR.D.H.; LEHNAN, R.G. - A general methodology for the analysis of experiments with repeated measurement of categorical data. **Biometrics, *33*:**133-58, 1977.

29. KOCH, G.G.; LANDIS, J.R.; FREEMAN, J.L.; FREEMAN, D.H.JR. - **Categorical data analysis. In: Statistical methodology in the pharmaceutical sciences.** D. A. Berry (ed). New York, 1989. p.391-475

30. KUPPER, L.L. & HOSEMAN, J.K. - The use of a correlated binomial model for the analysis of certain toxicological experiments. **Biometrics, 34:**69-76, 1978.

31. LAIRD, N.M. & WARE, J.H. - Rondom-effects models for longitudinal data. **Biometrics, 38:**963-74, 1982.

32. LANDIS, J.R.; MILLER, M.E.; DAVIS, C.S.; KOCH, G.G. - Some general methods for the analysis of categorical data in longitudinal studies. **Statist. Med, 7:**109-37, 1988.

33. LAVANGE, L.M.; KEYES, L.L.; KOCH, G.G.; MARGOLIS, P.A. - Application of sample survey methods for modelling rations to incidence densities. **Statist. Med., 13:**343-55, 1994.

34. LEFKOPOULOU, M.; MOORE, D.; RYAN, L. - The analysis of multiple correlated binary outcomes : application rodent teratology experiments. **J. Am. Statist. Assoc., 84:**810-5, 1989.

35. LEWIS, K.E. - **Evaluating clinical equivalence of treatments from a repeated measures design.** Chapel Hill, 1993. [Tese Mestrado - University of North Carolina]

36. LIANG, K. & ZEGER, S.L. - Longitudinal data analysis using generalized linear models. **Biometrika, 73:**13-22, 1986.

37. LIANG, K.Y.; ZEGER, S.L.; QAQISH, B.F. - Multivariate regression analysis for categorical data. **J. Roy. Statist. Soc., 54:**3-40, 1992 .

38. LIPSITZ, S.R.; FITZMAURICE, G.M.; ORAV, E.J.; LAIRD,N.M. - Performance of generalized estimating equations in practical situations. **Biometrics, 50:**270-8, 1994.

39. LIPSITZ, S.R.; LAIRD, N.M.; HARRINGTON, D.P. - Generalized estimating equations for correlated binary data : using the odds ratio as a measure of association. **Biometrika, 78:**156-60, 1991.

40. LIPSITZ, S.R.; KIM, K.; ZHAO, L. - Analysis of repeated categorical data using generalized estimating equations. **Statist. Med., 13:**1149-63, 1994.

41. MARQUES, E.H.F. - **Analysis of categorical data from longitudinal studies of subjects with possibly clustered structures**. Chapel Hill, 1987. [Tese Doutorado - University of North Carolina]

42. McCULLAGH, P. - Quasi-likelihood functions. **Ann. Statist., 11:**59-67, 1983.

43. McCULLAGH, P. & NELDER, J.A. - **Generalized linear models**. London, Chapman and Hall, 1989.

44. MILLER, M.E.; DAVIS, C.S.; LANDIS, J.R. - The analysis of longitudinal polytomous data: generalized estimating equations and connections with weighted least squares. **Biometrics, 49:**1033-44, 1993.

45. NEUHAUS, J.M. & JEWELL, N.P. - The effect of retrospective sampling on binary regression models for clustered data. **Biometrics, 46:**977-90, 1990.

46. PACK, A.R.C.; COXHEAD, L.J.; McDONALD, B.W. - The prevalence of overhanging margins in posterior amalgam resotrations and periodontal consequences. **J. Clin. Periodontol., 17:**145-52, 1990.

47. PARK, T. - Multivariate regression models for discrete and continuous repeated measurements. **Comm. Statist. Theory Methods, 23:**1547-64, 1994.

48. PARK, T. & WOOLSON, R.F. - Generalized multivariate models for longitudinal data. **Comm. Statist. Simulation Comput., 21:**925-46, 1992.

49. PIERCE, D.A. & SANDS, B.R. - Extra-Bernoulli variation in regression of binary data. **Technical Report, 46**, 1975.

50. PODGOR, M.J. & HELLER, R. - Associations of types of lens opacities between and within eyes of individuals: an application of second-order generalized estimating equations. **Statist. Med., 15:**145-56, 1996.

51. PRENTICE, R.L. - Correlated binary regression with covariates specific to each binary observation. **Biometrics, 44:**1033-48, 1988.

52. PRENTICE, R.L. & ZHAO, P. - Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. **Biometrics, 47:**825-39, 1991.

53. QAQISH, B.F. & LIANG, K. - Marginal models for correlated binary responses with multiple classes and multiple levels of nesting. **Biometrics, 48:**939-50, 1992.

54. QU, Y.; PIEDMONTE, M.R.; MENDENDORP, S.V. - Latent variable models for clustered ordinal data. **Biometrics, 51:**268-75, 1995.

55. QU, Y.; WILLIAMS, G.W.; BECK, G.J.; MEDENDORP, S.V.B. - Latent variable models for clustered dichotomous data with multiple subclusters. **Biometrics, 48:**1095-1102, 1992.

56. RAO, J.N.K. & SCOTT, A.J. - A Simple method for the analysis of clustered binary data. **Biometrics, 48:**577-85, 1992.

57. RASCH, G. - **Probabilistic models for some intelligence and attainment tests**. Copenhagem: Danmarks Paedagogiske Institut, 1960.

58. ROSNER, B. - Multivariate methods for clustered binary data With for clustered more than one level of nesting. **J. Am. Statist. Assoc. Appl. Case Studies, 84**, 1989.

59. ROSNER, B. - Multivariate methods in ophthalmology with application to other paired-data situations. **Biometrics, 40:**1025-35, 1984.

60. ROTNITZKY, A. & JEWELL, N.P. - Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. **Biometrika,** 77:485-97, 1990.

61. SNYDER, E.S. - **The analysis of binary data with large, unbalanced, and incomplete clusters using ratio means weighted regression methods**. Chapel Hill, 1993. [Tese Doutorado - University of North Carolina]

62. STANISH, W.M.; GILLINGS, D.B.; KOCH, G.G. - An application of multivariate ratio methods for probability of success as variable between sets of trials. **J. Roy. Statist. Soc., 10:**257-61, 1978.

63. STIRATELLI, R.; LAIRD, N.; WARE, J.H. - Random-effects models for serial observations with binary response. **Biometrika, 61:**439-47, 1984.

64. STRAM, D.O.; WEI, L.J.; WARE, J.H. - Analysis of repeated ordered categorical outcomes with possibly missing observations and time-dependent covariates. **J. Am. Statist. Assoc., 83:**631-7, 1988.

65. WAITVBERG, D. L. - Avaliação nutricional. In: Waitvberg D. L. - **Nutrição interal e parenteral na prática clínica.** 2 ed. São Paulo, Editora Ateneu, 1995. Cap13.

