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

Métodos Estatísticos para Análise de Dados Categorizados com Estruturas Complexas

Rosemeire Leovigildo Fiaccone

Profa. Dra. Eliana Heiser de Freitas Marques
Orientadora

Dez/98

Métodos Estatísticos para Análise de Dados Categorizados com Estruturas Complexas

Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por Rosemeire Leovigildo Fiaccone e aprovada pela comissão julgadora.

Campinas, 11 de dezembro de 1998

Profa. Dra. Eliana H. de Freitas Marques

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

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO IMECC DA UNICAMP

Fiaccone, Rosemeire Leovigildo

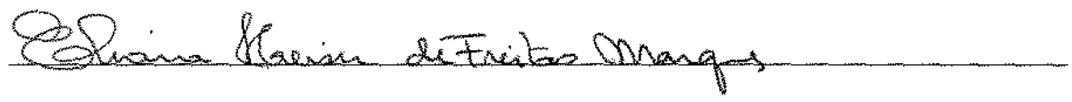
F44m Métodos estatísticos para análise de dados categorizados com estruturas complexas / Rosemeire Leovigildo Fiaccone -- Campinas, [S.P. :s.n.], 1998.

Orientador : Eliana Heiser de Freitas Marques

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

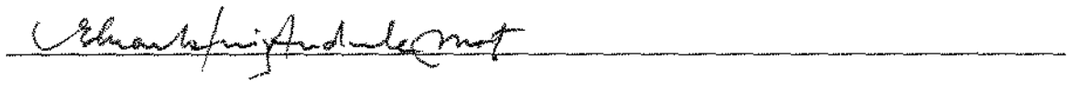
1. Estudos longitudinais. 2. Correlação (Estatística). I. Marques, Eliana Heiser de Freitas. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Dissertação de Mestrado defendida e aprovada em 11 de dezembro de 1998 pela Banca Examinadora composta pelos Profs. Drs.



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

Prof (a). Dr (a). GILBERTO ALVARENGA PAULA



Prof (a). Dr (a). EDUARDO LUIZ ANDRADE MOTA

"O caminho para o sucesso não é
fazer uma coisa 100% melhor, mas
100 coisas 1% melhor"
Júlio Lobos

AGRADECIMENTOS

Aos meus pais, Mário e Evany pelo apoio e incentivo.

Aos colegas e amigos do Departamento de Estatística da Ufba, pela compreensão e apoio.

À minha orientadora, Profa. Eliana H. Marques, pela orientação, dedicação, paciência, apoio e incentivo durante todo período de realização deste trabalho.

Um agradecimento muito especial a minha turma do Mestrado do ano de 1995, pelo convívio, amizade, apoio nas horas difícies e compartilhamento das horas de alegria.

A Fernando Lucambio e Rui Lyu pela disposição em me ajudar.

Aos novos amigos que fizeram parte do meu convívio em Campinas: Conceição, Desirê, Helena, Cinira, Daniela, Lusane, Rosi, Família Coniglo, Ritinha.

Aos colegas do Instituto de Saúde Coletiva e da Escola de Nutrição, em especial a Maurício Barreto e Ana Marluce pelo incentivo e colaboração.

Ao Prof. Dr. Luiz Roberto Moraes, pela confiança em colocar a minha disposição um dos conjuntos de dados usado nesta dissertação.

A Verônica, George, Carlos e Leila pela disposição em me ajudar.

A Jorge pelo carinho e compreensão.

Ao Prof. Dr. Jonh Preisser, pela confiança no envio de um programa.

A Profa. Dra. Andreas Ziegler pela colaboração com material bibliográfico.

Ao Prof. Dr. Vicent Carey com relação ao apoio a um dos programas utilizados nesta dissertação.

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

SUMÁRIO

RESUMO

Dados categorizados com estruturas complexas, resultantes de esquemas amostrais envolvendo conglomerados ou resultantes de respostas repetidas com as observações ocorrendo de forma agregada, têm sido frequente na literatura e têm gerado preocupações por parte dos pesquisadores, no que diz respeito aos métodos de estimação dos parâmetros de interesse. A realização deste trabalho tem por finalidade apresentar duas propostas avançadas: a metodologia da razão de médias provenientes de amostras complexas e as equações de estimação generalizadas para respostas correlacionadas, como novas alternativas para análise de dados não triviais. A motivação deste trabalho foi estudar essas novas ferramentas e no que diz respeito às aplicações, dar uma contribuição aos pesquisadores da área de saúde.

ABSTRACT

Categorical data with complex structures as a result of cluster sampling designs or repeated outcomes with observations occuring in some aggregated form, has been appearing recently in literature generating research and publications directed to methods of estimation of parameters, considering the possible correlation among the grouped observations. The purpose of dissertation is to present two advanced methodologies: a weighted regression method for analysis of multivariate categorical outcomes from cluster samples based on ratio means and the generalized estimating equations (GEE), extensions and diagnostics as new alternatives to analyse these non-standard data structures. The motivation for this study with respect to application was to contribute with new tools for research in the area of public health.

Capítulo I

Introdução

Pesquisas têm sido a fonte principal de informação para reflexões, decisões e ações em diferentes áreas do conhecimento. No campo social, por exemplo, o governo tem necessidade de possuir uma visão acurada da população em termos de localização, características pessoais, quantidade e qualidade de vida, com o propósito de formular sua política governamental de mudanças sociais. Já no campo da saúde, as pesquisas epidemiológicas visam o conhecimento sobre os determinantes do processo saúde/doença, propondo medidas específicas de prevenção, controle ou erradicação de doenças e fornecendo indicadores que sirvam de suporte ao planejamento, administração e avaliação das ações de saúde (Almeida, 1990). Neste contexto, tanto nessas áreas quanto em outras, a Estatística aparece como uma ferramenta básica para a análise dos dados levantados e consequentemente o conhecimento da área de estudo.

A variedade de métodos estatísticos relativamente novos, vem ao encontro das pesquisas metodológicas envolvendo dados com estruturas complexas que permitem levar

em consideração planejamentos complexos que incluem respostas multivariadas, respostas correlacionadas, processos de amostragem em um ou múltiplos estágios, etc. . Nas áreas de saúde, particularmente no cenário epidemiológico, nos anos mais recentes, têm, havido propostas de pesquisas usando a busca da melhor compreensão dos possíveis fatores que intervêm no processo saúde/doença, cujos planejamentos têm resultado em dados levantados, cuja estrutura é não trivial.

No âmbito da Estatística, técnicas inferenciais têm sido revisadas, aperfeiçoadas e estendidas 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. Essas estruturas complexas expressam o desenho do estudo ou a estratégia de coleta dos dados frequentemente empregada, podendo originar dependência entre as respostas para subconjuntos de observações com alguma característica comum de suas fontes.

Exemplos de situações de dependência entre as respostas aparecem tanto nos estudos longitudinais com medidas repetidas quanto nos estudos com coleta que refletem amostragem por conglomerado. Em ambos os casos, as observações formam um subconjunto com possível correlação intra-classe, podendo esses aglomerados de informações ser ou não tomados ao longo do tempo.

O problema central na analise de dados de conglomerado é considerar a dependência entre as sub-unidades do conglomerado. Como consequência tem havido uma evolução para métodos mais sofisticados de análise, envolvendo por exemplo, extensões de modelos lineares generalizados, correções de estatísticas baseadas no plano amostral, modelos hierárquicos, dentre outros (Rao & Scott, 1981, Wedderburn, 1974, Goldstein, 1987).

O objetivo original deste trabalho é apresentar duas propostas de análise de dados categorizados não triviais. Será dada ênfase a dados provenientes de amostras complexas, isto é, amostras que envolvem estratificação e/ou conglomerado, probabilidades distintas de seleção, como também propostas que levem em consideração direta ou indiretamente possíveis correlações em respostas categorizadas repetidas.

Este capítulo trata da relevância do processo de amostragem na estrutura dos dados, bem como a forma com que esses dados são apresentados, visando a escolha da técnica

estatística mais adequada para análise. Também neste capítulo é apresentada uma revisão bibliográfica sobre os temas em questão e uma sinopse sobre os dados que serão analisados na tese.

1.1 Considerações sobre dados categorizados

Segundo Koch *et al.* (1980), o planejamento da pesquisa é um dos determinantes mais importantes na estratégia de análise estatística, pois orienta a ligação entre o que é observado e como pode ser interpretado. A definição dos objetivos, do processo de amostragem, da mensuração dos dados, as suposições necessárias para a generalização dos resultados e a determinação do tipo de análise compõem, o corpo do planejamento de uma pesquisa.

As considerações importantes na determinação da análise apropriada são o processo de mensuração e amostragem dos dados, bem como o objetivo da pesquisa. O processo de mensuração diz respeito à forma pela qual os dados são expressos e a maneira pela qual são obtidos. A forma pela qual os dados são representados pode ser classificada de acordo com duas dimensões básicas : a escala de medida e a estrutura. A escala de mensuração da variável resposta expressa a natureza da informação disponível para a análise estatística. No caso de variáveis categorizadas, os dados incluem as classificações nominal, ordinal, enumeração discreta, entre outras.

No que diz respeito à estrutura da resposta de interesse, ela pode ser univariada ou multivariada.

A maneira pela qual os dados são coletados para levantamentos na área de saúde pode ser classificada de acordo com o modo de obtenção. Os questionários e entrevistas são os principais meios de coleta desses dados, que pode ser feita em uma única ocasião ou em duas ou mais ocasiões ou condições. Nesses tipos de coleta estão caracterizados, entre outros, os chamados estudos seccionais ou de corte-transversal e os estudos de medidas repetidas,

respectivamente. Uma classe importante dos estudos de medidas repetidas é chamada de estudos longitudinais, nos quais as medições dos dados são feitas em intervalos de tempo ou, mais geralmente, sob duas ou mais condições. Esta classe por ser o tema principal da tese será abordada no decorrer deste capítulo e também nos outros capítulos deste trabalho.

A outra importante consideração na determinação da análise é o processo de amostragem porque estabelece uma relação entre os dados e a população objetivo para a qual se deseja fazer inferência. Dados categorizados surgem de diferentes estruturas de amostragem. Em geral os dados se enquadram em uma das três estruturas de amostragem: dados históricos, dados experimentais e dados de inquéritos amostrais (Koch *et al.*, 1980).

Dados históricos são dados observacionais onde todos os elementos de uma certa população têm uma definição circunstancial, geográfica ou temporal. Por exemplo: inclusão de todas as ocorrências de uma doença infecciosa numa determinada área.

Dados experimentais são extraídos de estudos que envolvem alocação aleatória das unidades de investigação a tratamentos de interesse. Por exemplo: ensaios clínicos aleatorizados.

Dados de inquéritos amostrais envolvem seleção aleatória de unidades de investigação de uma grande população especificada. Por exemplo: pesquisas de opinião.

Pode existir uma combinação das duas formas anteriores, isto é, alocação aleatória de tratamentos a unidades de investigação selecionadas de uma amostra aleatória. A maior diferença nessas estruturas de amostragem é o uso da aleatorização para obtê-las. Dados históricos não envolvem aleatorização, por esse motivo é difícil assumir que eles sejam representativos de uma população conveniente.

Vale lembrar que a unidade de aleatorização pode ser simplesmente o indivíduo ou um conglomerado de indivíduos. Além disso, a aleatorização pode ser aplicada a subconjuntos, chamados estratos ou blocos, com probabilidades iguais ou não.

Ainda no processo de amostragem, o método de seleção mais comum é de amostragem aleatória simples, onde os indivíduos são escolhidos com igual probabilidade de seleção. Este método pode ser estendido para seleções de amostras separadas dentro de estratos pré-definidos. Neste caso as amostras são independentes umas das outras pelo fato de

que cada unidade amostral ocorre em um único estrato. Todo método de amostragem que afaste a propriedade de independência entre as observações ou altere a equiprobabilidade dos elementos que compõem a população de estudo é considerado um delineamento complexo.

Neste tipo de delineamento as unidades populacionais podem ser selecionadas envolvendo ambos, conglomerado e estratificação, bem como seleção em múltiplos estágios. A população pode ser estratificada dentro de vários subgrupos nos quais conglomerados de indivíduos devem ser escolhidos.

As estratégias de análise inferencial de dados categorizados podem envolver testes de hipótese ou a modelagem. Muitas questões em torno de dados categorizados podem ser respondidas pelo direcionamento de hipóteses específicas com referência à associação. Tais hipóteses frequentemente são investigadas com métodos de aleatorização. Também pode haver interesse em descrever a natureza da associação num conjunto de dados. Para isto, as técnicas de ajuste estatístico, utilizando estimação de máxima verossimilhança ou estimação de mínimos quadrados ponderados, são apropriadas para descrever esta variação em termos de um modelo estatístico parcimonioso.

Koch *et al.* (1975) e Freeman *et al.* (1976) adaptaram a metodologia de mínimos quadrados ponderados descrita por Grizzle, Starmer & Koch (1969) para analisar inquéritos amostrais complexos considerando o efeito do planejamento sobre a estatística analisada, nesse caso a razão, por se tratar de uma medida adequada em métodos de amostragem complexos. Para aplicação desta metodologia é necessário que os dados sejam arranjados em uma tabela de contingência de modo que haja um particionamento dos indivíduos de acordo aos níveis das variáveis explanatórias, fornecendo assim, uma estrutura de estratificação homogênea. Essas subdivisões devem ser identificadas como interseções dos níveis de diversas variáveis categorizadas. Se os valores da variável categorizada são conhecidos a priori e incluídos no planejamento amostral, então essas subdivisões são denominadas de estratos. Porém, em muitas populações complexas isto não é possível. Nesta situação as subdivisões são construídas após a amostra ter sido coletada recebendo a denominação de domínios (Freeman & Brock, 1977).

Como foi salientado anteriormente, a forma de obtenção dos dados é uma consideração importante pois identifica se a resposta de interesse foi observada em um único ou sucessivos pontos de tempo. Neste último, encaixam-se os estudos de medidas repetidas, os quais vêm recebendo bastante atenção devido, em grande parte, ao surgimento de pesquisas sobre métodos para tratar a dependência envolvendo respostas multivariadas categorizadas.

De um modo geral, a pesquisa longitudinal envolve observações de um conjunto de unidades de investigação classificadas em diferentes sub-populações segundo um ou mais fatores, ou tratamentos, ao longo de diversas condições de avaliação (como tempo, doses,...), que representam as unidades de observações (Singer & Andrade, 1986).

A análise de dados longitudinais apresenta algumas dificuldades. Por exemplo: a estrutura da dependência entre observações repetidas realizadas na mesma unidade de investigação. Um outro exemplo, a ocorrência de uma estrutura desbalanceada dos dados, resultante do não controle das circunstâncias em se obter as mensurações (Davis, 1993).

Um outro aspecto importante diz respeito à tomada de decisão com relação ao tipo de modelo, por exemplo, marginal ou condicional, que seja mais relevante para objetivo do estudo. A interpretação dos parâmetros será diferente conforme a escolha do modelo. Além disso, tanto a interpretação como os valores dos coeficientes do modelo a serem estimados são vinculados à natureza da estrutura de dependência das observações repetidas. Importante mencionar que, com respostas categorizadas, modelos não-lineares são comumente usados e os mesmos possuem uma estrutura na qual a resposta média não é separável da dependência entre observações repetidas como acontece em modelos lineares (Zeger, 1988). Assim, existem três distintas classes de modelos para análise de dados longitudinais: modelo marginal, condicional ou transicional e de efeitos aleatórios.

O modelo marginal descreve a distribuição da resposta média populacional em cada ocasião e a dependência dessas distribuições sobre as características das covariáveis. Os parâmetros no modelo marginal caracterizam a dependência da resposta média populacional sobre as covariáveis. O modelo transicional descreve a distribuição condicional de cada resposta como uma função explícita das respostas passadas e das covariáveis. Esse modelo combina as suposições a respeito da dependência da resposta sobre as covariáveis e da

correlação entre respostas repetidas em uma única equação. O modelo de efeitos aleatórios é muito útil quando o objetivo é produzir inferências em termos do indivíduo. Os parâmetros desse modelo descrevem como uma resposta esperada do indivíduo muda em função das mudanças nas suas covariáveis.

A base dos métodos clássicos de análise de dados longitudinais pertence a respostas contínuas e consiste de modelos paramétricos que assumem uma estrutura de erro normal multivariada. Koch *et al.* (1977) foram os primeiros a desenvolver um procedimento geral para analisar respostas repetidas categorizadas baseado na metodologia de mínimos quadrados ponderados de Grizzle, Starmer & Koch (1969), através da especificação de um modelo marginal. Esta metodologia pede a estratificação da amostra dentro de subgrupos que são homogêneos com respeito aos valores das covariáveis. Isto é, pelo fato da escala de mensuração ser categorizada, a formulação conceitual pode ser visualizada dentro do contexto de uma tabela de contingência (s x r), onde s são as sub-populações determinadas pela classificação cruzada de fatores de interesse e r são os perfis da resposta multivariada obtidas da classificação cruzada completa das variáveis respostas sobre o tempo. Duas limitações dessa metodologia são: a não inclusão de variáveis explanatórias contínuas e a exigência de tamanho suficientemente grande para as sub-populações.