66. WARE, J.H. - Linear models for the analysis of longitudinal studies. **Am. Statist., 39:** 1985.

67. WARE, J.H.; LIPSITZ, S.; SPEIZER, F.E. - Issues in the analysis of repeated categorical outcomes. **Statist. Med.,** 7:95-107, 1988.

68. WEI, L.J. & STRAM, D.O. - Analysing repeated measurements with possibly missing observations by modelling marginal distributions. **Statist. Med.,** 7:139-48, 1988.

69. WEDDERBURN, R.W.M. - Quasi-likelihood function, generalized linear models and the Gauss-Newton methods. **Biometrika, 61:**439-47, 1974.

70. ZEGER, S.L. - Commentary. **Statist. Med., 7:**161-68, 1988 .

71. ZEGER, S.L. & KARIM, M.R. - Generalized linear model with random effects; gibbs sampling approach. **J. Am. Statist. Assoc. Theory and Methods, 86**: 1991.

72. ZEGER, S.L. & LIANG, K. - Longitudinal data analysis for discrete and continuous outcomes. **Biometrics, 42:**121-30, 1986.

73. ZEGER, S.L.; LIANG, K.Y.; ALBERT, P.A. - Methods for longitudinal data a generalized estimating equation approach. **Biometrics, 44:**1049-60, 1988.

74. ZEGER, S.L.; LIANG, K.; SELF, S.G. - The analysis of binary longitudinal data with time-independent covariates. **Biometrika,** 72:31-8, 1985.

75. ZHAO, L.P. & PRENTICE, R.L. - Correlated binary regression using a quadratic exponencial model. **Biometrika,** 77:642-8, 1990.

APÊNDICES

APÊNDICE A

Funções de Estimação

1. A Função Escore

Considere $y = (y_1,\ldots,y_k)^T$ um vetor de observações da realização do vetor aleatório $Y = (Y_1, Y_2,\ldots, Y_k)$, independentemente distribuídas com média $\mu = (\mu_1,\ldots,\mu_k)^T$. Seja um Modelo Linear Generalizado (MLG) no qual Y_i tem a densidade na família exponencial dada por:

$$\pi(y;\theta_i,\phi_i) = \exp\{\phi_i[y\theta_i - a(\theta_i) + c(y,\phi_i)]\} \qquad (1)$$

onde b e c são funções conhecidas e $\phi_i > 0$, $i = 1, \ldots, k$, chamado parâmetro de escala, supostamente conhecido para cada observação.

Agora, seja:

$$\eta = X\beta \qquad (2)$$

a estrutura linear de um modelo de regressão onde $\eta = (\eta_1,\ldots,\eta_k)^T$, $\beta = (\beta_1,\ldots,\beta_p)^T$ e X é uma matriz $k \times p$ $(p < k)$ conhecida e de posto p. A função linear η dos parâmetros desconhecidos chama-se preditor linear.

Num MLG a média μ de uma observação é dada por uma função conhecida de η, monótona e duplamente diferenciável:

$$\mu_i = h^{-1}(\eta), \quad i = 1,2,\ldots,k \qquad (3)$$

sendo h uma *função de ligação.*

Portanto, para um modelo satisfazendo (1), (2) e (3), define-se a função escore como

$$U(\beta) = \frac{\partial L(\beta)}{\partial \beta} \tag{4}$$

onde $L(\beta)$ é a função de verossimilhança do modelo.

COX & HINKLEY (1979) por exemplo, demostraram que, sob as condições de regularidade, esta função escore tem valor esperado zero e estrutura de covariância igual à matriz de informação de Fisher. Assim , $E\{U(\beta)\} = 0$ e :

$$Cov\ \{U(\beta)\} = E\{U(\beta)\,U(\beta)'\} = E\left\{\frac{-\partial^2 U(\beta)}{\partial \beta' \partial \beta}\right\} = V \tag{5}$$

Por uma versão do teorema central do limite, a distribuição assintótica de $U(\beta)$ é normal p-dimensional, isto é, $N_p(0,V)$. Para grandes amostras, a estatística escore definida por $E = U(\beta)' V^{-1} U(\beta)$ tem, aproximadamente distribuição χ^2_p, supondo o modelo com os parâmetros β especificados verdadeiros (CORDEIRO, 1986; DOBSON, 1990).

2. A Função de Quase Verossimilhança

Sejam $Y = (Y_1, \ldots, Y_k)^T$ variáveis aleatórias com média $E(Y) = \mu$ e matriz de covariância $Cov\ (Y) = \phi^{-1} V(\mu)$, onde $V(\mu)$ é semi-definida positiva, cujos elementos são funções conhecidas de μ, e ϕ é uma constante de proporcionalidade ou parâmetro de escala.

Em geral, μ é uma função ligação conhecida h^{-1} de um conjunto de parâmetros $\beta = (\beta_1, \ldots, \beta_p)^T$ desconhecidos. Usualmente, esta função tem um componente linear envolvendo

uma matriz X de ordem $k \times p$, $\mu = h^{-1}(X\beta)$. Sejam $y = (y_1,\ldots,y_k)^T$ as respostas observadas.

Para a variável aleatória Y_i define-se o logaritmo da função de quase verossimilhança $l(\mu_i;y_i)$ pela relação:

$$\frac{\partial l(\mu_i,y_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{V(\mu_i)} \tag{6}$$

onde $Var(Y_i) = \phi^{-1}V(\mu_i)$. Para as k observações, o logaritmo da função de quase verossimilhança, $l(\mu_i,y_i)$ é definido pelo sistema de equações diferenciais parciais:

$$\begin{aligned} &\frac{\partial l(\mu,y)}{\partial \mu} = V(\mu_i)^{-}(y_i-\mu_i), \text{ ou similarmente} \\ &l(\mu_i;y_i) = \int^{\mu_i} \frac{y_i-\mu_i'}{V(\mu_i')}\, d\mu_i' + \text{função de } y_i \end{aligned} \tag{7}$$

onde $V(\mu)^{-}$ é uma inversa generalizada de $V(\mu)$. A expressão (6) é uma extensão da definição do WEDDERBURN (1974).

Integrando (1) com respeito à μ_i (CORDEIRO,1986), vem $l(\mu_i;y_i) = y_i\theta_i - a(\theta_i) + c(y_i;\theta)$, onde

$$\theta_i = \int V(\mu_i)^{-}\, d\mu_i, \quad a'(\theta_i) = \mu_i \quad e \quad a''(\theta_i) = \frac{d\mu_i}{d\theta_i} = V(\mu_i) \tag{8}$$

Portanto, a densidade de Y_i pode ser escrita na forma da família exponencial de distribuição com um parâmetro. Logo, admitir que as observações têm distribuição na família exponencial com um parâmetro equivale a supor uma relação variância-média para os dados (CORDEIRO, 1986).

Pode ser difícil decidir qual deve ser a distribuição populacional, mas a forma da relação variância-média é muito mais fácil de ser postulada. Isto é o que torna a quase verossimilhança de muita utilidade.

2.1. Propriedades da Função de Quase Verossimilhança

O logaritmo da função de quase verossimilhança tem propriedades similares ao log da função de verossimilhança.

Teorema 1: Sejam y e l definidos como em (6), supondo-se que μ é expressa como função dos parâmetros $\beta_1,\ldots,\beta_p$. Então l tem as seguintes propriedades:

i) $E\left(\frac{\partial l}{\partial \mu}\right) = 0$

ii) $E\left(\frac{\partial l}{\partial \beta_i}\right) = 0$

iii) $E\left(\frac{\partial l}{\partial \mu}\right)^2 = -E\left(\frac{\partial^2 l}{\partial \mu^2}\right) = \frac{1}{V(\mu)}$

iv) $E\left(\frac{\partial l}{\partial \beta_i}\frac{\partial l}{\partial \beta_i}\right) = -E\left(\frac{\partial^2 l}{\partial \beta_i \partial \beta_i}\right) = \frac{1}{V(\mu)}\cdot\frac{\partial \mu}{\partial \beta_i}\cdot\frac{\partial \mu}{\partial \beta_i}$

A função de quase verossimilhança é uma boa função de estimação porque ela satisfaz às propriedades da função escore (McCULLAGH, 1983). Uma outra propriedade importante, para o logarítimo da quase verossimilhança $l(\mu_i;y_i)$ e da verossimilhança $L(\mu_i;y_i)$ é apresentada no corolário abaixo:

Corolário: Se a distribuição de y é especificada em termos de μ, de maneira que o log

da verossimilhança L seja definido, então:

$$-E\left(\frac{\partial^2 l}{\partial \mu^2}\right) \leq -E\left(\frac{\partial^2 L}{\partial \mu^2}\right) \qquad (9)$$

sendo uma conseqüência imediata da desigualdade de Cramer-Rao. A expressão (9) torna-se uma igualdade quando L é o log da função de verossimilhança. O lado esquerdo de (9) é uma medida da informação quando se conhece apenas a relação variância-média dos dados, enquanto o lado direito é a informação obtida pelo conhecimento da distribuição dos dados. Para a prova do Teorema 1 e o Corolário, ver WEDDERBURN (1974).

As propriedades das funções de quase verossimilhança são muito semelhantes às da máxima verossimilhança, exceto para o parâmetro de perturbação, ϕ, quando é desconhecido, é tratado separadamente de β e não é estimado pelo método dos QMP.

3. Função de Estimação

Uma função de estimação é, de uma maneira simplificada, uma função dos dados e dos parâmetros de interesse. Em termos práticos, elas são construídas de modo que suas raízes, quando existem, sejam estimativas dos parâmetros envolvidos. Em geral, deseja-se a construção de estimadores consistentes e com distribuição assintótica conhecida. Um ponto importante no estudo dessas funções é o estabelecimento de condições que garantam que os estimadores obtidos possuam boas propriedades (ARTES, 1997). Condições estas dadas pelas seguintes definições:

Definição 1: Seja a existência de uma amostra de k vetores aleatórios independentes $y_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})^T$, $i = 1, 2, \ldots, k$. A cada unidade amostral i associa-se uma função de estimação ψ_i e define-se o conceito de função de estimação para a amostra através de

$$\Psi_k(y;\theta) = \sum_{i=1}^{k} \psi_i(y_i;\theta). \qquad (10)$$

Definição 2: Uma função $\Psi(y;\theta) = (\psi_1(y;\theta), \psi_2(y;\theta), \ldots, \psi_p(y;\theta))^T$, $\Psi : \mathbf{X} \times \Theta \to \mathbb{R}^p$ é dita regular se para todo $\theta = (\theta_1, \ldots, \theta_p)^T \in \Theta$,

i. a função é não viciada $E_\theta\{\Psi(y;\theta)\} = 0$;

ii. a derivada parcial $\partial\Psi(y;\theta)/\partial\theta_i$ existe quase certamente para $y \in \chi$;

iii. é possível permutar o sinal de integração e diferenciação da seguinte forma:

$$\frac{\partial}{\partial\theta_i}\int_{\chi}\Psi(y;\theta)p(y;\theta)d\nu(y) = \int_{\chi}\frac{\partial}{\partial\theta_i}\{\Psi(y;\theta)p(y;\theta)\}\nu(y)$$

iv. $E_\theta = \{\psi_i(y;\theta)\psi_j(y;\theta)\} \in \mathbb{R}$, para $i,j = 1,\ldots,p$ e

$$V_\Psi(\theta) = E_\theta\{\Psi(y;\theta)\Psi^T(y;\theta)\}$$

é positiva definida e

v. $E_\theta\left\{\frac{\partial\psi_i}{\partial\theta_r}(y;\theta)\frac{\partial\psi_j}{\partial\theta_s}(y;\theta)\right\} \in \mathbb{R}$ e $S_\Psi(\theta) = E_\theta\{\nabla_\theta\Psi(y;\theta)\}$ é não singular, onde ∇_θ representa o operador gradiente em relação a θ.

APÊNDICE B

Matrizes de Correlação

1. Valores das Correlações nas $\mathbb{R}_i(\alpha)$

Primeiro Conjunto de Dados

Variável: Índice de Massa Corpórea (IMC)

Matriz de Correlação Identidade			Matriz de Correlação Permutável			Matriz de Correlação Não Estruturada		
1.00	0	0	1.00	0.332	0.332	1.040	0.584	0.559
0	1.00	0	0.332	1.00	0.332	0.592	0.934	-0.084
0	0	1	0.332	0.332	1.00	0.496	-0.079	1.224

Variável: Superfície Corporal (SC)

Matriz de Correlação Identidade			Matriz de Correlação Permutável			Matriz de Correlação Não Estruturada		
1.00	0	0	1.00	0.313	0.313	1.119	0.804	0.240
0	1.00	0	0.313	1.00	0.313	0.713	1.135	-0.057
0	0	1	0.313	0.313	1.00	0.240	-0.056	0.900

Variável: Circunferência Média do Braço (CMB)

Matriz de Correlação Identidade			Matriz de Correlação Permutável			Matriz de Correlação Não Estruturada		
1.00	0	0	1.00	0.228	0.228	1.083	0.252	0.179
0	1.00	0	0.228	1.00	0.228	0.230	1.114	0.317
0	0	1	0.228	0.228	1.00	0164	0.287	1.096

Segundo Conjunto de Dados

Modelo A : $\text{logito}(RH_{ij}) = \text{Intercepto} + a\ \text{Hemoglobina}_{ij} + b\ \text{Tamanho}_{ij} + E_{ij}$

M. de Correlação Identidade			M.de Correlação Permutável			M. de Correlação AR-1			M. de Correlação Não Estruturada		
1.00	0	0	1.00	-0.112	-0.112	1.00	-0.080	-0.006	1.00	-0.028	-0.189
0	1.00	0	-0.112	1.00	-0.112	-0.080	1.00	-0.080	-0.028	1.00	0.187
0	0	1	-0.112	-0.112	1.00	0.006	-0.080	1.00	-0.189	-0.187	1.00

Modelo 2 : $\text{logito}(RH_{ij}) = \text{Intercepto} + b\ \text{Tamanho}_{ij} + E_{ij}$