O procedimento de equações de estimação generalizadas (Liang & Zeger 1986; Zeger & Liang 1986) é uma metodologia recente para análise de regressão de medidas repetidas que pode usar variáveis explanatórias contínuas ou discretas. É um método semi-paramétrico pois as equações de estimação foram deduzidas sem a especificação completa da distribuição conjunta das observações, entretanto inclui a especificação de uma estrutura de correlação de trabalho. O vetor multivariado de respostas repetidas pode ser discreto ou contínuo.

Quando a resposta de interesse é um vetor multivariado binário e o objetivo inclui não somente a descrição da dependência de cada resposta binária sobre as variáveis explanatórias como também a caracterização do grau de associação entre essas respostas, pode-se utilizar uma outra metodologia recente, Regressão Logística Alternada (Carey *et al.*, 1993). Ainda muito pouco explorada, essa metodologia é um caminho alternativo às equações

de estimação generalizadas de primeira e segunda ordem quando o tamanho dos conglomerados (número de respostas repetidas por indivíduo) torna-se muito grande.

1.2 Revisão da literatura

Dados categorizados com estruturas complexas, resultantes de esquemas amostrais envolvendo conglomerados em um ou mais estágios, têm sido frequentes na literatura e têm gerado preocupações por parte dos pesquisadores no que diz respeito aos métodos de estimação dos parâmetros de interesse.

Rao & Scott (1981, 1984) propuseram um método para corrigir a estatística qui-quadrado padrão em estudos com esquema amostral complexo, estimando pesos como função do efeito do delineamento amostral e usando os mesmos para corrigir esta estatística. Eles mostraram que a distribuição assintótica da estatística qui-quadrado é uma soma ponderada de variáveis aleatórias qui-quadrado independentes, onde os pesos são funções do efeito do delineamento amostral. Brier (1980) apresentou uma modificação simples na estatística qui-quadrado da razão de verossimilhança e na de Pearson para ajustar dados de uma tabela de contingência obtida de uma amostragem de conglomerado. Binder (1983) propôs um método de estimação assintótico da matriz de covariância dos parâmetros de regressão dentro da classe de modelos lineares generalizados para amostras de uma população finita de acordo ao delineamento amostral complexo, utilizando linearização em série de Taylor. Outra estratégia, já citada anteriormente, é a metodologia da razão. Koch *et al.* (1975), Freeman *et al.* (1976), Landis *et al.* (1987) são exemplos de alguns trabalhos que utilizam esse método juntamente com a metodologia de mínimos quadrados ponderados.

A análise de dados categorizados com estrutura complexa originada de medidas repetidas é também uma outra área de pesquisa muita ativa e novos desenvolvimentos têm surgido rapidamente. Em 1977, Koch *et al.* adaptaram a metodologia GSK, originada por Grizzle, Starmer e Koch (1969), a experimentos com medidas repetidas. Neste cenário surgiram trabalhos para análise de dados longitudinais categorizados como: Stanish *et al.*

(1978), Koch *et al.* (1985), Landis *et al.* (1988), Koch *et al.* (1989), Koch *et al.* (1992), dentre outros, baseando-se na metodologia de mínimos quadrados ponderados. Afora esta metodologia aparecem outros trabalhos, como por exemplo, Rosner (1984, 1989) que apresentou um modelo de regressão logística politômica para controlar o efeito do conglomerado e de covariáveis específicas quando existe correlação entre as unidades dentro do conglomerado. Donner & Donald (1988) e Donner (1989) propuseram um ajustamento na estatística qui-quadrado para o teste de homogeneidade de proporções entre grupos de indivíduos quando as observações correlacionadas ou múltiplas são feitas sobre cada indivíduo. Já Connoly & Liang (1988) sugeriram um procedimento baseado na classe de modelos de regressão logística condicional para dados binários correlacionados. A partir do final da década de 80 começaram a se intensificar na literatura artigos que utilizavam procedimentos semi-paramétricos na análise de dados longitudinais.

Em 1986, surgiu o método das equações de estimação generalizadas (EEG) de Liang & Zeger. As EEG são uma extensão das equações de estimação de modelos lineares generalizados para respostas multivariadas. É um método semi-paramétrico pois as equações de estimação são deduzidas sem a especificação completa da distribuição conjunta do vetor de resposta multivariado, necessitando apenas de suposições sobre o comportamento dos parâmetros de interesse e sobre a estrutura de correlação. Já Wei & Stram (1988) modelaram a distribuição marginal da resposta em cada tempo usando a classe de modelos lineares generalizados, obtendo assim coeficientes de regressão específicos em cada ponto de tempo. Segundo Zeger (1988), quando as covariáveis são dependentes do tempo, os métodos EEG e Wei e Stram apresentam estimativas dos coeficientes idênticas, usando uma estrutura de correlação de independência para as EEG.

Stram, Wei & Ware (1988) desenvolveram modelos marginais com respostas ordinais repetidas, ajustando regressões separadas em cada tempo. Essa técnica pode ser considerada como um método semi-paramétrico para o modelo do logito cumulativo de respostas longitudinais e como caso especial de independência das EEG.

Prentice (1988) estendeu o método das EEG para dados binários correlacionados com a formulação de um segundo conjunto de equações de estimação, com o objetivo de

estimar também o parâmetro de associação, no caso a correlação. Zhao & Prentice (1990) identificaram a classe de modelos exponenciais quadráticos para dados binários correlacionados, onde a função escore das equações de estimação é a máxima verossimilhança, introduzindo a extensão das EEG de segunda ordem. Posteriormente, Prentice & Zhao (1991) estenderam a estimação dos parâmetros da média e covariância a um vetor geral de respostas multivariadas.

Lipstiz, Laird & Harrington (1991) modificaram as equações de estimação de Prentice (1988) para permitir modelos de associação entre medidas repetidas via o uso da razão de chances. Em 1992, Liang Zeger & Qaqish nomearam as EEG de Liang & Zeger (1986) de EEG1 (equações de estimação generalizada de primeira ordem) e a extensão apresentada por Zhao & Prentice (1990) de EEG2, esta última é usada quando se deseja estimar também a correlação existente entre as medidas repetidas.

Fitzmaurice *et al.* (1993) propuseram um método no qual a verossimilhança completa é especificada com base na representação log-linear geral. Eles estudaram um modelo misto no qual os parâmetros de regressão descrevem a média marginal, porém, a associação é medida em termos da razão de chances condicionada a outras respostas. Contudo, a aplicação deste método é limitada a estudos onde o número de observações por indivíduo é igual. Carey *et al.* (1993) formularam o modelo de associação em termos da razão de chances marginal, denominado Regressão Logística Alternada, evitando assim alguns problemas de restrições associados com correlações em dados binários além da fácil interpretabilidade desta medida perante a razão de chances condicional. Uma outra aplicação desta metodologia é encontrada em Katz *et al.* (1993), onde estimou-se o grau de associação da diarréia em diferentes ambientes e inquéritos amostrais, com o objetivo de estimar o efeito do delineamento amostral e o grau de ocorrência da diarréia em casas e vilas habitadas por crianças na idade pré-escolar. Fitzmaurice (1995) apresentou um modelo para dados de série de tempo binário no qual as respostas repetidas sobre cada indivíduo podem ser desigualmente espaçadas no tempo. Este procedimento modela a associação entre respostas binárias usando padrões de razão de chances exponencial, isto é, análogo aos métodos comumente usados para dados contínuos de série de tempo. O autor também utilizou a metodologia de Regressão Logística Alternada.

Heagerty & Zeger (1996) propuseram equações de estimação para analisar dados categorizados ordinais correlacionados através de dois modelos de regressão: modelo de odds proporcional para média marginal e um modelo logístico para a razão de chances marginal descrevendo associação entre pares de respostas.

Ainda muito pouco explorado, o diagnóstico nas equações de estimação generalizadas começa a surgir na literatura, a exemplo, Preisser & Qaqish (1996), Ziegler & Arminger (1996), Ziegler *et al.* (no prelo), com objetivo de medir a influência de um subconjunto de observações sobre os parâmetros da regressão estimada e sobre os valores estimados do preditor linear.

A intenção neste trabalho não é comparar as metodologias existentes na análise de dados categorizados com estruturas complexas, seja no âmbito de medidas correlacionadas ou provenientes de esquemas amostrais complexos, e sim explorar as especificidades de duas destas metodologias avançadas da forma mais abrangente possível, que são: a metodologia da razão de médias provenientes de amostras complexas e as equações de estimação generalizadas (EEG) para respostas correlacionadas, tentando cobrir nos exemplos diferentes aspectos levantados pelos dados. Será explorada também, de maneira modesta, a metodologia de regressão logística alternada e a parte de diagnóstico nas EEG. É de interesse também, no que diz respeito às aplicações, dar uma contribuição aos pesquisadores da área de saúde no sentido de obter uma melhor visão dos fatores de risco associados às diferentes enfermidades.

O capítulo II aborda o método de regressão ponderada para análise de conglomerados grandes de dados binários e discretos de amostras extraídas pelo processo de conglomerados a um e dois estágios, baseando-se na razão de médias e utilizando a metodologia de mínimos quadrados ponderados para modelar essa razão de médias.

O capítulo III mostra um resumo da teoria das EEG e possíveis extensões na análise de dados categorizados correlacionados, além de uma breve explanação de diagnóstico nas EEG.

O capítulo IV apresenta aplicações das técnicas abordadas nos capítulos II e III, utilizando os dados descritos na próxima seção deste capítulo, com programas computacionais realizado pelos próprios pesquisadores e pelos softwares já disponíveis no mercado.

Como o objetivo deste trabalho é explorar as metodologias apresentadas de uma forma abrangente e também contribuir para um maior subsídio aos pesquisadores da área de saúde, as respostas de interesse utilizadas nas diferentes análises são de caracter epidemiológico.

1.3 Descrição dos dados

Dois conjuntos de dados serão analisados neste trabalho. O primeiro conjunto refere-se a um ensaio clínico aleatorizado, duplo-cego, placebo-controlado realizado pelo Instituto de Saúde Coletiva da Universidade Federal da Bahia, no período de dezembro de 1990 a dezembro de 1991, com o objetivo de avaliar o efeito da suplementação periódica de vitamina A sobre a morbidade e mortalidade em crianças menores de 5 anos - Estudo de Serrinha. O segundo conjunto refere-se a um projeto realizado pelo Departamento de Hidráulica e Saneamento da Universidade Federal da Bahia, no período de agosto de 1989 a novembro de 1990, com o objetivo de avaliar o impacto das ações de saneamento, em particular um sistema de coleta e transporte dos esgotos, na saúde da população da periferia de Salvador - AISAM.

Com o propósito de uma maior interação entre as técnicas estatísticas descritas e a epidemiologia, bem como uma melhor compreensão do tema abordado nos estudos a serem analisados nesse trabalho, são feitas algumas considerações epidemiológicas acerca das doenças diarréicas.

Dado que o processo saúde-doença se insere na complexidade dos fenômenos sociais, seu status está diretamente relacionado às condições ambientais domiciliares. As políticas públicas voltadas à melhoria e/ou ampliação da infra-estrutura urbana, com destaque para o sistema de abastecimento de água, coleta, acondicionamento e destinação dos dejetos líquidos e sólidos, tem implicações diretas sobre o processo de circulação de determinados agentes (patógenos) causadores de doenças dependentes do meio hídrico para desenvolver o ciclo de transmissão. Várias enfermidades associam-se à deficiência e/ou ausência de saneamento,

destacando-se no conjunto das doenças as diarréias infantis, as quais têm merecido a atenção de sanitaristas no mundo inteiro.

A importância de estudos que enfoquem a associação entre indicadores sócio-ambientais e de saúde através de indicadores de morbidade e mortalidade principalmente para doenças do grupo das infecto-contagiosas, revela-se na crescente produção de pesquisas, marcadamente nos países em desenvolvimento. Embora haja consenso sobre a importância dos indicadores sócio-ambientais, há que se levar em conta as limitações, divergências conceituais e metodológicas que caracterizam esses estudos.

Dentre inúmeras pesquisas, Costa *et al.* (1980), estudando o padrão de mortalidade das crianças na faixa etária de 7-14 anos em Salvador, observaram que as principais causas de morte foram atribuídas às diarréias. Concluíram que a maioria das mortes poderiam ter sido evitadas por medidas simples, envolvendo cuidados primários à saúde, saneamento e vacinação.

Segundo Moraes (1996), diversas doenças estão relacionadas ao saneamento inadequado. O impacto da melhoria de uma intervenção de saneamento sobre a saúde infantil tem sido estudado, embora poucos estudos tenham sido conduzidos em áreas urbanas. A incidência de diarréia, mortalidade, prevalência de infecção intestinal por nematóides e, mais recentemente, o estado nutricional têm sido utilizados como indicadores de saúde para avaliar o impacto da melhoria no saneamento. Evidenciando o papel do saneamento, o autor referido realizou um estudo de base longitudinal em Salvador, comparando três grupos de comunidades com diferentes condições de saneamento. Nesta pesquisa observou-se que a incidência de diarréia infantil foi significativamente menor no grupo residindo em área saneada do que entre os residentes em área desprovida de serviço de esgotamento sanitário.

Além de verificar aspectos ambientais sobre a ocorrência de diarréias infantis, algumas pesquisas têm sido realizadas com o objetivo de conhecer o papel da suplementação de vitamina A sobre esta doença. A redução expressiva na mortalidade face a suplementação com vitamina A, detectada por Sommer *et al.* (1986) entusiasmou alguns estudiosos e organizações internacionais de saúde, fomentando, a partir da década de 80, o desenvolvimento de estudos de intervenção para validar essas descobertas e explicar o

mecanismo da redução da mortalidade, especialmente por diarréia e infecção respiratória. Apesar disso, o estudo mencionado acima foi alvo de discussões por se tratar de um estudo não aleatorizado, não cego e não placebo controlado.

Um outro estudo de suplementação aleatorizado, duplo cego e placebo controlado, foi desenvolvido por West et al (1991) com crianças de Nepal (Sul da Ásia) de 6 a 72 meses de idade. A redução na taxa de mortalidade observada nesse estudo foi de 30% para a diarréia e disenteria no grupo suplementado. Outro estudo é o de Ghana (1993), que tem a peculiaridade de englobar áreas adjacentes em dois estudos concomitantes, com metodologias e objetivos diferentes. O estudo de sobrevivência avaliou o impacto do suplemento com vitamina A sobre a mortalidade de crianças de 6 a 90 meses de idade. A redução na mortalidade, em um período de 26 meses, para as crianças do grupo suplementado foi da ordem de 19% quando comparada com aquelas do grupo controle. Vale ressaltar que este estudo não encontrou diferença na prevalência média e na duração da diarréia, nem na prevalência média de sarampo e de sintomas relacionados à infecção respiratória em crianças que receberam o suplemento que pudesse explicar a redução encontrada.

O efeito protetor da suplementação com vitamina A na redução da morbidade infantil foi também verificado por Barreto *et al.* (1994). Esse estudo detectou que o suplemento mostrou maior impacto na redução da incidência dos episódios severos de diarréia para crianças do grupo suplementado, quando comparado com aquele verificado nas crianças do grupo placebo. Cabe comentar que os dois últimos estudos mencionados foram indicados como os melhores, juntamente com mais dois, no relatório de um encontro sobre vitamina A (Bellagio Meeting on Vitamin A Deficiency & Childhood Mortality, 1993).

1.3.1 Estudo de Serrinha

O estudo foi realizado na cidade de Serrinha, a 170Km noroeste de Salvador, Bahia. É uma cidade situada na zona do semi-árido, possuindo cerca de 30.000 habitantes e caracterizada por apresentar clima quente e seco, além de chuvas irregulares. Os serviços de saúde de Serrinha são deficientes e aquém das necessidades de sua população.

O desenho do estudo é do tipo longitudinal formado por uma coorte fixa, com o acompanhamento de 1240 crianças de 6 a 48 meses, com o objetivo de testar o efeito da suplementação de vitamina A sobre a diarréia e a infecção respiratória aguda. As crianças foram aleatorizadas e receberam vitamina A ou placebo a cada 4 meses por um período de um ano. Elas foram visitadas três vezes por semana nos seus lares por entrevistadores que coletaram dados a respeito da ocorrência de diarréia, bem como o número de dejeções líquidas e amolecidas por períodos de 24 horas e também informações sobre infecção respiratória. No caso de haver 3 ou mais dejeções líquidas/amolecidas uma investigação mais detalhada acerca de sinais de vômitos, presença de muco ou sangue nas fezes, febre, uso de medicamento, uso de reidratação oral, internação hospitalar, foi conduzida. No caso de ter havido relato de tosse, a frequência respiratória foi medida duas vezes. Se a criança apresentava um número médio superior a 40 bat./min ou se fosse observado chiado no peito, o caso era relatado e o pediatra do projeto investigava o episódio mais profundamente (Barreto *et al.*, 1994).

No início do estudo as crianças foram selecionadas de acordo com os seguintes critérios: idade entre 6 a 48 meses; consentimento dos pais, não existência de xeroftalmia[1] ativa, não ocorrência de sarampo nos últimos 30 dias, e não terem recebido alta dose de suplementação de vitamina A nos últimos 6 meses ou, ainda crianças, com peso não inferior a 60% daquele estabelecido pelo padrão do "National Center for Health Statistics" para cada idade. Também foram coletadas informações sócio-econômicas da família da criança.

O instrumento utilizado na pesquisa foi um questionário pré-testado para avaliar a sua consistência. O procedimento envolveu entrevistas domiciliares realizadas por entrevistadores de campo supervisionados, responsáveis cada um por 60 crianças (30

[1] Inflamação da córnea

visitas/dia), que utilizaram a técnica de três visitas semanais, onde foram coletados dados referentes à diarréia e à infecção respiratória.

Definiu-se como diarréia técnica o registro de três ou mais dejeções líquidas e/ou amolecidas em um período de 24 horas, e delimitou-se como um novo episódio de diarréia o intervalo de três ou mais dias sem diarréia. O intervalo de tempo estabelecido encaixa-se nas recomendações sugeridas em outros estudos, (Morris *et al.*, 1994) e (Baqui *et al.*,1991)

As análises que serão apresentadas neste trabalho utilizarão somente uma parte dos dados coletados para este estudo.

1.3.2 Estudo AISAM

O projeto AISAM - Avaliação do Impacto das Medidas de Saneamento Ambiental em Áreas Pauperizadas de Salvador - estuda os efeitos dos fatores ambientais, particularmente soluções de baixo custo para o transporte de excretas/esgotos sanitários, nas doenças diarréicas, infecções por nematóides e estado nutricional.

O estudo foi conduzido em áreas urbanas pauperizadas da periferia de Salvador, precisamente na Bacia do Rio Camurujipe, no período de agosto 1989 à novembro de 1990. Esta bacia atinge um total de 39 quilômetros quadrados, habitados por uma população de cerca de 800 mil pessoas de baixa renda, distribuídas em 34 agrupamentos ou comunidades. O Rio Camurujipe é o maior (15km de extensão) e o mais importante coletor de Salvador, para onde afluem tanto os excessos de chuvas quanto os afluentes de águas servidas, domésticas e industriais.

Segundo Moraes (1996), a metodologia utilizada para o estudo estratificou a área em três grupos de acordo com o tipo de intervenção de saneamento: comunidades que não tiveram nenhum tipo de medida adotada para o destino dos dejetos (Grupo 1 - Controle), outro cuja solução empregada para o esgotamento sanitário foi um sistema composto de rampas e escadarias drenantes[2] (Grupo 2) e o terceiro que, além destas, conta com uma rede

[2]As rampas e escadarias drenantes, com interior oco, funcionam como escoamento das águas de chuvas, circulação de pedestres e, neste caso, como solução de esgotamento sanitário.

coletora específica para os esgotos sanitários (Grupo 3). Três comunidades em cada grupo foram selecionadas ao acaso de uma lista de todas as comunidades, resultando assim num total de 9 comunidades. Em cada comunidade cerca de 120 casas foram selecionadas, ao acaso, de uma lista de todas as casas, para alcançar o tamanho da amostra proposto (130 crianças abaixo de 5 anos e 210 entre 5 a 14 anos para cada comunidade).

O desenho do estudo é do tipo longitudinal formado por uma coorte de 1162 crianças menores de 5 anos e pelo acompanhamento também de 1893 crianças de 5 a 14 anos.
Os instrumentos utilizados na pesquisa foram questionários pré-testados para levantar informações de saúde, demográficas, sociais, econômicas, físicas e antropológicas, aplicados por entrevistadores de campo supervisionados.

Em particular, para o estudo da morbidade de diarréia, todas as crianças menores de 5 anos com perda da consistência usual das fezes e aumento da frequência de evacuações foram notificadas pelas mães ou guardiãs da criança. Para tal foi utilizado um questionário com registro diário através de um calendário quinzenal com a fotografia da criança, onde as mães eram estimuladas a marcar diariamente com o sinal "+" ou "-" se cada uma das suas crianças apresentara ou não diarréia naquele dia. Também foram levantadas informações da causa e dos sintomas de diarréia e tratamento aplicado. Durante cada período de 2 semanas, pesquisadores de campo visitavam duas vezes a casa da criança para entrevistar e verificar se a mãe estava usando o calendário. Reuniões com os líderes e as mães de cada comunidade foram realizadas antes da coleta dos dados, com o objetivo de expor a importância do estudo, além de padronizar a percepção das mães quanto aos sintomas de diarréia.

Um episódio de diarréia foi definido como um ou mais dias com diarréia separado de qualquer outro episódio por pelo menos 2 dias livres do sintoma de diarréia.

Do ponto de vista estatístico, essas duas bases de dados enquadram-se perfeitamente na análise de dados com estruturas complexas. Seja na questão da maneira pela qual os dados são obtidos ou pelo processo de amostragem desses dados. Será dada ênfase a questão da medida repetida na pesquisa longitudinal, como também a questão de dados gerados por pesquisas envolvendo planos amostrais complexos.

A seguir será apresentada a metodologia de mínimos quadrados ponderados para medidas do tipo razão.

Capítulo II

Metodologia de Mínimos Quadrados Ponderados para Medidas do Tipo Razão

2.1 Introdução

Grizzle, Starmer e Koch, em 1969, propuseram um método alternativo ao de máxima verossimilhança para análise de dados categorizados com base na teoria de modelos lineares e mínimos quadrados ponderados, hoje conhecido como método GSK. O objetivo da análise, descrito resumidamente, é a modelagem de tabelas de contingência multi-dimensionais geradas a partir de classificações cruzadas de variáveis qualitativas, juntamente com testes de hipóteses apropriados.

O método de mínimos quadrados ponderados propõe uma metodologia bastante ampla na modelagem de dados categorizados. As estimativas obtidas podem ser um vetor de proporções, escores médios ou outras funções mais complicadas dos dados. A escolha da função de resposta baseia-se em alguns critérios, tais como: objetivo da pesquisa, facilidade computacional na estimação dos parâmetros e busca do melhor ajuste para o modelo linear.

A idéia geral é modelar a distribuição da variável resposta (representada nas colunas de uma tabela de contingência), entre os níveis das variáveis explanatórias (representada pelas linhas da tabela), sob uma estrutura de amostragem aleatória estratificada. Esta metodologia pode ser facilmente adaptada não somente para estruturas de amostragem mais complexas como também para lidar com variáveis de respostas múltiplas, isto é, medidas repetidas.

Koch *et al.* (1977) descrevem a aplicação da metodologia de mínimos quadrados ponderados para medidas repetidas de dados categorizados. Em aplicações desse tipo o interesse geralmente detém-se na análise da distribuição marginal da resposta em cada ponto de tempo ou condição. Nesse caso, vão existir múltiplas funções por grupo e a estrutura de correlação induzida pelas medidas repetidas deve ser levada em consideração. A estrutura de covariância baseada na distribuição multinomial é uma candidata natural para lidar com a correlação das medidas repetidas.

Quando um inquérito amostral envolve uma estrutura complexa de seleção de unidades amostrais em dois ou mais estágios é necessário que os métodos estatísticos para analisar tais dados incorporem essa estrutura de amostragem. Koch *et al.* (1975) e Freeman *et al.* (1976) adaptaram a metodologia de mínimos quadrados ponderados para analisar dados multivariados com estrutura complexa considerando o efeito do planejamento amostral sobre a estatística de interesse. Já Landis *et al.* (1987) usaram esta mesma metodologia para modelar logitos cumulativos com planejamento amostral complexo.

A metodologia de mínimos quadrados ponderados é baseada no modelo $E_A(\underset{\sim}{F}) = X\underset{\sim}{\beta}$ onde X é a matriz de planejamento, $\underset{\sim}{\beta}$ é o vetor de parâmetros de regressão e F é uma função de interesse. Para amostras complexas, F pode ser um vetor de estimativas do tipo razão, que são funções dos estimadores de Horvitz-Thompson para totais populacionais (Davies, 1994).

Nas seções seguintes, serão abordadas a metodologia básica do método GSK, a título de revisão, e extensões.

2.2 Análise em tabela de contingência

O conjunto de observações de dados categorizados pode ser resumido numa tabela de contingência, que é uma representação resultante da classificação cruzada de duas ou mais variáveis categorizadas.

Suponha que existem s sub-populações indexadas por $i=1,2,...,s$ das quais se extraem amostras independentes de tamanho n_i e seja $j=1,2,...,r$ o índice que representa os níveis ou categorias da variável resposta ou dependente em cada sub-população.

O esquema descrito acima pode ser resumido em uma tabela de contingência sxr :

Tabela 1: Forma bidimensional de uma tabela de contingência genérica

Sub-população	Níveis de Resposta					Total
	1	2	3	.....	r	
1	y_{11}	y_{12}	y_{13}	.....	y_{1r}	$n_{1.}$
2	y_{21}	y_{22}	y_{23}	.....	y_{2r}	$n_{2.}$
3	y_{31}	y_{32}	y_{33}	.....	y_{3r}	$n_{3.}$
⋮	⋮	⋮	⋮		⋮	⋮
s	y_{s1}	y_{s2}	y_{s3}	.....	y_{sr}	$n_{s.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	.....	$n_{.4}$	$n_{..}$

As principais distribuições utilizadas na modelagem probabilística de tabelas de contingência são a distribuição multinomial e a de Poisson. No caso da distribuição de Poisson as caselas são independentes que diferem da situação da multinomial, porém esses modelos probabilísticos estão intimamente associados entre si (Breslow & Day, 1987). Os parâmetros que indexam essas distribuições possuem estimadores consistentes, não viciados e assintoticamente normais. Essas propriedades assintóticas propiciam a utilização do método delta no cálculo da distribuição assintótica de funções particulares desses estimadores.

Considere o conjunto de dados categorizados apresentado na tabela acima. Os totais marginais $n_{1.}$, $n_{2.}$, ..., $n_{s.}$ constituem os tamanhos de amostra em cada sub-população e as

variáveis aleatórias y_{ij} em cada casela representam o número de indivíduos na amostra correspondente à sub-população *i* que apresentam a resposta *j*. Essas amostras são conceitualmente representativas de sub-populações infinitas e as tendências de cada indivíduo em apresentar a *j*-ésima resposta são consideradas mutuamente independentes.

Considerando válidas as afirmações acima, o vetor aleatório $\underset{\sim}{Y}_i = (y_{i1}, y_{i2}, \ldots, y_{ir})'$ tem distribuição multinomial com parâmetros $n_{i.}$ e $\underset{\sim}{\pi}_i = (\pi_{i1}, \ldots, \pi_{ir})'$ onde π_{ij} é a probabilidade de um indivíduo selecionado ao acaso da *i*-ésima sub-população apresentar a *j*-ésima categoria da resposta. A função de probabilidade de $\underset{\sim}{Y}_i$ é

$$P(Y_{i1} = y_{i1}, \ldots, Y_{ir} = y_{ir}) = (n_i)! \prod_{j=1}^{n} \frac{\pi_{ij}^{y_{ij}}}{(y_{ij})!} \qquad (2.1)$$

com $\sum_{j=1}^{r} y_{ij} = n_{i.}$ e $\sum_{j=1}^{r} \pi_{ij} = 1$, $\pi_{ij} \in (0,1)$ para todo *i*=1,2,..., *s* e *j*=1,2,..., *r*.

O estimador não viciado para o parâmetro π_{ij} é a proporção amostral $p_{ij} = \frac{y_{ij}}{n_{i.}}$.

Sendo assim, a estimativa do vetor $\underset{\sim}{\pi}_i$ é $\underset{\sim}{p}_i = (p_{i1}, p_{i2}, \ldots, p_{ir})' = \left(\frac{y_{i1}}{n_{i.}}, \frac{y_{i2}}{n_{i2}}, \ldots, \frac{y_{ir}}{n_{ir}} \right)'$ para a *i*-ésima sub-população, *i*=1,2,..., *s* e os elementos da matriz de variância-covariância são:

$$Var(p_{ij}) = \frac{\pi_{ij}(1-\pi_{ij})}{n_{i.}}, \quad \text{cuja estimativa é} \quad v(p_{ij}) = \frac{p_{ij}(1-p_{ij})}{n_{i.}} \qquad (2.2)$$

$$Cov(p_{ij}, p_{ij'}) = \frac{-\pi_{ij}\pi_{ij'}}{n_{i.}} \text{ para } j \neq j' \text{ e} \qquad (2.3)$$

$$Cov(p_{ij}, p_{i'j}) = 0 \text{ para } i \neq i' \text{ (sub-populações independentes).} \qquad (2.4)$$

Portanto o vetor de parâmetros das s sub-populações é denotado por $\underset{\sim}{\pi}=(\pi_1',\ldots,\pi_s')'$ e a estimativa da proporção amostral é $\underset{\sim}{p}=(p_1',p_2',\ldots,p_s')'$.

Então,

$$E(\underset{\sim}{p})=\underset{\sim}{\pi} \quad \text{e} \quad V(\underset{\sim}{p})=\begin{bmatrix} \underset{\sim}{V}_1(\underset{\sim}{\pi}_1) & 0_{rxr} & \ldots\ldots & 0_{rxr} & 0_{rxr} \\ 0_{rxr} & \underset{\sim}{V}_2(\underset{\sim}{\pi}_2) & \ldots\ldots & 0_{rxr} & 0_{rxr} \\ \vdots & \vdots & & \vdots & \vdots \\ \vdots & \vdots & & \vdots & \vdots \\ 0_{rxr} & 0_{rxr} & \ldots\ldots & 0_{rxr} & \underset{\sim}{V}_s(\underset{\sim}{\pi}_s) \end{bmatrix}=\underset{\sim}{v}(\underset{\sim}{\pi}) \qquad (2.5)$$

é uma matriz bloco diagonal com elementos $\underset{\sim}{V}_i(\underset{\sim}{\pi}_i)_{rxr}$ na diagonal principal para i=1,2,...,s, onde cada

$$\underset{\sim}{V}_i(\underset{\sim}{\pi}_i)_{rxr}=\begin{bmatrix} \pi_{i1}(1-\pi_{i1}) & -\pi_{i1}\pi_{i2} & \ldots\ldots & -\pi_{i1}\pi_{ir} \\ -\pi_{i2}\pi_{i1} & \pi_{i2}(1-\pi_{i2}) & \ldots\ldots & -\pi_{i2}\pi_{ir} \\ \vdots & \vdots & & \vdots \\ -\pi_{ir}\pi_{i1} & -\pi_{ir}\pi_{i2} & \ldots\ldots & \pi_{ir}(1-\pi_{ir}) \end{bmatrix}$$

é a matriz de variância-covariância da i-ésima sub-população.

Em notação matricial escreve-se

$$\underset{\sim}{V}_i(\underset{\sim}{\pi}_i)=\left\{\left(\underset{\sim}{D}_{\underset{\sim}{\pi}_i}-\pi_i\pi_{i'}\right)/\underset{\sim}{n}_i\right\}, \qquad (2.6)$$

onde $\underset{\sim}{D}_{\underset{\sim}{\pi}_i}$ é uma matriz diagonal com os elementos $\underset{\sim}{\pi}_i$ na diagonal.

Uma vez obtidas as estimativas das funções desejadas, $F_1(\underset{\sim}{p}),F_2(\underset{\sim}{p}),\ldots,F_u(\underset{\sim}{p})$, um estimador consistente da matriz de variância-covariância de $\underset{\sim}{F}$ é a matriz uxu de forma

$$\underset{\sim}{v}_{\underset{\sim}{F}}=\underset{\sim}{H}\underset{\sim}{V}(\underset{\sim}{p})\underset{\sim}{H}', \quad \text{onde } \mathrm{H}_{uxrs}=\left.\frac{\partial \underset{\sim}{\mathrm{F}}(\underset{\sim}{\pi})}{\partial \underset{\sim}{\pi}}\right|_{\underset{\sim}{\pi}=\underset{\sim}{\mathrm{p}}} \qquad (2.7)$$

é a matriz das derivadas parciais de 1ª ordem de funções F calculadas em $\underset{\sim}{p}$.