M. de Correlação Identidade			M. de Correlação Permutável			M. de Correlação AR-1			M. de Correlação Não Estruturada		
1.00	0	0	1.00	-0.116	-0.116	1.00	-0.080	-0.006	1.00	0.024	-0.203
0	1.00	0	-0.112	1.00	-0.116	-0.080	1.00	-0.080	0.024	1.00	- 0.181
0	0	1	-0.116	-0.116	1.00	0.006	-0.080	1.00	-0.203	-0.181	1.00

2. Exemplos de Estimadores de α

Exemplo 1: Seja $\alpha = (\alpha_1, \ldots, \alpha_{n_i-1})^T$, onde $\alpha_j = corr\,(y_{ij}, y_{i,j-1})$ para $j = 1, \ldots, n_i-1$. O estimador natural de α_j, dado β e ϕ é:

$$\hat{\alpha}_j = \phi \sum_{i=1}^{k} \hat{r}_{ij} \hat{r}_{i,j-1} / (k-p) \qquad (1)$$

Agora seja $\mathbb{R}_i(\alpha)$ uma matriz triagonal com $\mathbb{R}_{j,j-1} = \alpha_j$. Isto é equivalente ao modelo 1-dependente. Como um caso especial, toma-se $s=1$ e $\alpha_j = \alpha$ ($j = 1, \ldots, n_i - 1$). Então o α pode ser estimado por:

$$\hat{\alpha} = \sum_{j=1}^{n_i-1} \hat{\alpha}_j / (n_i - 1) \qquad (2)$$

A extensão para o modelo m-dependente é bastante complexa.

Exemplo 2: Seja $s=1$ e assume-se que $\alpha_j = corr\,(y_{ij}, y_{ij'})$ para todo $j \neq j'$. Esta é a estrutura de correlação permutável obtida do modelo de efeitos aleatórios com níveis aleatórios para cada indivíduo (LAIRD & WARE, 1982). Dado ϕ, α pode ser estimado como:

$$\hat{\alpha} = \phi \sum_{i=1}^{k} \sum_{j<j'} \hat{r}_{ij} \hat{r}_{ij'} \Big/ \left\{ \sum_{i=1}^{k} \frac{1}{2} n_i (n_i - 1) - p \right\} \qquad (3)$$

LIANG & ZEGER (1986) e LIPSITZ et al. (1994) trazem outros exemplos e demostrações para estimar as matrizes $\mathbb{R}_i(\alpha)$.

APÊNDICE C

LISTAGENS DO PROGRAMA RMGEE

Variáveis: Hemoglobina e Tamanho (1= não reduziu o tamanho; 0=reduziu o tamanho)

```
REPEATED MEASURES ANALYSIS USING GENERALIZED ESTIMATING EQUATIONS
LIANG AND ZEGER (1986) BIOMETRIKA 73:13-22

MISSING VALUE CODE:    -1.0
NUMBER OF VARIABLES TO BE READ IN (EXCLUSIVE OF SUBJECT ID):  4
INDEX OF THE RESPONSE VARIABLE:  1
NUMBER OF COVARIATES:  3
INDICES OF COVARIATES:  2  3  4
NUMBER OF COVARIATES IN JOINT TEST OF SIGNIFICANCE:  0
INTERCEPT: NO
TYPE OF RESPONSE VARIABLE:  3    (BINOMIAL)
CORRELATION STRUCTURE:  1     (INDEPENDENT)
TOTAL NUMBER OF SUBJECTS:           30
TOTAL NUMBER OF OBSERVATIONS:       90
NUMBER OF COMPLETE OBSERVATIONS:    90
MINIMUM NUMBER OF OBSERVATIONS PER SUBJECT:  3
MAXIMUM NUMBER OF OBSERVATIONS PER SUBJECT:  3

INICIAL ESTIMATE OF PARAMETER VECTOR BETA
     .1332646      -.009756      .093809

ESTIMATE OF BETA AT ITERACTION  1
   -1.4726700      -.038560      .369783

ESTIMATE OF BETA AT ITERACTION  2
   -1.7819000      -.095248      .886354

ESTIMATE OF BETA AT ITERACTION  3
    -1.790566      -.150039    1.394312

ESTIMATE OF BETA AT ITERACTION  4
    -1.835700      -.168894    1.625808

ESTIMATE OF BETA AT ITERACTION  5
    -1.853710      -.170434    1.658708

ESTIMATE OF BETA AT ITERACTION  6
    -1.853830      -.170447    1.659266
```

```
WORKING CORRELATION MATRIX
    1.000000      .000000      .000000
     .000000     1.000000      .000000
     .000000      .000000     1.000000

COVARIANCE MATRIX OF BETA
 .441455D+01 -.412535D+00 -.187208D+00
-.412535D+00  .397125D-01  .343355D-02
-.187208D+00  .343355D-02  .122537D+01

STANDARD ERRORS OF ESTIMATED PARAMETERS
    2.29500      .199280     1.106966

STANDARDIZED PARAMETERS (ESTIMATE/S.E.)
    -.80770     -.855313     1.498932
```

****

Variáveis: Hemoglobina e Tamanho (1=não reduziu o tamanho; 0=reduziu o tamanho)

```
REPEATED MEASURES ANALYSIS USING GENERALIZED ESTIMATING EQUATIONS
LIANG AND ZEGER (1986) BIOMETRIKA 73:13-22

MISSING VALUE CODE:    -1.0
NUMBER OF VARIABLES TO BE READ IN (EXCLUSIVE OF SUBJECT ID):  4
INDEX OF THE RESPONSE VARIABLE:  1
NUMBER OF COVARIATES:  3
INDICES OF COVARIATES:  2  3  4
NUMBER OF COVARIATES IN JOINT TEST OF SIGNIFICANCE:  0
INTERCEPT: NO
TYPE OF RESPONSE VARIABLE:  3    (BINOMIAL)
CORRELATION STRUCTURE:  2    (EXCHANGEABLE)
TOTAL NUMBER OF SUBJECTS:          30
TOTAL NUMBER OF OBSERVATIONS:      90
NUMBER OF COMPLETE OBSERVATIONS:   90
MINIMUM NUMBER OF OBSERVATIONS PER SUBJECT:  3
MAXIMUM NUMBER OF OBSERVATIONS PER SUBJECT:  3

INICIAL ESTIMATE OF PARAMETER VECTOR BETA
     .133246     -.009756      .093809

ESTIMATE OF BETA AT ITERACTION  1
    -.377100     -.118461      .064375

ESTIMATE OF BETA AT ITERACTION  2
   -1.175501     -.136507      .650543
```

```
ESTIMATE OF BETA AT ITERACTION  3
    -2.035631     -.135015     .583206

ESTIMATE OF BETA AT ITERACTION  4
    -2.64925      -.125735     2.142656

ESTIMATE OF BETA AT ITERACTION  5
    -2.753237     -.122735     2.220019

ESTIMATE OF BETA AT ITERACTION  6
    -2.742844     -.122287     2.206249

ESTIMATE OF BETA AT ITERACTION  7
    -2.747593     -.122287     2.210173

ESTIMATE OF BETA AT ITERACTION  8
    -2.746745     -.122283     2.209275

WORKING CORRELATION MATRIX
    1.000000      -.111918      -.111918
    -.111918      1.000000      -.111918
    -.111918      -.111918      1.000000

COVARIANCE MATRIX OF BETA
 .290591D+01 -.264953D+00 -.284724D+00
-.264953D+00  .250775D-01  .175300D-01
-.284724D+00  .175300D-01  .183696D+01

STANDARD ERRORS OF ESTIMATED PARAMETERS
    2.04300       .158359      1.355344

STANDARDIZED PARAMETERS (ESTIMATE/S.E.)
    -1.344466     -.772189     1.630048
```

****

Variáveis: Hemoglobina e Tamanho (1=não reduziu o tamanho; 0=reduziu o tamanho)

```
REPEATED MEASURES ANALYSIS USING GENERALIZED ESTIMATING EQUATIONS
LIANG AND ZEGER (1986) BIOMETRIKA 73:13-22

MISSING VALUE CODE:   -1.0
NUMBER OF VARIABLES TO BE READ IN (EXCLUSIVE OF SUBJECT ID):  4
INDEX OF THE RESPONSE VARIABLE:  1
NUMBER OF COVARIATES:  3
INDICES OF COVARIATES:  2  3  4
```

```
NUMBER OF COVARIATES IN JOINT TEST OF SIGNIFICANCE:  0
INTERCEPT: NO
TYPE OF RESPONSE VARIABLE:  3   (BINOMIAL)
CORRELATION STRUCTURE:  3    (UNSPECIFIED)
TOTAL NUMBER OF SUBJECTS:          30
TOTAL NUMBER OF OBSERVATIONS:      90
NUMBER OF COMPLETE OBSERVATIONS:   90
MINIMUM NUMBER OF OBSERVATIONS PER SUBJECT:  3
MAXIMUM NUMBER OF OBSERVATIONS PER SUBJECT:  3

INICIAL ESTIMATE OF PARAMETER VECTOR BETA
     .1332646     -.009756     .093809

ESTIMATE OF BETA AT ITERACTION  1
    -.0702240     -.139072     .048778

ESTIMATE OF BETA AT ITERACTION  2
    -.8903560     -.150763     .614599

ESTIMATE OF BETA AT ITERACTION  3
    -1.786965     -.142777    1.522459

ESTIMATE OF BETA AT ITERACTION  4
    -2.501956     -.127779    2.143104

ESTIMATE OF BETA AT ITERACTION  5
    -2.762335     -.119120    2.328659

ESTIMATE OF BETA AT ITERACTION  6
    -2.772698     -.117214    2.319918

ESTIMATE OF BETA AT ITERACTION  7
    -2.777021     -.117042    2.323029

ESTIMATE OF BETA AT ITERACTION  8
    -2.776997     -.117009    2.322711

WORKING CORRELATION MATRIX
    1.000000     -.028842     -.189415
     .028842     1.000000      .187013
    -.189415     -.187013     1.000000

COVARIANCE MATRIX OF BETA
 .314478D+01 -.288772D+00 -.177158D+00
-.288772D+00  .273745D-01  .157543D-01
-.177158D+00  .157543D-01  .142278D+01

STANDARD ERRORS OF ESTIMATED PARAMETERS
    2.053000      .156452     1.192803
```

```
STANDARDIZED PARAMETERS (ESTIMATE/S.E.)
    -1.352653      -.707206     1.947272

**********************************************************************
```

Variáveis: Hemoglobina e Tamanho (1= reduziu o tamanho; 0=não reduziu o tamanho)
Testando a interação para Hemog × Tamanho

```
REPEATED MEASURES ANALYSIS USING GENERALIZED ESTIMATING EQUATIONS
LIANG AND ZEGER (1986) BIOMETRIKA 73:13-22

MISSING VALUE CODE:    -1.0
NUMBER OF VARIABLES TO BE READ IN (EXCLUSIVE OF SUBJECT ID):  4
INDEX OF THE RESPONSE VARIABLE:  1
NUMBER OF COVARIATES:  3
INDICES OF COVARIATES:  2  3  4
NUMBER OF COVARIATES IN JOINT TEST OF SIGNIFICANCE:  2
INDICES OF COVARIATES IN JOINT SIGNIFICANCE TEST:  3  4
INTERCEPT: NO
TYPE OF RESPONSE VARIABLE:  3    (BINOMIAL)
CORRELATION STRUCTURE:  3     (UNSPECIFIED)
TOTAL NUMBER OF SUBJECTS:           30
TOTAL NUMBER OF OBSERVATIONS:       90
NUMBER OF COMPLETE OBSERVATIONS:    90
MINIMUM NUMBER OF OBSERVATIONS PER SUBJECT:  3
MAXIMUM NUMBER OF OBSERVATIONS PER SUBJECT:  3

INICIAL ESTIMATE OF PARAMETER VECTOR BETA
     .227074      -.009756      -.093809

ESTIMATE OF BETA AT ITERACTION  1
    -.119002      -.139072      -.048778

ESTIMATE OF BETA AT ITERACTION  2
    -.275758      -.150763      -.614599

ESTIMATE OF BETA AT ITERACTION  3
    -.264507      -.142777     -1.522459

ESTIMATE OF BETA AT ITERACTION  4
    -.358852      -.127779     -2.143104

ESTIMATE OF BETA AT ITERACTION  5
    -.433675      -.119120     -2.328659

ESTIMATE OF BETA AT ITERACTION  6
    -.452780      -.117214     -2.319918
```

```
ESTIMATE OF BETA AT ITERACTION  7
    -.453992      -.117042     -2.323029

ESTIMATE OF BETA AT ITERACTION  8
    -.454286      -.117009     -2.322711

WORKING CORRELATION MATRIX
    1.000000       .028842      -.189415
     .028842      1.000000      -.187013
    -.189415      -.187013      1.000000

COVARIANCE MATRIX OF BETA
 .314478D+01 -.288772D+00 -.177158D+00
-.288772D+00  .273745D-01  .157543D-01
-.177158D+00  .177543D-01  .142278D+01

STANDARD ERRORS OF ESTIMATED PARAMETERS
    1.773353       .165452      1.192803

STANDARDIZED PARAMETERS (ESTIMATE/S.E.)
    -.256173      -.707206     -1.947272

SUBSET OF PARAMETERS FOR JOINT SIGNIFICANCE TEST THAT ALL ARE
EQUAL TO ZERO
    -.117009     -2.322711

COVARIANCE MATRIX FOR SUBSET OF PARAMETERS
 .273745D-01  .157543D-01
-.407090D+00  .707358D+00

BETA'*SIGMA-INVERSE*BETA=  4.098     DF= 2     P=  .129
```