Quando as funções $\underset{\sim}{F}(\underset{\sim}{p})$ são não-lineares em $\underset{\sim}{p}$, usa-se o método de linearização em série de Taylor, no qual se decompõe a função $\underset{\sim}{F}(\underset{\sim}{\pi})$ em torno de $\underset{\sim}{p}$ até o termo de 1ª ordem, ou seja,

$$\underset{\sim}{F}(\underset{\sim}{\pi}) = \underset{\sim}{F}(\underset{\sim}{p}) + (\underset{\sim}{\pi} - \underset{\sim}{p}) \left. \frac{\partial \underset{\sim}{F}(\underset{\sim}{\pi})}{\partial \underset{\sim}{\pi}} \right|_{\underset{\sim}{\pi}=\underset{\sim}{p}} + O(|\underset{\sim}{\pi} - \underset{\sim}{p}|) \quad , \text{ onde } \quad O(|\underset{\sim}{\pi} - \underset{\sim}{p}|) \to \infty, \text{ quando}$$

$n_i \to \infty$, i=1,2,3,....,s e a estimativa da matriz de variância-covariância de $\underset{\sim}{F}$ é dada por

$$\underset{\sim}{v}_{\underset{\sim}{F}} = \left[\left. \frac{\partial \underset{\sim}{F}(\underset{\sim}{\pi})}{\partial \underset{\sim}{\pi}} \right|_{\underset{\sim}{\pi}=\underset{\sim}{p}} \right] \underset{\sim}{V}(\underset{\sim}{p}) \left[\left. \frac{\partial \underset{\sim}{F}(\underset{\sim}{\pi})}{\partial \underset{\sim}{\pi}} \right|_{\underset{\sim}{\pi}=\underset{\sim}{p}} \right]' .$$

Se as sub-populações, formadas pela classificação cruzada dos níveis das variáveis explanatórias, têm tamanhos de amostras suficientes, então a variação entre as funções de resposta pode ser examinada por um modelo de regressão linear com mínimos quadrados ponderados:

$$\underset{\sim}{E}_A(\underset{\sim}{F}) = \underset{\sim}{E}_A[\underset{\sim}{F}(\underset{\sim}{p})] = \underset{\sim}{F}(\underset{\sim}{\pi}) = \underset{\sim}{X}\underset{\sim}{\beta}, \tag{2.8}$$

onde $E_A(.)$ representa o valor esperado assintótico de $\underset{\sim}{F}(\underset{\sim}{p})$, $\underset{\sim}{X}$ é uma matriz uxt de especificação do modelo, de posto completo $t \leq u$ e $\underset{\sim}{\beta}$ é um vetor tx1 de parâmetros desconhecidos que descrevem a variação entre as funções respostas.

2.2.1 Estimação e testes de hipóteses

A estimativa de mínimos quadrados ponderados de $\underset{\sim}{\beta}$, $\underset{\sim}{b}$, e sua matriz de variância-covariância $\underset{\sim}{V}_{\underset{\sim}{b}}$ são dadas por:

$$\underset{\sim}{b} = (\underset{\sim}{X}\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{X} \quad \text{e} \tag{2.9}$$

$$\underset{\sim}{V}_{\underset{\sim}{b}} \cong (\underset{\sim}{X}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{X})^{-1}. \tag{2.10}$$

Segundo Koch e Imrey (1985), $\underset{\sim}{b}$ tem distribuição assintoticamente normal multivariada com vetor de média $E_A(\underset{\sim}{b}) = \underset{\sim}{\beta}$ e variância $\underset{\sim}{V}_{\underset{\sim}{b}}$, sendo este um estimador consistente para $V_A(\underset{\sim}{b})$.

O ajuste do modelo pode ser verificado através da estatística de Wald, ou seja,

$$Q_W = (\underset{\sim}{F} - \underset{\sim}{X}\underset{\sim}{b})'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}(\underset{\sim}{F} - \underset{\sim}{X}\underset{\sim}{b}). \tag{2.11}$$

Sob a hipótese nula de que o modelo é adequado, Q_W tem distribuição χ^2 com $(u\text{-}t)$ graus de liberdade para sub-populações moderadamente grandes, isto é, $n_i \geq 25$.

Se o modelo $\underset{\sim}{F}(\underset{\sim}{\pi}) = \underset{\sim}{X}\underset{\sim}{\beta}$ descreve satisfatoriamente a variação entre os elementos de $\underset{\sim}{F}(\underset{\sim}{\pi})$, pode-se direcionar questões a respeito dos parâmetros com o uso de teste de hipóteses. Cada hipótese pode ser escrita na forma $H_0: \underset{\sim}{C}\underset{\sim}{\beta} = \underset{\sim}{0}$. A estatística de teste empregada é a de Wald, expressa da seguinte forma,

$$Q_C = (\underset{\sim}{C}\underset{\sim}{b})'(\underset{\sim}{C}\underset{\sim}{V}_{\underset{\sim}{b}}\underset{\sim}{C})^{-1}(\underset{\sim}{C}\underset{\sim}{b}), \tag{2.12}$$

onde Q_C tem distribuição χ^2 com graus de liberdade igual ao número de linhas linearmente independentes de C, sob a hipótese nula.

Valores preditos com base no modelo adotado são obtidos por $\hat{\underset{\sim}{F}} = \underset{\sim}{X}\underset{\sim}{b}$, e a matriz de variância-covariância correspondente estimada por : $\underset{\sim}{V}_{\underset{\sim}{F}} = \underset{\sim}{X}\underset{\sim}{V}_{\underset{\sim}{b}}\underset{\sim}{X}'$.

A metodologia básica apresentada acima pode facilmente ser adaptada a estudos com medidas repetidas desde que seja preservada a estrutura geral dos dados numa tabela de

contingência. Por exemplo: se uma variável resposta com C níveis categorizados é medida em t pontos de tempo ou condições, os r perfis de resposta multivariada formados pela classificação-cruzada serão $r = C^t$. Assim podem existir t(C-1) proporções marginais correlacionadas, logitos generalizados ou cumulativos ou mesmo t escores médios correlacionados (se a resposta é ordinal). E a representação na tabela será:

Tabela 2: Tabela de contingência para dados com medidas repetidas

sub-população	Indivíduos dentro das subpopulações	Tempos ou Condições			
		1	2	...	t
1	1	y_{111}	y_{121}	...	y_{1t1}
1	2	y_{112}	y_{122}	...	y_{1t2}
⋮	⋮	⋮	⋮		⋮
⋮	⋮	⋮	⋮		⋮
1	n_1	y_{11n_1}	y_{12n_1}	...	y_{1tn_1}
2	1	y_{211}	y_{221}	...	y_{2t1}
2	2	y_{212}	y_{222}	...	y_{2t2}
⋮	⋮	⋮	⋮		⋮
⋮	⋮	⋮	⋮		⋮
2	n_2	y_{21n_2}	y_{22n_2}	...	y_{2tn_2}
s	1	y_{s11}	y_{s21}	...	y_{st1}
s	2	y_{s12}	y_{s22}	...	y_{st2}
⋮	⋮	⋮	⋮		⋮
⋮	⋮	⋮	⋮		⋮
s	n_s	y_{s1n_s}	y_{s2n_s}	...	y_{stn_s}

onde y_{ijk} representa a resposta do k-ésimo indivíduo na i-ésima sub-população para a j-ésima condição; $i = 1,2,\ldots,s$; $j = 1,2,\ldots,t$ e $k = 1,2,\ldots,n_i$. As respostas possíveis de cada y_{ijk} são

indexadas por $c = 0,1,2,\ldots,C$ para a classificação do correspondente indivíduo dentro de alguma das (C+1) categorias de uma escala ordinal, nominal ou binária (C=1).

2.3 Razão de médias para análise de dados binários de uma amostragem aleatória por conglomerados

Nos dias de hoje, inclusive por razão das facilidades computacionais, encontra-se disponível ao pesquisador uma vasta gama de procedimentos estatísticos para análise de dados. Porém, a escolha não cuidadosa do método a ser implementado pode levar a que os resultados obtidos indiquem ou resultem em inferências erradas sobre a população em estudo. Em particular, o papel da amostragem num determinado estudo deve ser levado em consideração quando da escolha do método de análise, pois a complexidade do desenho amostral está frequentemente conectada com a complexidade do procedimento de estimação. É comum em diversas áreas como a de saúde, pesquisa de mercado e ciências sociais, deparar-se com estudos onde a variável resposta de interesse é categorizada e possui uma estrutura de amostragem envolvendo conglomerados em um ou mais estágios.

É sabido que se as estimativas são baseadas em amostragem probabilística complexa, e suas variâncias são frequentemente diferentes daquelas baseadas em amostragem aleatória simples e, também, que a estratificação pode ajudar a reduzir a variabilidade, enquanto que a técnica de conglomerado e a probabilidade de seleção desigual podem aumentar esta variabilidade. Por outro lado, em muitas pesquisas, a técnica de conglomerados é a que cabe pela estrutura dos dados na população. Ao se utilizar medidas ou estimativas do tipo razão, esta variabilidade pode ser controlada, principalmente quando existir variação no tamanho dos conglomerados (Hansen *et al.*, 1953). Também o uso do peso de amostragem, que reflete algumas características do delineamento como, por exemplo, a probabilidade de seleção desigual, vem ajudar (Landis *et al.*, 1982).

Métodos de estimação de razão têm sido historicamente usados não somente na análise de dados com estrutura complexa (Koch *et al.*, 1975, Freeman *et al.*, 1976 e Landis *et al.*, 1987), onde o interesse é estimar taxas ou proporções para subgrupos populacionais definidos pela classificação cruzada de variáveis explanatórias, como também em situações para manusear dados faltantes (Stanish *et al.*, 1978). Recentemente Lavange *et al.* (1994) propuseram o uso do método da razão multivariada para análise de densidades de incidência em um estudo observacional de infecção respiratória baixa em crianças durante o seu primeiro ano de vida. Este método é válido desde que amostras de tamanho grande estejam disponíveis para assegurar esta análise. O método da razão é usado por Snyder (1993) para produzir estimativas da resposta média geral e erro-padrão, levando em consideração a variação aleatória no tamanho da amostra relativo ao procedimento de amostragem por conglomerado. Para o caso de medidas repetidas, o método da razão pode ser facilmente estendido, porém nenhuma informação a respeito da estrutura de dependência é fornecida.

O estimador da razão de médias, seu erro-padrão e um teste estatístico assintótico para contrastes de duas ou mais razões de médias serão abordados (Snyder, 1993). Estas idéias serão estendidas para um vetor de razões de médias (ou um vetor do log de razões de médias), correspondendo à classificação cruzada de covariáveis categorizadas para serem modeladas usando o método dos mínimos quadrados ponderados.

Primeiramente, será apresentado o método da razão para estimar uma proporção proveniente de uma amostra de conglomerados a um estágio. Em seguida, uma extensão deste método para razão oriunda de variáveis discretas, correspondente à classificação cruzada de covariáveis categorizadas, será apresentada na forma de subseções. As covariáveis podem representar as características do conglomerado como um todo ou das subunidades dentro do conglomerado. Por fim, mostra-se este método para a situação de uma amostragem de conglomerado em dois estágios.

2.3.1 Definição da razão de médias

O estimador da razão de médias para a média populacional geral por elementos de um atributo de interesse é definido nesta seção para observações de uma resposta binária de indivíduos numa amostragem de conglomerados a um estágio ou mesmo para medidas repetidas de um indivíduo num estudo longitudinal. O método de amostragem assumido para os conglomerados é amostragem aleatória simples com reposição (ou equivalentemente sem reposição para uma população grande). Embora a notação usada nesta seção pareça ser complexa, sua utilidade dar-se-à nas seções seguintes.

Seja $i = 1, 2, ..., N$ o índice referente aos conglomerados amostrados, $j = 1, 2, ..., M_i$ o índice referente aos elementos no i-ésimo conglomerado, $t = 1, 2, ..., v_{ij}$ o índice das observações múltiplas do j-ésimo elemento. N representa o número de conglomerados selecionados, M_i o número total de elementos no i-ésimo conglomerado e v_{ij} o número total de observações potenciais para o j-ésimo elemento no i-ésimo conglomerado.

Sejam Y_{ijt} uma resposta binária, que assume o valor 1 se a t-ésima observação para o j-ésimo elemento no i-ésimo conglomerado é relevante e tem o atributo de interesse e 0 caso contrário, e X_{ijt} uma resposta binária, que assume o valor 1 se a t-ésima observação para o j-ésimo elemento no i-ésimo conglomerado é relevante (ou observada) e 0 caso contrário. Em algumas aplicações, todas as observações para cada indivíduo são relevantes e nesse caso, $X_{ijt} = 1$.

Define-se

$$Y_{i..} = \sum_{j=1}^{M_i} \sum_{t=1}^{v_{ij}} Y_{ijt} \quad \text{e} \quad X_{i..} = \sum_{j=1}^{M_i} \sum_{t=1}^{v_{ij}} X_{ijt}$$

como o número total de observações relevantes com o atributo e o número total de observações relevantes para o i-ésimo conglomerado, respectivamente. Como o método de amostragem é aleatório simples com reposição, os vetores $(Y_{i..}, X_{i..})$ são independentes e identicamente distribuídos.

O estimador da razão de médias para a proporção de observações com o atributo para todos os conglomerados é definido como:

$$R=\frac{\sum_{i=1}^{N} Y_{i..}/N}{\sum_{i=1}^{N} X_{i..}/N}=\frac{\overline{Y}}{\overline{X}} \quad \text{onde} \quad \overline{Y}=\frac{\sum_{i=1}^{N} Y_{i..}}{N} \quad \text{e} \quad \overline{X}=\frac{\sum_{i=1}^{N} X_{i..}}{N}. \tag{2.13}$$

R pode ser entendido como o número médio estimado por conglomerados de ocorrências de observações relevantes com o atributo, dividido pelo número médio estimado por conglomerados de ocorrências relevantes, ou ainda a proporção estimada de ocorrências com o atributo de interesse entre observações relevantes.

A matriz de covariância estimada correspondente à estimativa da razão é calculada via aproximação da série de Taylor de primeira ordem, isto é, o estimador da variância de R pode ser calculado notando que R é uma função não linear de duas estatísticas, podendo assim ser expandida via série de Taylor em torno de $\mu_x = E(X)$ e $\mu_y = E(Y)$ como segue:

$$R=\frac{\mu_Y}{\mu_X}+\frac{1}{\mu_X}(\overline{Y}-\mu_Y)-\frac{\mu_Y}{\mu_X^2}(\overline{X}-\mu_X)+O(1/N). \tag{2.14}$$

O valor esperado assintótico de 1ª ordem em série de Taylor para R, $E_A(R)$, é $\frac{\mu_Y}{\mu_X}=\theta$, a razão de médias na população.

$$R=\frac{\mu_Y}{\mu_X}+\frac{1}{\mu_X}\{(\overline{Y}-\theta\overline{X})-(\mu_Y-\theta\mu_X)\}+O(1/N).$$

A variância de R, baseada na linearização em série de Taylor é

$$V(R)=\left(\frac{\mu_Y}{\mu_X}\right)\left\{\frac{var(\overline{Y})}{\mu_Y^2}-\frac{2cov(\overline{Y},\overline{X})}{\mu_X\mu_Y}+\frac{var(\overline{X})}{\mu_X^2}\right\} \tag{2.15}$$

$$=\frac{1}{\mu_X^2}\{var(\overline{Y}-\theta\overline{X})\}.$$

Um estimador consistente para $V(R)$ é dado por

$$v(R) = \frac{R^2}{N}\left\{\frac{s_Y^2}{\overline{Y}^2} - \frac{2s_{XY}}{\overline{Y}\,\overline{X}} + \frac{s_X^2}{\overline{X}^2}\right\} \qquad (2.16)$$

$$= \frac{1}{N\overline{X}^2}\sum_{i=1}^{N}\frac{(Y_{i.} - RX_{i.})^2}{N-1}, \text{ onde}$$

$$s_Y^2 = \sum_{i=1}^{N}\frac{(Y_{i.} - \overline{Y})^2}{N-1},$$

$$s_X^2 = \sum_{i=1}^{N}\frac{(X_{i.} - \overline{X})2}{N-1},$$

$$s_{XY} = \sum_{i=1}^{N}\frac{(X_{i.} - \overline{X})(Y_{i.} - \overline{Y})}{N-1}.$$

Em muitas situações o interesse pode ser comparar duas ou mais razões de médias para subgrupos definidos pelos níveis das características dos conglomerados. Por exemplo, comparar a prevalência de diarréia entre o grupo vitaminado e o placebo. Então uma estatística que pode ser usada com este propósito, isto é, comparar duas razões R e R' é a seguinte:

$$Q = \frac{[ln(R/R')]^2}{\{v[ln(R/R')]\}'}, \text{ onde} \qquad (2.17)$$

$v[ln(R/R')] = \frac{v(R)}{R^2} + \frac{v(R')}{R'^2}$. Q é a estatística Score, e tem distribuição aproximadamente qui-quadrado com 1 grau de liberdade para amostras grandes, sob a hipótese nula de que o quociente de R e R' é 1.