****

Testando interação: Hemog × Tamanho (1= reduziu o tamanho; 0=não reduziu o tamanho)

```
REPEATED MEASURES ANALYSIS USING GENERALIZED ESTIMATING EQUATIONS
LIANG AND ZEGER (1986) BIOMETRIKA 73:13-22

MISSING VALUE CODE:    -1.0
NUMBER OF VARIABLES TO BE READ IN (EXCLUSIVE OF SUBJECT ID):  5
INDEX OF THE RESPONSE VARIABLE:  1
NUMBER OF COVARIATES:  4
INDICES OF COVARIATES:  2  3  4  5
NUMBER OF COVARIATES IN JOINT TEST OF SIGNIFICANCE:  2
INDICES OF COVARIATES IN JOINT SIGNIFICANCE TEST:  3  4
```

```
INTERCEPT: NO
TYPE OF RESPONSE VARIABLE:  3    (BINOMIAL)
CORRELATION STRUCTURE:  2    (EXCHANGEABLE)
TOTAL NUMBER OF SUBJECTS:          30
TOTAL NUMBER OF OBSERVATIONS:      90
NUMBER OF COMPLETE OBSERVATIONS:   90
MINIMUM NUMBER OF OBSERVATIONS PER SUBJECT:  3
MAXIMUM NUMBER OF OBSERVATIONS PER SUBJECT:  3

INICIAL ESTIMATE OF PARAMETER VECTOR BETA
     .224984     -.009723     -.092948      .005959

ESTIMATE OF BETA AT ITERACTION  1
    -.443579     -.118434      .064928      .006949

ESTIMATE OF BETA AT ITERACTION  2
    -.539074     -.136755     -.642173      .057074

ESTIMATE OF BETA AT ITERACTION  3
    -.469475     -.135678    -1.569167      .076391

ESTIMATE OF BETA AT ITERACTION  4
    -.515134     -.126598    -2.130045      .045006

ESTIMATE OF BETA AT ITERACTION  5
    -.535168     -.123196    -2.213405      .018378

ESTIMATE OF BETA AT ITERACTION  6
    -.538597     -.122659    -2.202418      .014131

ESTIMATE OF BETA AT ITERACTION  7
    -.539898     -.122509    -2.206330      .014389

ESTIMATE OF BETA AT ITERACTION  8
    -.539979     -.122496    -2.205520      .014193

ESTIMATE OF BETA AT ITERACTION  9
    -.540031     -.122490    -2.205755      .014221

ESTIMATE OF BETA AT ITERACTION  10
    -.540031     -.122490    -2.205700      .014210

WORKING CORRELATION MATRIX
    1.000000     -.111757     -.111757
    -.111757     1.000000     -.111757
    -.111757     -.111757     1.000000
```

```
COVARIANCE MATRIX OF BETA
 .319823D+01 -.273765D+00 -.362075D+00 -.663483D+00
-.273765D+00  .243769D-01  .297853D-01  .516980D-01
-.362075D+00  .297853D-01  .175524D+01 -.150569D+00
-.663483D+00  .516980D-01 -.150569D+00  .465900D+00

STANDARD ERRORS OF ESTIMATED PARAMETERS
    1.788360      .156131     1.324853      .862568

STANDARDIZED PARAMETERS (ESTIMATE/S.E.)
    -.301970     -.784532    -1.664864      .020818

SUBSET OF PARAMETERS FOR JOINT SIGNIFICANCE TEST THAT ALL ARE
EQUAL TO ZERO
    -.122490    -2.205700

COVARIANCE MATRIX FOR SUBSET OF PARAMETERS
 .243769D-01  .297853D-01
 .297853D-01  .175524D+01

COVARIANCE MATRIX INVERSE
 .418911D+02 -.710867D+00
-.710867D+00  .581787D+00

BETA'*SIGMA-INVERSE*BETA=  3.075    DF= 2    P=  .215

**********************************************************************
```

Variáveis: Hemoglobina, Tamanho e Child (1=reduziu o tamanho; 0=não reduziu o tamanho)
Testando interação: Hemog × Tamanho

```
REPEATED MEASURES ANALYSIS USING GENERALIZED ESTIMATING EQUATIONS
LIANG AND ZEGER (1986) BIOMETRIKA 73:13-22

MISSING VALUE CODE:    -1.0
NUMBER OF VARIABLES TO BE READ IN (EXCLUSIVE OF SUBJECT ID):  5
INDEX OF THE RESPONSE VARIABLE:  1
NUMBER OF COVARIATES:  4
INDICES OF COVARIATES:  2  3  4  5
NUMBER OF COVARIATES IN JOINT TEST OF SIGNIFICANCE:  2
INDICES OF COVARIATES IN JOINT SIGNIFICANCE TEST:  3  4
INTERCEPT: NO
TYPE OF RESPONSE VARIABLE:  3     (BINOMIAL)
CORRELATION STRUCTURE:  1      (INDEPENDENT)
TOTAL NUMBER OF SUBJECTS:          30
TOTAL NUMBER OF OBSERVATIONS:      90
```

```
NUMBER OF COMPLETE OBSERVATIONS:    90
MINIMUM NUMBER OF OBSERVATIONS PER SUBJECT:  3
MAXIMUM NUMBER OF OBSERVATIONS PER SUBJECT:  3

INICIAL ESTIMATE OF PARAMETER VECTOR BETA
     .224984      -.009723      -.092948       .005959

ESTIMATE OF BETA AT ITERACTION  1
   -1.111026      -.038430      -.366384       .023489

ESTIMATE OF BETA AT ITERACTION  2
    -.912670      -.095174      -.878099       .055218

ESTIMATE OF BETA AT ITERACTION  3
    -.415013      -.150640     -1.381218       .083499

ESTIMATE OF BETA AT ITERACTION  4
    -.225041      -.170167     -1.610810       .093587

ESTIMATE OF BETA AT ITERACTION  5
    -.209090      -.171812     -1.643573       .094463

ESTIMATE OF BETA AT ITERACTION  6
    -.208948      -.171826     -1.644132       .094470

WORKING CORRELATION MATRIX
    1.000000      -.000000      -.000000
    -.000000      1.000000      -.000000
    -.000000      -.000000      1.000000

COVARIANCE MATRIX OF BETA
 .475544D+01 -.418464D+00 -.324826D+00 -.888740D+00
-.418464D+00  .380190D-01  .177480D-01  .730847D-01
-.324826D+00  .177480D-01  .120344D+01 -.208058D+00
-.888740D+00  .730847D-01 -.208058D+00  .461549D+00

STANDARD ERRORS OF ESTIMATED PARAMETERS
    2.180698       .194985     1.097016       .679374

STANDARDIZED PARAMETERS (ESTIMATE/S.E.)
    -.095817      -.881230    -1.498730       .139055

SUBSET OF PARAMETERS FOR JOINT SIGNIFICANCE TEST THAT ALL ARE
EQUAL TO ZERO
    -.171826    -1.644132

COVARIANCE MATRIX FOR SUBSET OF PARAMETERS
 .380190D-01  .177480D-01
 .177480D-01  .120344D+01
```

```
COVARIANCE MATRIX INVERSE
 .264850D+02 -.390592D+00
-.390592D+00  .836708D+00

BETA'*SIGMA-INVERSE*BETA=  2.823    DF= 2    P=  .244
```

****

Testando interação: Hemog × Tamanho (1=reduziu o tamanho; 0= não reduziu o tamanho)

```
REPEATED MEASURES ANALYSIS USING GENERALIZED ESTIMATING EQUATIONS
LIANG AND ZEGER (1986) BIOMETRIKA 73:13-22

MISSING VALUE CODE:   -1.0
NUMBER OF VARIABLES TO BE READ IN (EXCLUSIVE OF SUBJECT ID):  5
INDEX OF THE RESPONSE VARIABLE:  1
NUMBER OF COVARIATES:  4
INDICES OF COVARIATES:  2  3  4  5
NUMBER OF COVARIATES IN JOINT TEST OF SIGNIFICANCE:  2
INDICES OF COVARIATES IN JOINT SIGNIFICANCE TEST:  3  4
INTERCEPT: NO
TYPE OF RESPONSE VARIABLE:  3    (BINOMIAL)
CORRELATION STRUCTURE:  3     (UNSPECIFIED)
TOTAL NUMBER OF SUBJECTS:          30
TOTAL NUMBER OF OBSERVATIONS:      90
NUMBER OF COMPLETE OBSERVATIONS:   90
MINIMUM NUMBER OF OBSERVATIONS PER SUBJECT:  3
MAXIMUM NUMBER OF OBSERVATIONS PER SUBJECT:  3

INICIAL ESTIMATE OF PARAMETER VECTOR BETA
     .224984     -.009723      -.092948      .005959

ESTIMATE OF BETA AT ITERACTION  1
    -.133803     -.138877       .050898      .056317

ESTIMATE OF BETA AT ITERACTION  2
    -.284984     -.151165      -.608697      .049161

ESTIMATE OF BETA AT ITERACTION  3
    -.269357     -.143008     -1.517509      .022221

ESTIMATE OF BETA AT ITERACTION  4
    -.349687     -.127376     -2.150326     -.044882

ESTIMATE OF BETA AT ITERACTION  5
    -.413207     -.118221     -2.352410     -.092438
```

```
ESTIMATE OF BETA AT ITERACTION  6
    -.430184      -.115934      -2.351236     -.106609

ESTIMATE OF BETA AT ITERACTION  7
    -.431892      -.115619      -2.355727     -.107507

ESTIMATE OF BETA AT ITERACTION  8
    -.432264      -.115552      -2.355660     -.107923

ESTIMATE OF BETA AT ITERACTION  9
    -.432307      -.115541      -2.355844     -.107961

WORKING CORRELATION MATRIX
    1.000000       .029391      -.191849
     .029391      1.000000      -.188721
    -.191849      -.188721      1.000000

COVARIANCE MATRIX OF BETA
 .335922D+01 -.292595D+00 -.224175D+00 -.624171D+00
-.224175D+00  .264608D-01  .261515D-01  .472000D-01
-.224175D+00  .261515D-01  .134177D+01 -.200751D+00
-.624171D+00  .473000D-01 -.200751D+00  .485140D+00

STANDARD ERRORS OF ESTIMATED PARAMETERS
    1.832816       .162668      1.158346       .696520

STANDARDIZED PARAMETERS (ESTIMATE/S.E.)
    -.235870      -.710288     -2.033800      -.155001

SUBSET OF PARAMETERS FOR JOINT SIGNIFICANCE TEST THAT ALL ARE
EQUAL TO ZERO
    -.115541     -2.355844

COVARIANCE MATRIX FOR SUBSET OF PARAMETERS
 .264608D-01  .261515D-01
 .261515D-01  .134177D+01

COVARIANCE MATRIX INVERSE
 .385341D+02 -.751043D+00
-.751043D+00  .759925D+00

BETA'*SIGMA-INVERSE*BETA=  4.323    DF= 2    P=  .115


**************************************************************
```

Variável: Tamanho (1= não rediziu o tamanho; 0=reduziu o tamanho)

```
REPEATED MEASURES ANALYSIS USING GENERALIZED ESTIMATING EQUATIONS
LIANG AND ZEGER (1986) BIOMETRIKA 73:13-22

MISSING VALUE CODE:    -1.0
NUMBER OF VARIABLES TO BE READ IN (EXCLUSIVE OF SUBJECT ID):  4
INDEX OF THE RESPONSE VARIABLE:  1
NUMBER OF COVARIATES:  2
INDICES OF COVARIATES:  2  3
NUMBER OF COVARIATES IN JOINT TEST OF SIGNIFICANCE:  0
INTERCEPT: NO
TYPE OF RESPONSE VARIABLE:  3    (BINOMIAL)
CORRELATION STRUCTURE:  3     (UNSPECIFIED)
TOTAL NUMBER OF SUBJECTS:           30
TOTAL NUMBER OF OBSERVATIONS:       90
NUMBER OF COMPLETE OBSERVATIONS:    90
MINIMUM NUMBER OF OBSERVATIONS PER SUBJECT:  3
MAXIMUM NUMBER OF OBSERVATIONS PER SUBJECT:  3

INICIAL ESTIMATE OF PARAMETER VECTOR BETA
     .0238095      .101190

ESTIMATE OF BETA AT ITERACTION  1
   -1.633576      .029854

ESTIMATE OF BETA AT ITERACTION  2
    -2.598197     .771935

ESTIMATE OF BETA AT ITERACTION  3
    -3.428161     1.712757

ESTIMATE OF BETA AT ITERACTION  4
    -3.982871     2.456497

ESTIMATE OF BETA AT ITERACTION  5
    -4.107918     2.456497

ESTIMATE OF BETA AT ITERACTION  6
    -4.088257     2.435116

ESTIMATE OF BETA AT ITERACTION  7
    -4.092819     2.439918

ESTIMATE OF BETA AT ITERACTION  8
    -4.091818     2.438840
```

```
WORKING CORRELATION MATRIX
    1.000000      .023962     -.202917
     .023962     1.000000     -.180947
    -.202917     -.180947     1.000000

COVARIANCE MATRIX OF BETA
 .967220D-01 -.166081D-01
-.166081D-01  .154078D+01

STANDARD ERRORS OF ESTIMATED PARAMETERS
    1.26700      1.241280

STANDARDIZED PARAMETERS (ESTIMATE/S.E.)
   -4.091818     1.964779
```

LISTAGEM DA MACRO DO SAS (1ª versão)