Agora, quando se quer comparar duas ou mais razões de médias para subgrupos definidos pelos níveis de uma característica referente a um mesmo conglomerado, a estatística usada é a mesma, apenas uma modificação é feita na variância estimada, isto é,

$$Q = \frac{[ln(R/R')]^2}{\{v[ln(R/R')]\}'},$$

onde $v[ln(R/R')] = \frac{v(R)}{R^2} + \frac{v(R')}{R'^2} - 2\frac{cov(R,R')}{RR'}$.

O intervalo de confiança de *(1-α)*% para a estimativa da razão de taxa, $\frac{R}{R'}$, é dado por

$$exp\left\{ln\left(\frac{R}{R'}\right) \mp z_{\frac{1-\alpha}{2}}\sqrt{v\left[ln\left(\frac{R}{R'}\right)\right]}\right\}. \qquad (2.18)$$

Se o interesse for além de uma simples estimativa pontual, ou comparações de duas ou mais razões de médias, isto é, o ajuste de algum modelo, pode-se, por exemplo, ajustar um modelo log-linear para as razões estimadas com o propósito de testar a significância dos efeitos das covariáveis, aplicando o método dos mínimos quadrados ponderados. Este tópico será abordado ainda neste capítulo.

Num estudo longitudinal, onde as medidas repetidas de um indivíduo formam um conglomerado, o método da razão descrito até aqui pode ser usado desde que os indivíduos do estudo tenham sido selecionados ao acaso com reposição da população alvo. Esta é a única suposição feita até o momento. Nota-se que nenhuma suposição da estrutura de correlação entre as observações repetidas de um indivíduo é considerada no cálculo do estimador da variância.

Na verdade, para cada indivíduo calcula-se um desvio das suas observações repetidas em relação a uma média geral ponderada, obtendo-se assim uma medida única para cada indivíduo. Ou seja, denotando i como indivíduo e $j = 1, \ldots, m_i$ como as observações repetidas, $M = \sum_{i=1}^{n} m_i$ é o número total de observações no estudo, e portanto para cada indivíduo temos

$$Y_i^* = \sum_{j=1}^{m_i}\left(Y_{ij} - \frac{M}{m_i}\overline{Y}\right) \quad \text{e} \quad X_i^* = \sum_{j=1}^{m_i}\left(X_{ij} - \frac{M}{m_i}\overline{X}\right). \qquad (2.19)$$

Logo, o estimador da variância de R é o mesmo mencionado anteriormente, isto é,

$$v(R) = \frac{1}{n(n-1)\overline{X}^2}\sum_{i=1}^{n}(Y_i^* - RX_i^*)^2. \qquad (2.20)$$

2.4 Extensão da razão de médias para resposta discreta de uma amostragem por conglomerado

Tendo como motivação a estrutura dos dados dos estudos descritos em 1.3 e pela utilidade de quantidades como incidência de um evento durante um período de seguimento ou densidade de incidência ou prevalência de uma doença em estudos epidemiológicos, a metodologia desenvolvida resumidamente na seção anterior, e apresentada em Snyder (1993), é estendida para análise de resposta discreta bivariada de conglomerados. Aqui, outras definições para o estimador da razão de médias podem ser usadas para estimar quantidades similares de variáveis discretas ou contínuas de interesse. Como tais medidas expressam uma razão de soma de variáveis aleatórias, sem a suposição de uma distribuição exata dessas variáveis, o método descrito anteriormente fornece um caminho para estimar e modelar tais medidas, ajustando-as às variáveis explicativas ou fatores de risco de interesse.

A seguir é apresentada a metodologia para o caso de razão de variáveis aleatórias discretas. Vale lembrar que os dois conjuntos de dados mencionados no capítulo I servirão de base para ilustração desta metodologia e serão analisados no capítulo IV.

i. AISAM- Avaliação do Impacto das Medidas de Saneamento Ambiental em Áreas Pauperizadas de Salvador. Pesquisa desenvolvida no período de Agosto de 1989 a Dezembro 1990, tendo como objetivo avaliar o impacto das ações de saneamento na saúde da população de Salvador, cuja área de estudo encontra-se na periferia.

ii. SERRINHA- Efeito da Suplementação de Vitamina A na Diarréia e Infeção Respiratória Aguda. Pesquisa desenvolvida na período de Dezembro de 1990 a Dezembro de 1991 na cidade de Serrinha, com o objetivo de avaliar a redução na morbidade de crianças.

As duas pesquisas diferem quanto ao processo de amostragem, porém, as medidas a serem utilizadas no método da razão são semelhantes. Como a densidade de incidência é uma razão de soma de duas variáveis aleatórias, o método de estimação da razão representa bem essa medida epidemiológica.

2.4.1 Definição de razão de médias para respostas discretas

O estimador da razão de médias para representar a média populacional de uma quantidade de interesse será definido nesta seção para observações discretas provenientes de uma amostragem por conglomerados, com tamanhos diferentes, a um estágio.

No estudo de Serrinha i =1, 2, ..., N, representa o índice dos conglomerados amostrados (no exemplo seriam as crianças). Assim N representa o número total de crianças selecionadas para o estudo.

Sejam

Y_i = variável discreta que assume um determinado valor se o i-ésimo conglomerado apresenta a condição de interesse;

X_i = variável discreta que assume um determinado valor se ocorre a observação do i -ésimo conglomerado.

Para os dados de Serrinha a condição de interesse é se a criança apresentar episódio de diarréia, isto é, apresentar 3 ou mais dejeções líquidas / amolecidas no período de 24 horas seguido de um intervalo de pelo menos 3 dias sem diarréia. Portanto

Y_i = número de episódios de diarréia para a i-ésima criança selecionada;

X_i = número de dias de acompanhamento da i-ésima criança selecionada.

Define-se

$$Y = \sum_{i=1}^{N} Y_i \quad \text{e} \quad X = \sum_{i=1}^{N} X_i$$

onde Y é o número total de episódios de diarréia e X é o número total de crianças-dia no estudo.

Como o método de amostragem é aleatório simples com reposição, os *(Y, X)* são independentes e identicamente distribuídos. O estimador da razão de médias para a medida epidemiológica de interesse, é definido como:

$$R = \frac{\sum_{i=1}^{N} Y_i / N}{\sum_{i=1}^{N} X_i / N} = \frac{Y/N}{X/N} = \frac{\bar{Y}}{\bar{X}}. \qquad (2.21)$$

Na verdade, o estimador da razão definido no contexto da medida de interesse, para os dados em questão, expressa uma quantidade mais ampla, isto é, uma taxa.

Para este conjunto de dados, R é interpretado como a densidade de incidência de diarréia entre todas as crianças selecionadas no estudo, ou ainda, o número de episódios por crianças-dia de acompanhamento.

Já com relação ao estudo AISAM, considere $i = 1,2, ..., N$ o índice que representa os conglomerados amostrais, $j = 1,2,..., M_i$ o índice de todos os elementos no i-ésimo conglomerado amostrado e $t = 1,2, ..., v_{ij}$ o índice das observações múltiplas do j-ésimo elemento no i-ésimo conglomerado. Assim N é o número de conglomerados selecionados, no caso as comunidades, M_i é o número de elementos no i-ésimo conglomerado, isto é, número total de domicílios na comunidade i e v_{ij} é o número de observações múltiplas, isto é, número de crianças potenciais por domicílio j na comunidade i.

Sejam ,

Y_{ijt} variável discreta que assume um determinado valor se a t-ésima observação para o j-ésimo elemento do i-ésimo conglomerado apresentar a condição de interesse;

X_{ijt} variável discreta que assume um determinado valor se a t-ésima observação para o j-ésimo elemento do i-ésimo conglomerado estiver presente na quinzena de acompanhamento.

Para este exemplo a condição de interesse é se a criança apresenta 1 ou mais dias de diarréia detectada pela mãe, separado de qualquer outro episódio com pelo menos 2 dias livres do sintoma. Portanto,

Y_{ijt} número de episódios de diarréia para a t-ésima criança do j-ésimo domicilio na i-ésima comunidade;

X_{ijt} número de quinzenas observadas para a t-ésima criança do j-ésimo domicilio na i-ésima comunidade.

Como cada quinzena representa um período de 14 dias fixos e houve 26 quinzenas de acompanhamento no estudo, pode-se transformar X_{ijt} no número de dias de acompanhamento.

Definindo então

$$Y_{i..} = \sum_{j=1}^{M_i} \sum_{t=1}^{v_{ij}} Y_{ijt} \quad \text{e} \quad X_{i..} = 14 \sum_{j=1}^{M_i} \sum_{t=1}^{v_{ij}} X_{ijt}$$

como o número total de episódios de diarréia para a i-ésima comunidade e o número total de crianças-dia de acompanhamento no estudo, respectivamente.

Uma vez que o método de amostragem é aleatório simples com reposição, os $(Y_{i..}, X_{i..})$ são independentes e identicamente distribuídos. O estimador da razão de médias para a medida epidemiológica de interesse é definido como:

$$R = \frac{\sum_{i=1}^{N} Y_{i..}/N}{\sum_{i=1}^{N} X_{i..}/N} = \frac{\overline{Y}}{\overline{X}}. \qquad (2.22)$$

R pode ser interpretado como a densidade de incidência de diarréia geral ou, ainda, o número de episódios por crianças-dia no estudo.

De forma análoga à descrita anteriormente, a representação de R numa série de Taylor de primeira ordem em relação a média populacional (μ_y, μ_x) é a mesma de (2.14). Assim a variância de R e seu estimador são dados por (2.15) e (2.16), respectivamente.

Existe interesse em considerar as seguintes idéias: examinar separadamente as possibilidades de se calcular a razão de médias de acordo com o delineamento amostral apresentado na seção 2.4.2 e, posteriormente, o cálculo da razão de médias para os subgrupos definidos pela classificação cruzada das características do delineamento amostral simultaneamente, objetivando assim a formulação de um modelo de regressão para as razões de médias, seção 2.4.3.

2.4.2 Estratificação da razão de médias pelas características do delineamento amostral

A razão de médias pode ser calculada separadamente para subgrupos de observações correspondentes à classificação cruzada dos níveis das covariáveis que representam o conglomerado, ou das que representam os elementos dentro do conglomerado, ou ainda das que representam as unidades de análise. Nesta situação, o método de amostragem assumido é amostragem aleatória estratificada com reposição. Para o estudo na cidade de Salvador (AISAM), a razão de médias pode ser calculada utilizando somente as características dos conglomerados (comunidades) ou para alguma característica, que representa os elementos dentro do conglomerado, isto é, os domicílios, ou ainda para as características referentes às crianças do domicílio selecionado. Assim, por exemplo, a razão de médias pode representar as comunidades sem nenhuma intervenção de saneamento (Grupo 1) ou os domicílios com piso de terra ou as crianças do sexo masculino.

Seja então, h =1,2, ..., H o índice referente aos estratos formados pela classificação cruzada das características dos conglomerados. No estudo em questão, uma característica que pode ser considerada para o conglomerado é o tipo de saneamento disponível nas comunidades; i =1,2, ..., N_h , o índice dos conglomerados amostrados no estrato h, j =1,2, ..., M_{hi} , o índice de todos os elementos no i-ésimo conglomerado do estrato h, e t =1,2, ... , v_{hij} , o índice das observações múltiplas para o elemento j no conglomerado i do estrato h. Assim H

representa o número de estratos formados por uma característica dos conglomerados (no caso H=3), N_h o número de comunidades amostradas no estrato h, M_{hi} representa o número de domicílios no i-ésimo conglomerado do estrato h e v_{hij} o número total de crianças potenciais por domicílio j no conglomerado i do estrato h.

Sejam ,

Y_{hijt} variável discreta que assume um determinado valor se a t-ésima observação para o j-ésimo elemento no i-ésimo conglomerado do h-ésimo estrato apresentar a condição de interesse;

X_{hijt} variável discreta que assume um determinado valor se a t-ésima observação para o j-ésimo elemento no i-ésimo conglomerado do h-ésimo estrato estiver presente na quinzena de acompanhamento.

Da mesma forma descrita anteriormente, a condição de interesse é se a criança apresenta 1 ou mais dias de diarréia detectada pela mãe, separado de qualquer outro episódio de pelo menos 2 dias livre do sintoma. Assim,

Y_{hijt} = número de episódios de diarréia para a t-ésima criança do j-ésimo domicílio na i-ésima comunidade na h-ésima condição de saneamento;

X_{hijt} = número de quinzenas observadas para a t-ésima criança do j-ésimo domicílio na i-ésima comunidade na h-ésima condição de saneamento.

Define-se

$$Y_{hi..} = \sum_{j=1}^{M_i} \sum_{t=1}^{v_{ij}} Y_{hijt} \quad \text{e} \quad X_{hi..} = 14 \sum_{j=1}^{M_i} \sum_{t=1}^{v_{ij}} X_{hijt}$$

onde $Y_{hi..}$ é o número total de episódios de diarréia na i-ésima comunidade da h-ésima condição de saneamento e $X_{hi..}$ é o número total de crianças-dia observadas na i-ésima comunidade da h-ésima condição de saneamento.

O estimador da razão de médias para a medida epidemiológica de interesse é definida como

Um intervalo de confiança ao nível de *(1-α)* para a razão $\frac{\theta_h}{\theta_{h'}}$, baseado em amostras grandes, é dado por

$$exp\left\{ln\left(R_h/R_{h'}\right) \pm z_{1-\frac{\alpha}{2}} v\left[ln\left(R_h/R_{h'}\right)\right]^{1/2}\right\}. \qquad (2.29)$$

Alternativamente, a razão de médias pode ser calculada para subgrupos definidos pelos níveis das características dos elementos, isto é, os domicílios. Para o estudo AISAM, as incidências de diarréia podem ser calculadas separadamente, por exemplo, para domicílios que tratam água (fervida ou filtrada) ou para os que não tratam água, ou as incidências de diarréia podem ser calculadas para domicílios com sanitário ou sem sanitário. É importante lembrar que, neste caso, as razões de médias são correlacionadas uma vez que podem ser provenientes do mesmo conglomerado.

De forma similar ao início desta seção, define-se como i =1,2, ... ,N o índice referente aos conglomerados amostrais (as comunidades selecionadas), j =1,2, ... ,M_i o índice dos elementos (domicílios) selecionados no i-ésimo conglomerado, t =1,2, ..., v_{ij} o índice das observações múltiplas (crianças menores de 5 anos) do elemento j no conglomerado i, e k =1,2, ... , K o índice de alguma característica do domicílio. Assim, N é o número de comunidades selecionadas, M_i é o número de domicílios na i-ésima comunidade, v_{ij} o número total de crianças potenciais para o domicílio j na comunidade i, e K é o número de níveis de uma ou mais características referente ao domicílio.

As variáveis discretas concernentes a esta condição são construídas da seguinte forma:

Y_{ikjt} número de episódios de diarréia para a t-ésima criança do j-ésimo domicílio do tipo k da i-ésima comunidade;

X_{ikjt} número de quinzenas observadas para a t-ésima criança do j-ésimo domicílio do tipo k da i-ésima comunidade.

Define-se então,

$$R_h = \frac{\sum_{i=1}^{N} Y_{hi..}/N}{\sum_{i=1}^{N} X_{hi..}/N} = \frac{\overline{Y}_h}{\overline{X}_h}. \qquad (2.23)$$

A quantidade R_h é a incidência de diarréia da h-ésima condição de saneamento, ou ainda, o número de episódios por crianças-dia para h-ésima condição de saneamento. Para os dados do AISAM, R_h pode ser, por exemplo, a incidência de diarréia para as comunidades sem nenhuma intervenção de saneamento.

O valor esperado assintótico de R_h na série de Taylor de primeira ordem, $E_A(R_h)$ é $\frac{\mu_{hy}}{\mu_{hx}} = \theta_h$, a razão de médias na população do h-ésimo estrato.

A representação de R_h numa série de Taylor de primeira ordem em relação a média populacional (μ_{hy}, μ_{hx}) dos $(Y_{hi..}, X_{hi..})$ é dada por:

$$R_h = \frac{\mu_{hY}}{\mu_{hX}} + \frac{1}{\mu_{hX}}(\overline{Y}_h - \mu_{hY}) - \frac{\mu_{hY}}{\mu_{hX}^2}(\overline{X}_h - \mu_{hX}) + O(1/N). \qquad (2.24)$$

A variância de R_h, baseada no método de linearização da série de Taylor, é dada abaixo:

$$V(R_h) \cong \left(\frac{\mu_{hY}}{\mu_{hX}}\right)^2 \left\{\frac{Var(\overline{Y}_h)}{\mu_{hY}^2} - \frac{2Cov(\overline{Y}_h, \overline{X}_h)}{\mu_{hX}\mu_{hY}} + \frac{Var(\overline{X}_h)}{\mu_{hX}^2}\right\}. \qquad (2.25)$$

Um estimador consistente para $V(R_h)$ para grandes amostras é dado por:

$$v(R_h) \cong \frac{R_h^2}{N_h}\left\{\frac{s_{hY}^2}{\overline{Y}_h^2} - \frac{2s_{hXY}}{\overline{Y}_h\overline{X}_h} + \frac{s_{hX}^2}{\overline{X}_h^2}\right\} \qquad (2.26)$$

onde

$$s_{hY}^2 = \sum_{i=1}^{N_h}(Y_{hi..} - \overline{Y}_h)^2/N_h - 1,$$

$$s_{hX}^2 = \sum_{i=1}^{N_h}(X_{hi..} - \overline{X}_h)^2/N_h - 1,$$

$$s_{hXY} = \sum_{i=1}^{N_h} (X_{hi..} - \overline{X}_h)(Y_{hi..} - \overline{Y}_h)/N_h - 1,$$

$$\overline{Y}_h = \sum_{i=1}^{N_h} Y_{hi..}/N_h,$$

$$\overline{X}_h = \sum_{i=1}^{N_h} X_{hi..}/N_h.$$

Um estimador para grandes amostras de $V(ln\ R_h)$ baseado na linearização da série de Taylor é dado por

$$v(ln\ R_h) = \frac{v(R_h)}{R_h^2}. \tag{2.27}$$

Um intervalo de confiança ao nível de $(1-\alpha)$ para θ_h, baseado em grandes amostras é

$$exp\left\{ ln\ R_h \pm z_{1-\frac{\alpha}{2}} \left[v(R_h)^{1/2} / R_h \right] \right\}. \tag{2.28}$$

Do mesmo modo como descrito anteriormente, pode existir o interesse em comparar duas ou mais razões de médias para subgrupos definidos pelos níveis da característica dos conglomerados. Para o conjunto de dados AISAM, pode ser de interesse comparar as incidências de diarréia entre as comunidades sem nenhuma e com alguma intervenção de saneamento.

Um teste estatístico baseado em grandes amostras pode ser utilizado para comparar duas razões de médias R_h e $R_{h'}$, e é dado por

$$Q = \frac{\left\{ ln\left(R_h / R_{h'} \right) \right\}^2}{\left\{ v\left[ln\left(R_h / R_{h'} \right) \right] \right\}}, \text{ onde } v\left[ln\left(R_h / R_{h'} \right) \right] = \left\{ \frac{v(R_h)}{R_h^2} + \frac{v(R_{h'})}{R_{h'}^2} \right\}. \text{ Sob a hipótese}$$

nula de que o quociente de R_h e $R_{h'}$ é 1, a estatística Q, tem aproximadamente uma distribuição qui-quadrado com um grau de liberdade para grandes amostras.

$$Y_{ik..} = \sum_{j=1}^{M_i} \sum_{t=1}^{v_{ij}} Y_{ikjt} \quad \text{e} \quad X_{ik..} = 14 \sum_{j=1}^{M_i} \sum_{t=1}^{v_{ij}} X_{ikjt}$$

como o número total de episódios de diarréia por domicílio do tipo k na i-ésima comunidade e o número total de crianças-dia observadas para domicílios do tipo k na i-ésima comunidade respectivamente.

Os vetores $(Y_{i1..}, X_{i1..}, \ldots, Y_{ik..}, X_{ik..})'$ são também independentes e identicamente distribuídos, como consequência do método de amostragem para o conglomerado ser aleatório simples com reposição. O estimador da razão de médias para a medida epidemiológica de interesse é definido abaixo:

$$R_k = \frac{\sum_{i=1}^{N} Y_{ik..}/N}{\sum_{i=1}^{N} X_{ik..}/N} = \frac{\overline{Y}}{\overline{X}}. \qquad (2.30)$$

R_k neste caso é a incidência de diarréia para domicílios do tipo k, ou número de episódios por crianças-dia para domicílios do tipo k. Por exemplo, R_k pode ser a incidência de diarréia para os domicílios com sanitário ou para os domicílios com piso de terra.

O valor esperado assintótico de R_k na série de Taylor de primeira ordem, $E_A(R_k)$ é $\frac{\mu_{kY}}{\mu_{kX}} = \theta$, a razão de médias para a k-ésima sub-população.

A representação de R_k numa série de Taylor é dada por:

$$R_k = \theta_k + \theta_k \frac{(\overline{Y}_k - \mu_{kY})}{\mu_{kY}} - \theta_k \frac{(\overline{X}_k - \mu_{kX})}{\mu_{kX}} + O(1/N). \qquad (2.31)$$

A variância de R_k, baseada no método de linearização da série de Taylor, é dada por:

$$V(R_k) \cong \left(\frac{\mu_{kY}}{\mu_{kX}}\right)^2 \left\{ \frac{V(\overline{Y}_k)}{\mu_{kY}^2} - \frac{2Cov(\overline{Y}_k, \overline{X}_k)}{\mu_{kY}\mu_{kX}} + \frac{V(\overline{X}_k)}{\mu_{kX}^2} \right\}. \qquad (2.32)$$

uma vez que $(Y_{ik.}, X_{ik.})$ são independentes e identicamente distribuídos. Um estimador para a variância de R_k é dada por:

$$v(R_k)=\frac{R_k^2}{N}\left\{\frac{s_{kY}^2}{\overline{Y}_k^2}-\frac{2s_{kXY}}{\overline{Y}_k\overline{X}_k}+\frac{s_{kX}^2}{\overline{X}_k^2}\right\}, \text{ onde} \qquad (2.33)$$

$$s_{kY}^2=\sum_{i=1}^{N}(Y_{ik..}-\overline{Y}_k)^2/N-1,$$

$$s_{kX}^2=\sum_{i=1}^{N}(X_{ik..}-\overline{X}_k)^2/N-1,$$

$$s_{kXY}=\sum_{i=1}^{N}(X_{ik..}-\overline{X}_k)(Y_{ik..}-\overline{Y}_k)^2/N-1,$$

$$\overline{Y}_k=\sum_{i=1}^{N}Y_{ik..}/N,$$

$$\overline{X}_k=\sum_{i=1}^{N}X_{ik..}/N.$$

A covariância de R_k e $R_{k'}$, baseada na linearização da série de Taylor, é dada por:

$$\mathrm{Cov}(R_k,R_{k'})\cong\theta_k\theta_{k'}\left\{\frac{\mathrm{Cov}(\overline{Y}_k,\overline{Y}_{k'})}{\mu_{kY}\mu_{k'Y}}-\frac{\mathrm{Cov}(\overline{Y}_k,\overline{X}_{k'})}{\mu_{kY}\mu_{k'Y}}-\frac{\mathrm{Cov}(\overline{Y}_{k'},\overline{X}_k)}{\mu_{kY}\mu_{k'Y}}+\frac{\mathrm{Cov}(\overline{X}_k,\overline{X}_{k'})}{\mu_{kY}\mu_{k'Y}}\right\} \qquad (2.34)$$

uma vez que $(Y_{ik..},X_{ik..},Y_{ik'..},X_{ik'..})'$ são independentes e identicamente distribuídos. Um estimador da $\mathrm{Cov}(R_k,R_{k'})$ é dado por

$$cov(R_k,R_{k'})=\frac{R_kR_{k'}}{N}\left\{\frac{s_{kk'Y}}{\overline{Y}_k\overline{Y}_{k'}}-\frac{s_{kk'YX}}{\overline{Y}_k\overline{X}_{k'}}-\frac{s_{k'kYX}}{\overline{Y}_{k'}\overline{X}_k}+\frac{s_{kk'X}}{\overline{X}_k\overline{X}_{k'}}\right\}, \qquad (2.35)$$

onde

$$s_{kk'Y}=\sum_{i=1}^{N}(Y_{ik..}-\overline{Y}_k)(Y_{ik'..}-\overline{Y}_{k'})/N-1,$$

$$s_{kk'YX}=\sum_{i=1}^{N}(Y_{ik..}-\overline{Y}_k)(X_{ik'..}-\overline{X}_{k'})/N-1,$$

$$s_{k'kYX}=\sum_{i=1}^{N}(Y_{ik'..}-\overline{Y}_{k'})(X_{ik..}-\overline{X}_k)/N-1,$$

$$s_{kk'X}=\sum_{i=1}^{N}(X_{ik..}-\overline{X}_k)(X_{ik'..}-\overline{X}_{k'})/N-1.$$

Havendo interesse em comparar duas ou mais razões de médias para subgrupos definidos pelos níveis das características dos domicílios, utiliza-se a mesma estatística Q apresentada anteriormente.

Razões de médias podem ser também calculadas para subgrupos de observações definidos pelos níveis de uma ou mais características das crianças. Para o estudo AISAM, as incidências de diarréia podem ser determinadas, por exemplo, para crianças do sexo masculino ou feminino separadamente. Sendo assim, define-se

Y_{ijlt} como o número de episódios de diarréia para a t-ésima criança com característica l do j-ésimo domicílio da i-ésima comunidade;

X_{ijlt} como o número de quinzenas observadas para a t-ésima criança com característica l do j-ésimo domicílio da i-ésima comunidade.

E,

$$Y_{i.l.} = \sum_{j=1}^{M_i} \sum_{t=1}^{v_{ij}} Y_{ijlt} \quad \text{e} \quad X_{i.l.} = 14 \sum_{j=1}^{M_i} \sum_{t=1}^{v_{ij}} X_{ijlt}$$

onde $Y_{i.l.}$ é o número total de episódios de diarréia para crianças do tipo l na i-ésima comunidade e $X_{i.l.}$ é o número total de crianças-dia para crianças do tipo l na i-ésima comunidade.

O estimador da razão de médias para a medida epidemiológica de interesse é

$$R_l = \frac{\sum_{i=1}^{N} Y_{i.l.}/N}{\sum_{i=1}^{N} X_{i.l.}/X_{i.l.}} = \frac{\overline{Y}}{\overline{X}}, \tag{2.36}$$

assim, a quantidade R_l é a incidência de diarréia para crianças com característica l, ou número de episódios por crianças-dia para crianças com a característica l. Para exemplificar, R_l pode ser a incidência de diarréia para crianças do sexo feminino.

A esperança e variância assintóticas de R_l e a covariância entre R_l e $R_{l'}$ podem ser obtidas substituindo o índice k e k' das formulas anteriores por l e l'.

2.4.3 Subgrupos de razão de médias definidos pelos níveis das características das comunidades, domicílios e crianças

Razões de médias também podem ser calculadas para subgrupos de observações definidas pela classificação cruzada (ou cruzamentos) dos níveis das características dos conglomerados, dos elementos dentro do conglomerado e dos indivíduos simultaneamente. Este cruzamento poderá ser feito caso exista um número adequado de observações em cada sub-população de interesse para assegurar a (log) normalidade dos estimadores das razões de médias.

Sejam h =1,2, ..., H o índice referente aos níveis das características dos conglomerados, i =1,2, ... ,N_h o índice referente aos conglomerados amostrais no estratos (as comunidades selecionadas); k =1,2, ..., K o índice dos níveis das características dos elementos dentro dos conglomerados; l =1,2, ..., L o índice dos níveis das características dos indivíduos; j =1,2, ..., M_{hi} o índice dos elementos estudados no conglomerado i do estrato h; t =1,2, ...,v_{hij} o índice das observações múltiplas por elemento j no conglomerado i do estrato h. Então N_h é o número de comunidades selecionadas no h-ésimo estrato, M_{hi} é o número de domicílios na i-ésima comunidade do h-ésimo estrato, v_{hij} é o número total de crianças potenciais para o domicílio j na comunidade i do estrato h, K é o número de níveis do tipo de domicílio, L é o número de níveis da característica da criança e H é o número de estratos formados por uma característica das comunidades.

Sejam

Y_{hikjlt} número de episódios de diarréia para a t-ésima criança com característica l do j-ésimo domicílio do tipo k na i-ésima comunidade do estrato h;

X_{hikjlt} número de quinzenas observadas para a t-ésima criança com característica l do j-ésimo domicílio do tipo k na i-ésima comunidade do estrato h.

Define-se

$$Y_{hik.l.} = \sum_{j=1}^{M_{hi}} \sum_{t=1}^{v_{hij}} Y_{hikjlt} \quad \text{e} \quad X_{hik.l.} = 14 \sum_{j=1}^{M_{hi}} \sum_{t=1}^{v_{hij}} X_{hikjlt}$$

onde $Y_{hik.l.}$ é o número total de episódios de diarréia para crianças com característica *l* de domicílios de tipo *k* na *i*-ésima comunidade do *h*-ésimo estrato ou condição de saneamento e $X_{hik.l.}$ é o número total de dias observados para as crianças com característica *l* de domicílios de tipo *k* na *i*-ésima comunidade do *h*-ésimo estrato ou condição de saneamento

Também para cada estrato *h*, os vetores $(Y_{hi1.1.}, X_{hi1.1.}, \ldots, Y_{hiK.L.}, X_{hiK.L.})$ são independentes e identicamente distribuídos como conseqüência do método de amostragem aleatório simples, com reposição, assumido para os conglomerados. Então a razão de médias para a medida epidemiológica de interesse é:

$$R_{hkl} = \frac{\sum_{i=1}^{N_h} Y_{hik.l.}/N_h}{\sum_{i=1}^{N_h} X_{hik.l.}/N_h} = \frac{\overline{Y}_{hkl}}{\overline{X}_{hkl}} . \qquad (2.37)$$

A quantidade R_{hkl} é a densidade de incidência de diarréia, ou número de episódios por crianças-dia em crianças com a característica *l* do domicílio do tipo *k* para a comunidade com a característica *h*. Por exemplo, R_{hkl} pode ser a incidência de diarréia para crianças do sexo masculino cujo domicílio tem sanitário nas comunidades sem nenhuma intervenção de saneamento.

Abaixo é apresentada a formulação matricial para o cálculo das razões de médias para subgrupos de observações definidas pelo cruzamento dos níveis das características das comunidades, domicílios e crianças, simultaneamente.

Seja $\underset{\sim}{f}_{hikjlt} = (Y_{hikjlt}, X_{hikjlt})'$ uma observação bivariada discreta onde Y_{hikjlt} e X_{hikjlt} são respostas discretas descritas acima.

Sejam

$$\underset{\sim}{f}_{hikl} = \sum_{j} \sum_{t} (Y_{hikjlt}, X_{hikjlt})',$$

$\underset{\sim}{f}_{hi} = (\underset{\sim}{f}_{hi1}, \underset{\sim}{f}_{hi2}, \ldots, \underset{\sim}{f}_{hiKL})$ e

$$\overline{\underset{\sim}{f}}_{h} = \frac{\sum_{i=1}^{N_h} \underset{\sim}{f}_{hi}}{N_h}.$$

Para cada h, os $\underset{\sim}{f}_{hi}$ são independentes e identicamente distribuídos como conseqüência do método de amostragem assumido aleatório simples, com reposição, para os conglomerados. Um estimador não viciado da matriz de variância-covariância de $\overline{\underset{\sim}{f}}_{h}$ é dado por:

$$\underset{\sim}{v}_{\overline{\underset{\sim}{f}}_{h}} = \frac{1}{N_h(N_h-1)} \sum_{i=1}^{N_h} (\underset{\sim}{f}_{hi} - \overline{\underset{\sim}{f}}_{h})(\underset{\sim}{f}_{hi} - \overline{\underset{\sim}{f}}_{h})'. \qquad (2.38)$$

Se $\overline{\underset{\sim}{f}} = (\overline{\underset{\sim}{f}}_{1}, \overline{\underset{\sim}{f}}_{2}, \ldots, \overline{\underset{\sim}{f}}_{H})'$,

então

$$\underset{\sim}{v}_{\overline{\underset{\sim}{f}}} = BL(\underset{\sim}{v}_{\overline{\underset{\sim}{f}}_{h}}). \qquad (2.39)$$

onde $BL()$ é uma matriz bloco diagonal, e $\underset{\sim}{v}_{\overline{\underset{\sim}{f}}_{h}}$ são os blocos com seus elementos na diagonal.

Portanto, o vetor de razões de médias , ou ainda o vetor de incidência de diarréia, resultante da classificação cruzada dos níveis das características das comunidades, domicílios e crianças pode ser escrito como:

$$\underset{\sim}{R} = \underset{\sim}{R}_{hkl} = exp\, \underset{\sim}{A}\, ln\, \overline{\underset{\sim}{f}}, \qquad (2.40)$$

onde $\underset{\sim}{A} = [1 -1] \otimes I_{HKL}$ e I_{HKL} é a matriz identidade HKL x HKL e $\otimes$ é o símbolo para o produto Kronecker.