Variáveis: Hemoglobina e Tamanho (1=nãoreduziu; 0=reduziu)

```
The SAS System

Regression analysis using GEE:
==============================

Data File: VARIZES

Outcome variable: SANGROU

Covariates: INTERCPT HEMO    TAMANHO

Link:  3 (Logit)

Variance: 3 (Binomial)

Denominator _1_

Correlation:  5  (AR - 1)

Total number of records read:      90

Total number of clusters:    30

Maximum and minimum custer size    3 and    3

Averages of Outcome variable and Covariates (over all)

                 SANGROU  INTERCPT      HEMO    TAMANHO

Observations:  0.0777778         1 10.815556  0.4666667
               0.0777778         1 10.815556  0.4666667

Inicial estimate of regression coefficients:
INTERCEPT  0.1332646
HEMO      -0.009756
TAMANHO    0.093809

===> Iteration:     1
```

```
  Estimate

INTERCEPT    -0.696951
HEMO         -0.092016
TAMANHO       0.102681

===> Iteration:      2

             Estimate

INTERCEPT    -1.289613
HEMO         -0.126262
TAMANHO       0.683636

===> Iteration:      3

             Estimate

INTERCEPT    -1.743967
HEMO         -0.150768
TAMANHO       1.368466

===> Iteration:      4

             Estimate

INTERCEPT    -2.007191
HEMO         -0.161088
TAMANHO       1.744989

===> Iteration:      5

             Estimate

INTERCEPT    -2.057119
HEMO         -0.163373
TAMANHO       1.820581

===> Iteration:      6

             Estimate

INTERCEPT    -2.053823
HEMO         -0.163681
TAMANHO       1.820772

===> Iteration:      7
```

```
              Estimate

INTERCEPT    -2.053689
HEMO         -0.163694
TAMANHO       1.820772

Convergence after   7 teration(s).

Working Correlation:
         1 -0.080002 0.0064003
-0.080002          1 -0.080002
0.0064003 -0.080002          1

Scale parameter: 1.0419768


Mean Squared Error: 0.0687238

Variance estimate (naive):

           INTERCEPT      HEMO     TAMANHO

INTERCEPT 7.9889677 -0.615283 -1.46847
HEMO       -0.925932 0.0607321  0.030375
TAMANHO    0.0344678 0.105445 1.3663477

Variance estimate (robust):

           INTERCEPT      HEMO     TAMANHO
INTERCEPT 4.7919967 -0.353156  -1.167988
HEMO       -0.868653  0.0340095 -0.0024812
TAMANHO    -0.073926 -0.0116479  1.3342324


NOTE: Covariances are above diagonal and correlations are below
diagonal.

Estimate, s.e. and z-score:

            Estimate   s.e.-Naive  s.e.-Robust  z-Robust
INTERCEPT -2.053689        2.526        2.189     -0.94
HEMO      -0.163694        0.246        0.184     -0.89
TAMANHO    1.820772        1.169        1.155      1.58

(c)  M. Rezaul Karim, 1989
Department of Biostatistics, The Johns Hopkins University


**************************************************************
```

Variável : Tamanho (1=não reduziu; 0=reduziu)

```
The SAS System

Regression analysis using GEE:
==============================

Data File: VARIZES

Outcome variable: SANGROU

Covariates: INTERCPT    TAMANHO

Link:  3 (Logit)

Variance: 3 (Binomial)

Denominator _1_

Correlation:  6  (Unspecified)

Total number of records read:       90

Total number of clusters:    30

Maximum and minimum custer size     3 and     3

Averages of Outcome variable and Covariates (over all)

                 SANGROU  INTERCPT        TAMANHO

Observations:  0.0777778         1    0.4666667
               0.0777778         1    0.4666667

Inicial estimate of regression coefficients:
INTERCEPT  0.0238095
TAMANHO    0.10119

===> Iteration:      1
```

```
                Estimate

INTERCEPT     -1.633576
TAMANHO        0.029854

===> Iteration:      2

                Estimate

INTERCEPT     -2.598197
TAMANHO        0.771935

===> Iteration:      3

                Estimate

INTERCEPT     -3.428161
TAMANHO        1.712757

===> Iteration:      4

                Estimate

INTERCEPT     -3.982871
TAMANHO        2.324034

===> Iteration:      5

                Estimate

INTERCEPT     -4.107918
TAMANHO        2.456497

===> Iteration:      6

                Estimate

INTERCEPT     -4.08827
TAMANHO        2.435116

===> Iteration:      7

                Estimate

INTERCEPT     -4.092819
TAMANHO        2.439918

===> Iteration:      8
```

```
             Estimate

INTERCEPT    -4.091818
TAMANHO       2.43884

Convergence after   8 iteration(s).

Working Correlation:
         1 0.0239369 -0.202872
0.0239369         1 -0.1809
-0.202872 -0.1806          1

Scale parameter: 1.11012


Mean Squared Error: 0.0698853

Variance estimate (naive):

           INTERCEPT      TAMANHO

INTERCEPT 1.371328      -1.42878
TAMANHO  -0.95356        1.6371694

Variance estimate (robust):

           INTERCEPT    TAMANHO
INTERCEPT 1.6042812    -1.524167
TAMANHO  -0.969444      1.5407754


NOTE: Covariances are above diagonal and correlations are below
diagonal.

Estimate, s.e. and z-score:

            Estimate   s.e.-Naive  s.e.-Robust  z-Robust
INTERCEPT -4.0971818      1.171        1.267      -3.23
TAMANHO    2.4388399      1.280        1.241       1.96

(c)  M. Rezaul Karim, 1989
Department of Biostatistics, The Johns Hopkins University


**********************************************************************
```

Variável : Tamanho (1=não reduziu; 0=reduziu)

The SAS System

Regression analysis using GEE:
==========

Data File: VARIZES

Outcome variable: SANGROU

Covariates: INTERCPT TAMANHO

Link: 3 (Logit)

Variance: 3 (Binomial)

Denominator _1_

Correlation: 5 (AR - 1)

Total number of records read: 90

Total number of clusters: 30

Maximum and minimum custer size 3 and 3

Averages of Outcome variable and Covariates (over all)

	SANGROU	INTERCPT	TAMANHO
Observations:	0.0777778	1	0.4666667
	0.0777778	1	0.4666667

Inicial estimate of regression coefficients:
INTERCEPT 0.0238095
TAMANHO 0.10119

===> Iteration: 1

	Estimate
INTERCEPT	-1.740726
TAMANHO	0.181687

===> Iteration: 2

```
                Estimate

INTERCEPT    -2.70489
TAMANHO       0.794367

===> Iteration:      3

                Estimate

INTERCEPT    -3.41146
TAMANHO       1.484322

===> Iteration:      4

                Estimate

INTERCEPT    -3.767735
TAMANHO       1.847594

===> Iteration:      5

                Estimate

INTERCEPT    -3.833911
TAMANHO       1.915893

===> Iteration:      6

                Estimate

INTERCEPT    -3.833714
TAMANHO       1.915972


Convergence after    6 iteration(s).

Working Correlation:
          1  -0.080263  0.0064421
-0.080263           1 -0.080263
 0.0064421 -0.080263          1

Scale parameter: 1.0458394


Mean Squared Error: 0.0691883


Variance estimate (naive):
```

```
           INTERCEPT      TAMANHO

INTERCEPT 1.1202822    -1.140933
TAMANHO  -0.92924       1.3456664

Variance estimate (robust):

           INTERCEPT   TAMANHO
INTERCEPT 1.1562797   -1.17899
TAMANHO  -0.954004     1.3208601


NOTE: Covariances are above diagonal and correlations are below
diagonal.

Estimate, s.e. and z-score:

            Estimate   s.e.-Naive  s.e.-Robust  z-Robust
INTERCEPT -3.833714       1.058        1.075      -5.56
TAMANHO    1.915972       1.160        1.149       1.67

(c)  M. Rezaul Karim, 1989
Department of Biostatistics, The Johns Hopkins University

**********************************************************************
```