Usando aproximação por série de Taylor de primeira ordem determina-se a matriz de variância-covariância estimada de $\underset{\sim}{R}$, ou seja,

$$\underset{\sim}{V}_{\underset{\sim}{R}} \cong \underset{\sim}{H}\underset{\sim}{V}_{\underset{\sim}{f}}\underset{\sim}{H}' \quad \text{onde} \quad \underset{\sim}{H} = \underset{\sim}{D}_{\underset{\sim}{R}}\underset{\sim}{A}\underset{\sim}{D}_{\underset{\sim}{f}}^{-1} \tag{2.41}$$

e $\underset{\sim}{D}_{\underset{\sim}{R}}$ e $\underset{\sim}{D}_{\underset{\sim}{f}}$ são matrizes diagonais com os elementos de $\underset{\sim}{R}$ e $\overline{\underset{\sim}{f}}$ como seus respectivos elementos diagonais.

2.5 Regressão ponderada com razão de médias

Comparações de razões de médias correspondentes ao cruzamento dos níveis das características das comunidades, domicílios e crianças podem ser examinadas. Além disso, alguns desses níveis das características citadas acima podem ser correlacionados. Portanto, o estudo simultâneo dos efeitos dos níveis das características dos mesmos pode ser feito utilizando um procedimento de regressão, considerando inclusive interações entre duas ou mais características de interesse.

O método de regressão para as razões de médias (ou funções das razões de médias) é possível utilizando a metodologia de Grizzle, Starmer & Koch (1969) (GSK) e extensões desta, como discutido em Koch *et al.* (1985). Este método de regressão será apresentado nas próximas seções deste capítulo.

2.5.1 Um modelo linear para R

Seja $\underset{\sim}{R} = (R_{111}, \ldots, R_{HKL})'$ um vetor de razões de médias para sub-populações definidas pela classificação cruzada dos níveis das características que representam as comunidades, os domicílios e as crianças, ou seja, um vetor de dimensão HxKxL, onde H representa o número de estratos formados por uma característica das comunidades, K o número de níveis de uma característica do domicílio e L o número de níveis de uma característica da criança. Um modelo linear para $\underset{\sim}{R}$ é dado por :

$$E_A(\underset{\sim}{R}) = \underset{\sim}{X}\underset{\sim}{\beta}, \tag{2.42}$$

onde $E_A()$ denota o valor da esperança assintótica, $\underset{\sim}{X}$ a matriz de planejamento HKL x μ , e $\underset{\sim}{\beta}$ um vetor μ x 1 de parâmetros desconhecidos a serem estimados. A matriz de variância-covariância de $\underset{\sim}{R}$, $\underset{\sim}{V}_{\underset{\sim}{R}}$, deve ser estimada consistentemente por (2.41). O estimador de mínimos quadrados ponderados $\underset{\sim}{b}$ de $\underset{\sim}{\beta}$ é dado por:

$$\underset{\sim}{b} = (\underset{\sim}{X}' \underset{\sim}{V}_{\underset{\sim}{R}}^{-1} \underset{\sim}{X})^{-1}(\underset{\sim}{X}' \underset{\sim}{V}_{\underset{\sim}{R}}^{-1} \underset{\sim}{R}). \tag{2.43}$$

Em grandes amostras, o estimador $\underset{\sim}{b}$ tem aproximadamente uma distribuição normal com

$$E_A(\underset{\sim}{b}) = \underset{\sim}{\beta}, \tag{2.44}$$

e a correspondente matriz de covariância para cada estimador consistente é

$$\underset{\sim}{V}_{\underset{\sim}{b}} = (\underset{\sim}{X}' \underset{\sim}{V}_{\underset{\sim}{R}}^{-1} \underset{\sim}{X})^{-1}. \tag{2.45}$$

O ajuste do modelo em (2.42) pode ser verificado através da estatística de mínimos quadrados ponderado dos resíduos

$$Q = (\underset{\sim}{R} - \underset{\sim}{X}\underset{\sim}{b})' \underset{\sim}{V}_{\underset{\sim}{R}}^{-1} (\underset{\sim}{R} - \underset{\sim}{X}\underset{\sim}{b}). \qquad (2.46)$$

Quando o modelo em (2.42) é válido, Q tem aproximadamente uma distribuição qui-quadrado, com $(HKL-\mu)$ graus de liberdade, assumindo que todos os HKL subgrupos apresentam tamanhos de amostras grandes.

2.5.2 Um modelo linear para *log(R)*

Seja $\underset{\sim}{F} = (F_{111}, \ldots, F_{HKL})' = (\underset{\sim}{log}\, R_{111}, \ldots, \underset{\sim}{log}\, R_{HKL})' = \underset{\sim}{log}\, \underset{\sim}{R}$ um vetor do log das razões de médias correspondendo às sub-populações definidas pelo cruzamento dos níveis H, K e L das características dos conglomerados, dos elemento e dos indivíduos, respectivamente. Uma estimativa para a matriz de variância-covariância de $\underset{\sim}{F}$, baseada na série de Taylor, é dada por:

$$\underset{\sim}{V}_{\underset{\sim}{F}} = \underset{\sim}{D}_{\underset{\sim}{R}}^{-1} \underset{\sim}{V}_{\underset{\sim}{R}} \underset{\sim}{D}_{\underset{\sim}{R}}^{-1}, \qquad (2.47)$$

onde $\underset{\sim}{D}_{\underset{\sim}{R}}$ é uma matriz HKL x HKL, com os elementos do vetor $\underset{\sim}{R}$ na diagonal, e $\underset{\sim}{V}_{\underset{\sim}{R}}$ é a estimativa da matriz de covariância para $\underset{\sim}{R}$ dada em (2.41). O vetor de funções $\underset{\sim}{F}$ pode ser ajustado usando a metodologia GSK, descrita na seção anterior.

2.6 Razão de médias para resposta discreta de uma amostragem de conglomerados em dois estágios

A seção 2.4 trata de um método estatístico para analisar respostas discretas bivariadas selecionadas de uma amostra de conglomerados em um estágio. Ou seja, todos os elementos de cada conglomerado selecionado são incluídos na análise. É claro que isto só é possível quando não

existe um alto custo extra na coleta e manutenção das informações de todos esses elementos. Esta seção abordará a metodologia de análise de respostas discretas bivariadas selecionadas de uma amostragem por conglomerado em dois estágios. Isto é, cada conglomerado selecionado na amostra é sub-amostrado. Modificações do método descrito anteriormente surgirão devido à necessidade de técnicas analíticas complexas para análise de conglomerados sub-amostrados. Por exemplo, será usado um total amostral ponderado para estimar os totais dos conglomerados. O uso de ponderação evita o vício da estimativa do total. Como anteriormente, serão abordados a questão da amostragem em dois estágios com estratificação, com a formulação matricial para a razão de médias e um modelo log-linear utilizando mínimos quadrados ponderados. Um breve resumo da amostragem por conglomerado em 2 estágios juntamente com o estimador adotado para este trabalho encontra-se no apêndice A.

2.6.1 Definição da razão de médias para resposta discreta

O processo de amostragem de conglomerados em dois estágios reflete a proposta do estudo AISAM, onde as comunidades foram selecionadas por uma amostragem aleatória simples, com reposição, no primeiro estágio, e os domicílios foram selecionados pela amostragem aleatória simples, sem reposição, no segundo estágio. E assim foram coletadas as informações de todas as crianças de 0 a 5 anos no domicílio amostrado.

Sejam $i = 1,2, \ldots, N$ o índice referente às comunidades selecionadas, $j = 1,2, \ldots, n_i$ o índice dos domicílios selecionados e $t = 1,2, \ldots, v_{ij}$ o índice das observações de crianças para o domicílio j na comunidade i. Assim N representa o número de comunidades selecionadas ($N = 9$), n_i representa o número de domicílios amostrados na i-ésima comunidade (em torno de 120 domicílios por comunidade) e v_{ij} o número de crianças no domicílio j na comunidade i (cerca de 130 por comunidade).

Define-se

Y_{ijt2} número de episódios de diarréia para a t-ésima criança do j-ésimo domicílio amostrado da i-ésima comunidade selecionada;

X_{ijt2} número de quinzenas observadas para a t-ésima criança do j-ésimo domicílio amostrado da i-ésima comunidade selecionada.

O subíndice 2, em Y_{ijt2} e X_{ijt2}, enfatiza que essas variáveis aleatórias resultam de um processo de amostragem de conglomerado em dois estágios.

Define-se também

$$\overline{Y}_{i..2} = \frac{\sum_{j=1}^{n_i}\left(\sum_{t=1}^{v_{ij}} Y_{ijt2}\right)}{n_i} \quad \text{e} \quad \overline{X}_{i..2} = \frac{14\sum_{j=1}^{n_i}\left(\sum_{t=1}^{v_{ij}} X_{ijt2}\right)}{n_i}$$

como sendo o número médio de episódios de diarréia para a i-ésima comunidade e número médio de dias observados para a i-ésima comunidade, respectivamente. As medidas $\overline{Y}_{i..2}$ e $\overline{X}_{i..2}$ são provenientes de uma amostragem de conglomerado em dois estágios, portanto devem ser acrescidas de um peso de amostragem a fim de evitar o vício em tais medidas, pois os conglomerados podem diferir muito quanto aos seus tamanhos.

Assim,

$$\overline{Y}_{i..2w} = w_i \overline{Y}_{i..2} \quad , \quad \overline{X}_{i..2w} = w_i \overline{X}_{i..2} \ ^{1}$$

e $w_i = \dfrac{v_i}{\varphi_i v_+}$ é o peso de amostragem onde v_i é o número total de elementos do i-ésimo conglomerado selecionado, v_+ é o número total de elementos na população e φ_i é a probabilidade de selecionar o i-ésimo conglomerado. Esse peso de amostragem serve para corrigir o estimador da média por conglomerado tornando-o um estimador não-viciado para a média geral.

---

[1] Esses estimadores são idênticos ao estimador $\overline{z}_i$ apresentado no apêndice A

Como o método de amostragem para os conglomerados é aleatório simples, com reposição; $(\overline{Y}_{i..2w}, \overline{X}_{i..2w})$ são independentes e identicamente distribuídos. Assim o estimador da razão de médias dos conglomerados (média geral) é :

$$R_{2w} = \frac{\sum_{i=1}^{N} \overline{Y}_{i..2w}}{\sum_{i=1}^{N} \overline{X}_{i..2w}} = \frac{\overline{Y}_{2w}}{\overline{X}_{2w}} \tag{2.48}$$

R_{2w} é uma estimativa do número médio de episódios de diarréia por crianças-dia de acompanhamento.

Utilizando a notação matricial, R_{2w} pode ser formulado da seguinte forma

$$f_{ijt2} = (Y_{ijt2}, X_{ijt2})'$$

$$\underset{\sim}{\overline{f}}_{i2} = \sum_{j=1}^{n_i} \frac{1}{n_i} \sum_{t=1}^{m_{ij}} \underset{\sim}{f}_{ijt2}$$

$$\underset{\sim}{\overline{f}}_{i2w} = w_i \underset{\sim}{\overline{f}}_{i2} \quad \text{e}$$

$$\underset{\sim}{\overline{f}}_{2w} = \frac{\sum_{i=1}^{n} \underset{\sim}{\overline{f}}_{i2w}}{N}.$$

Então R_{2w} é dado por

$$R_{2w} = exp\, \underset{\sim}{A}\, ln\, \underset{\sim}{\overline{f}}_{2w} \quad \text{onde } \underset{\sim}{A} = [1,-1]. \tag{2.49}$$

A representação de R_{2w} em série de Taylor da média populacional $\mu = (\mu_y, \mu_x)'$ de $(\overline{Y}_{2w}, \overline{X}_{2w})$ é dada por

$$R_{2w} = \frac{\mu_Y}{\mu_X} + \frac{1}{\mu_X}(\overline{Y}_{2w} - \mu_Y) - \frac{\mu_Y}{\mu_{X^2}}(\overline{X}_{2w} - \mu_X) + O(1/N). \tag{2.50}$$

A esperança assintótica da representação de R_{2w} na série de Taylor de primeira ordem, $E_A(R)$ é $\frac{\mu_y}{\mu_x} = \theta$, a razão de médias na população. O estimador da variância de R_{2w} baseado na linearização em série de Taylor, é dado por:

$$v(R_{2w}) = \underset{\sim}{D}_{R_{2w}} \underset{\sim}{A} \underset{\sim}{D}^{-1}_{\bar{f}_{2w}} \underset{\sim}{V}_{\bar{f}_{2w}} \underset{\sim}{D}^{-1}_{\bar{f}_{2w}} \underset{\sim}{A}^{-1} \underset{\sim}{D}_{R_{2w}} \tag{2.51}$$

onde,

$$\underset{\sim}{V}_{\bar{f}_{2w}} = \frac{1}{N(N-1)} \sum_{i=1}^{N} (\underset{\sim}{\bar{f}}_{i2w} - \underset{\sim}{\bar{f}}_{2w})(\underset{\sim}{\bar{f}}_{i2w} - \underset{\sim}{\bar{f}}_{2w})' . \tag{2.52}$$

Observa-se que a expressão (2.52) depende de w_i. Mais especificamente, esta expressão depende do conhecimento do número total de elementos no conglomerado i (v_i), da probabilidade de selecionar o conglomerado i (φ_i) e do número total de elementos na população ($v_.$). Na prática algumas dessas quantidades não são conhecidas. Certos esquemas de amostragem evitam a necessidade de conhecer alguma ou todas essas quantidades para análise. Duas opções serão discutidas à seguir.

Seja $\varphi_i = \frac{1}{\Gamma}$ a probabilidade de seleção de conglomerados dentro da amostra usando amostragem aleatória simples com reposição no primeiro estágio, e π uma probabilidade constante de selecionar elementos dentro de cada conglomerado no segundo estágio. Assim, a probabilidade de seleção comum para cada elemento na população é dada por $\varphi = \frac{1}{\Gamma}\pi$, onde π é uma probabilidade fixa de selecionar um elemento em cada conglomerado, garantindo assim que os conglomerados estejam igualmente representados na amostra, isto é, π é uma probabilidade proporcional ao tamanho dos conglomerados, $\pi = \frac{n_i}{v_i}$.

Portanto

$$\varphi = \frac{1}{\Gamma}\frac{n_i}{\nu_i} \quad \text{onde} \quad n_i = \pi\nu_i$$

isto implica que $w_i = \dfrac{\nu_i}{\varphi_i \nu_+} = \dfrac{n_i \Gamma}{\pi \nu_+} \quad \left(\varphi_i = \dfrac{1}{\Gamma}\right)$, então

$$R_{2w} = \frac{\overline{Y}_{2w}}{\overline{X}_{2w}} = \frac{\sum_{i=1}^{N}\left\{\dfrac{n_i\Gamma}{\pi\nu_+}\sum_{j=1}^{n_i}\left(\sum_{t=1}^{\nu_{ij}} Y_{ijt2} \Big/ n_i\right)\right\}}{\sum_{i=1}^{N}\left\{\dfrac{n_i\Gamma}{\pi\nu_+}\sum_{j=1}^{n_i}\left(\sum_{t=1}^{\nu_{ij}} X_{ijt2} \Big/ n_i\right)\right\}} = \frac{\dfrac{\Gamma}{\pi\nu_+}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{t=1}^{\nu_{ij}} Y_{ijt2}}{\dfrac{\Gamma}{\pi\nu_+}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{t=1}^{\nu_{ij}} X_{ijt2}}$$

ou,

$$R = \frac{\sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{t=1}^{\nu_{ij}} Y_{ijt2}}{\sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{t=1}^{\nu_{ij}} X_{ijt2}}. \qquad (2.53)$$

A expressão acima é similar a expressão da seção 2.4.3, exceto que aqui a soma sobre j inclui somente indivíduos selecionados ao acaso dentro da amostra no segundo estágio. Logo a estimativa de R é a mesma descrita na seção 2.4, embora a matriz de variância-covariância dependa do número total de elementos dentro de cada conglomerado, ν_i.

Neste esquema, contudo, deve-se conhecer ν_i e além disso, atribuir um valor para π, sendo esta sua principal desvantagem.

Um outro esquema de amostragem que evita a especificação de quantidades populacionais na análise é o planejamento "alto-ponderado". Seja $\varphi_M = \dfrac{\nu_M}{\nu_+}$ a probabilidade de selecionar um conglomerado M no primeiro estágio. Isto é, os conglomerados são selecionados proporcionalmente ao tamanho. Seja também n um número de elementos constante selecionado de cada conglomerado i amostrado, i =1,2, ... ,N, no segundo estágio.

Então a probabilidade comum de seleção para cada indivíduo é dada por $\varphi = \varphi_i \pi = \frac{\nu_\gamma}{\nu_+} \frac{n}{\nu_\gamma}$,

mas $w_i = \frac{\nu_i}{\varphi_i \nu_+} = \frac{\nu_i}{\frac{\nu_i}{\nu_+} \nu_i} = 1$ ($\varphi_M = \varphi_i$ pois um conglomerado $M = i$ foi selecionado).

Neste esquema é necessário que se conheça o tamanho de cada conglomerado a fim de determinar φ_M.

Portanto

$$R_{2w} = R = \frac{\overline{Y}}{\overline{X}} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{n_i} \sum_{t=1}^{\nu_{ij}} Y_{ijt2}}{\sum_{i=1}^{N} \sum_{j=1}^{n_i} \sum_{t=1}^{\nu_{ij}} X_{ijt2}}, \qquad (2.54)$$

e os métodos desenvolvidos na seção 2.4.3 podem ser usados para conglomerados sub-amostrados sem modificação.

2.6.2 Subgrupos de razões de médias para resposta discreta definidos pelos níveis das características dos conglomerados, das subunidades dentro do conglomerado e das unidades de análise

A proposta de dois estágios se aplica à base de dados em questão. Seja então:

h =1,2, ... , H o índice referente aos estratos formados pela condição de saneamento das comunidades;

k =1,2, ... , K o índice dos níveis da característica do domicílio, por exemplo, presença ou não de sanitário no domicílio;

l =1,2, ... , L o índice dos níveis da característica da criança;

i =1,2, ... , N_h índice dos conglomerados (comunidades) amostrais no h-ésimo estrato;

j =1,2, ... , n_{hi} o índice dos domicílios selecionados na i-ésima comunidade do h-ésimo estrato;

t =1,2, ... , v_{hij} índice das crianças do j-ésimo domicílio na i-ésima comunidade do h-ésimo estrato.

Onde N_h é o número de comunidades selecionadas no h-ésimo estrato, n_{hi} é o número de domicílios selecionados na i-ésima comunidade do h-ésimo estrato, H é o número referente as condições de saneamento das comunidades, K é o número níveis de uma característica dos domicílios e L é o número de níveis de uma característica da criança.

Define-se

$Y_{hikjlt2}$ número de episódios de diarréia para a t-ésima criança com característica l do j-ésimo domicílio do tipo k na i-ésima comunidade do estrato h (resultante de uma amostragem de conglomerado em 2 estágios);

$X_{hikjlt2}$ número de quinzenas observadas para a t-ésima criança com característica l do j-ésimo domicílio do tipo k na i-ésima comunidade do estrato h.

Sejam,

$$\underset{\sim}{\overline{f}}_{hikl2} = \sum_{j=1}^{n_{hi}} \frac{1}{n_{hi}} \sum_{t=1}^{v_{hij}} (Y_{hikjlt2},\ 14X_{hikjlt2})',$$

$$\underset{\sim}{\overline{f}}_{hi2} = (\underset{\sim}{\overline{f}}'_{hi112}, \underset{\sim}{\overline{f}}'_{hi122}, \ldots, \underset{\sim}{\overline{f}}'_{hiKL2})',$$

$$\underset{\sim}{\overline{f}}_{hi2w} = w_{hi}\underset{\sim}{\overline{f}}_{hi2} \quad \text{e} \quad \underset{\sim}{\overline{f}}_{h2w} = \frac{\sum_{i=1}^{N_h} \underset{\sim}{\overline{f}}_{hi2w}}{N_h} \quad \text{onde} \quad w_{hi} = \frac{v_{hi}}{\varphi_{hi} v_{+}}.$$

É importante observar que o peso de amostragem agora se refere não somente ao conglomerado selecionado como também ao estrato que o mesmo pertence, não alterando,

entretanto, sua forma de calculá-lo (na aplicação os estratos estão igualmente representados, pois o número de conglomerados selecionados dentro dos estratos foi uniforme).

Um estimador consistente para a variância de $\underset{\sim}{\overline{f}}_{h2w}$ é dado por

$$\underset{\sim}{v}_{\overline{f}_{h2w}} = \frac{1}{N_h(N_h-1)}\sum_{i=1}^{N_h}(\underset{\sim}{\overline{f}}_{hi2w} - \underset{\sim}{\overline{f}}_{h2w})(\underset{\sim}{\overline{f}}_{hi2w} - \underset{\sim}{\overline{f}}_{h2w})' . \qquad (2.55)$$

Sejam também

$$\underset{\sim}{\overline{f}}_{2w} = (\underset{\sim}{\overline{f}}_{12w}', \ldots, \underset{\sim}{\overline{f}}_{H2W}')' \qquad (2.56)$$

e

$$\underset{\sim}{v}_{\overline{f}_{2W}} = BL\left(\underset{\sim}{v}_{\overline{f}_{h2W}}\right).$$

Assim, o vetor de razões de médias resultante da classificação cruzada das características das comunidades, domicílios e crianças pode ser escrito como

$$\underset{\sim}{R}_{2w} = exp\ \underset{\sim}{A}\ ln\ \underset{\sim}{\overline{f}}_{2W} \qquad (2.57)$$

onde $\underset{\sim}{A} = [1 -1] \otimes I_{HKL}$ e I é a matriz identidade HKL x HKL e $\otimes$ é o símbolo para o produto Kronecker.

Usando expansão em série de Taylor de primeira ordem determina-se a variância estimada de $\underset{\sim}{R}_{2w}$, ou seja,

$$\underset{\sim}{V}_{\underset{\sim}{R}_{2w}} \cong \underset{\sim}{H}\underset{\sim}{V}_{\underset{\sim}{\overline{f}}_{2w}}\underset{\sim}{H}' \quad \text{onde} \quad \underset{\sim}{H} = \underset{\sim}{D}_{\underset{\sim}{R}_{2w}}\ \underset{\sim}{A}\ \underset{\sim}{D}^{-1}_{\underset{\sim}{\overline{f}}_{2w}} . \qquad (2.58)$$

2.7 Regressão ponderada com razão de médias para resposta discreta

O exame simultâneo dos efeitos dos níveis das características das comunidades, domicílios e crianças na razão de médias pode ser realizado através do procedimento de regressão, além da condição de avaliar interações entre duas ou mais algumas dessas características. É importante frisar que a combinação de duas ou mais características a nível de domicílio ou criança pode ser usada para representar um nível na construção do vetor de razões.

2.7.1 Modelo para *log(R)*

Seja $\underset{\sim}{F} = (F_{111}, \ldots, F_{HKL})' = \underset{\sim}{log}\,\underset{\sim}{R}$ um vetor do log das razões de médias, correspondendo às sub-populações definidas pelo cruzamento dos níveis dos H tipos de estrato, k tipos de domicílios e L características das crianças. Uma estimativa para a matriz de covariância de $\underset{\sim}{F}$,baseada na série de Taylor, é dada por:

$$\underset{\sim}{V}_{\underset{\sim}{F}} = \underset{\sim}{D}^{-1}_{\underset{\sim}{R}} \underset{\sim}{V}_{\underset{\sim}{R}} \underset{\sim}{D}^{-1}_{\underset{\sim}{R}} \qquad (2.59)$$

onde $\underset{\sim}{D}_{\underset{\sim}{R}}$ é uma matriz HKL x HKL com os elementos do vetor $\underset{\sim}{R}$ na diagonal, e $\underset{\sim}{V}_{\underset{\sim}{R}}$ é a estimativa da matriz de covariância para $\underset{\sim}{R}$ dada em (2.41). O vetor de funções $\underset{\sim}{F}$ pode ser ajustado utilizando a metodologia GSK, junto com sua matriz de covariância descrito nas seções 2.5.1 e 2.5.2.

Dependendo do tipo de vetor de razão que o pesquisador tem interesse pode-se modelar o logito de R. Assim, $\underset{\sim}{F} = (F_{111}, \ldots, F_{HKL})' = \underset{\sim}{logit}\,\underset{\sim}{R} = \underset{\sim}{ln}(\underset{\sim}{R}) - \underset{\sim}{ln}(1 - \underset{\sim}{R})$ e a variância neste caso seria $\underset{\sim}{V}_{\underset{\sim}{F}} = \underset{\sim}{D}^{-1}_{\underset{\sim}{R}} [\underset{\sim}{D}_{1-\underset{\sim}{R}}]^{-1} \underset{\sim}{V}_{\underset{\sim}{R}} [\underset{\sim}{D}_{1-\underset{\sim}{R}}]^{-1} \underset{\sim}{D}^{-1}_{\underset{\sim}{R}}$.

A seguir, no próximo capítulo é apresentada a metodologia das EEG.

Capítulo III

As Equações de Estimação Generalizadas, Extensões e Diagnóstico

3.1 Introdução

Em estudos longitudinais, nos quais a variável resposta é contínua, o método de análise comumente usado envolve modelos lineares com erros correlacionados. Já com resposta discreta contendo dependência, poucas técnicas são disponíveis devido, em parte, a falta de uma distribuição discreta multivariada tão flexível quanto a normal multivariada, que é a base da teoria de modelos lineares. A dificuldade fundamental na análise de dados longitudinais discretos refere-se às interpretações dos coeficientes nos modelos de regressão, pois estas são vinculadas a suposições a respeito da natureza da dependência no tempo, ou seja, diferentes suposições a respeito da fonte de correlação levam a distintas interpretações. O mesmo não acontece com modelos lineares, pois embora a estimação dos parâmetros nesse modelo deva considerar a correlação dos dados, as interpretações dos coeficientes de regressão são independentes da estrutura de correlação adotada. Segundo Diggle *et al.* (1994), existem três

estratégias com alicerce em modelos lineares generalizados para analisar dados longitudinais: modelo marginal, de efeitos aleatórios e condicional ou transicional. A escolha do procedimento a ser utilizado deve ser cuidadosa, pois depende do objetivo do estudo, do tipo de desenho, da natureza das variáveis, bem como a escolha da fonte de correlação.

Com os modelos de efeitos aleatórios e transicional é possível estimar os parâmetros desconhecidos na regressão usando o método de máxima verossimilhança ou até máxima verossimilhança condicional. Com dados normais, os dois primeiros momentos determinam completamente a verossimilhança na modelagem marginal, porém, o mesmo não acontece com outros membros da classe de modelos lineares generalizados. Nesse caso, para determinar a verossimilhança completa são necessárias suposições adicionais de momentos de ordem superior, porém, a verossimilhança torna-se intratável devido à presença de muitos parâmetros de perturbação. Por esta razão, uma solução em problemas desse tipo é usar o método de Equações de Estimação Generalizadas (EEG) que especifica somente a distribuição marginal e uma matriz de variância-covariância de trabalho para o vetor de observações repetidas.

Liang & Zeger (1986) e Zeger & Liang (1986) apresentaram um procedimento unificado para respostas contínuas e discretas, baseado nas EEG para a estimação dos parâmetros de regressão, sem a especificação da verossimilhança completa e sim com suposições do comportamento dos parâmetros de interesse. Esta metodologia é uma extensão multivariada da quase-verossimilhança, onde somente a relação entre a função do valor esperado da resposta e o preditor linear (função de ligação) e a relação entre a variância e a média, além do parâmetro de escala, necessitam ser especificadas. Esta técnica produz estimativas consistentes e assintoticamente normais sob a especificação correta da função de ligação.

Prentice (1988) formalizou uma extensão das EEG criando equações de estimação para o parâmetro de correlação, permitindo assim que o mesmo fosse avaliado em função das covariáveis. Zhao & Prentice (1990) e Prentice & Zhao (1991) identificaram uma classe de modelos exponenciais quadráticos para a qual uma EEG particular é a equação escore. Na verdade este é um método alternativo para modelar parâmetros marginais de uma distribuição binária multivariada através de uma transformação dos parâmetros de uma distribuição

condicional em parâmetros marginais.

Liang *et al.* (1992) especificaram a dependência do tempo em termos da razão de chances marginal, descrevendo um conjunto equivalente de equações para estimação dos parâmetros da média e associação marginal simultaneamente. Eles denominaram de equações de estimação generalizadas de segunda ordem (EEG2). A medida que o tamanho do conglomerado aumenta, as EEG2 tornam-se menos atrativas computacionalmente e, como caminho alternativo, Carey *et al.* (1993) desenvolveram um procedimento chamado Regressão Logística Alternada (RLA), que caracteriza a dependência em termos da razão de chances e usa as EEG de primeira ordem para estimar os parâmetros de regressão.

Uma breve ilustração de técnicas de diagnóstico nas equações de estimação generalizadas será apresentada com objetivo de medir a influência do conglomerado ou das observações sobre as estimativas dos parâmetros do modelo de regressão.

A metodologia das EEG, juntamente com algumas de suas extensões e diagnóstico, será abordada no decorrer deste capítulo.

3.2 Fundamentos

Fisher (1921) propôs a noção de verossimilhança como um método de estimação e como um critério para comparar duas hipóteses rivais a serem testadas.

A verossimilhança desempenha um papel central nos estudos teóricos de inferência estatística. Entretanto, em muitas situações práticas, esta metodologia tradicional não é satisfatória e pode até mesmo apresentar resultados falaciosos, como por exemplo quando o número de parâmetros de perturbação é da ordem de grandeza do número de observações. Para contornar situações como esta, surgiram extensões da definição de verossimilhança, as chamadas pseudo-verossimilhanças, que podem ser mais satisfatórias na inferência de problemas complexos. Essas extensões tem como principal objetivo reduzir a dimensão do vetor de parâmetros de perturbação.

Um tipo de pseudo-verossimilhança que não exige o conhecimento da função de distribuição dos dados e sim a noção de relações funcionais para os seus dois primeiros

momentos é a quase-verossimilhança. A importância da teoria de quase-verossimilhança é que se pode estimar parâmetros, realizar testes de hipóteses e construir regiões de confiança de forma similar à teoria de verossimilhança, supondo porém, restrições mais fracas para os dados, ou seja, considerando apenas relações funcionais para os seus dois primeiros momentos no lugar de uma família de distribuição (Cordeiro, 1992).

Outro método de estimação amplamente difundido e aplicável sob suposições dos dois primeiros momentos é o de mínimos quadrados. É sabido que no caso da família normal as estimativas de máxima verossimilhança e de mínimos quadrados coincidem.

Godambe (1991) mostrou que as semelhanças entre os métodos de estimação: máxima verossimilhança, mínimos quadrados e variância *mínima* não-viciada, quando estudados completamente, atendem à uma teoria de estimação na qual combina-se o poder desses três métodos e ao mesmo tempo eliminam-se suas limitações. Esta teoria, foi denominada de teoria de Função de Estimação.

Uma função de estimação g é uma função dos dados e de parâmetros desconhecidos θ de interesse. Essa função é construída de modo que sua raiz, quando existe, é uma estimativa do parâmetro envolvido, isto é, $g(y,\hat{\theta}) = 0$. Liang & Zeger (1995) argumentam que as funções de estimação são uma ferramenta útil para minimizar a influência dos parâmetros de perturbação, isto é, uma estratégia alternativa a qual permite combinar os dados e parâmetros desconhecidos em uma função de estimação, especificando somente parte da distribuição de probabilidade. Por exemplo: no modelo logístico linear para dados longitudinais, necessita-se especificar somente a média e a covariância das observações repetidas para cada indivíduo numa função de estimação. Godambe (1960) estabeleceu um critério de otimalidade para as funções de estimação pertencentes a uma classe de funções de estimação não-viciada ($E[g(y,\hat{\theta})] = 0 \; \forall\theta$) com objetivo de que as raízes dessas funções possuíssem boas propriedades assintóticas.

Quando se fala em Modelos Lineares Generalizados sabe-se que a estimação dos parâmetros de interesse pela função de verossimilhança é possível. Esta estratégia de estimação é ótima no sentido que a solução das equações escores tem variância *mínima* assintótica entre todas as estimativas que são obtidas de equações de estimação não-viciadas. Ainda no contexto

de modelos lineares generalizados, caso a restrição de que a distribuição marginal dos dados pertença à família exponencial seja relaxada de maneira que a variância possa ser uma função arbitrária da média, tem-se então os modelos de quase-verossimilhança introduzidos por Wedderburn (1974). Para estes modelos, uma equação quase-escore é deduzida via função de quase-verossimilhança (McCullagh & Nelder, 1989) e os estimadores obtidos como solução dessa função quase-escore são ótimos no sentido que eles têm variância mínima entre a classe de estimadores lineares não-viciados. Sendo assim, a função quase-escore é uma função de estimação ótima no sentido de Godambe (1960).

Seja y_i a resposta do i-ésimo entre k indivíduos, e x_{ij} o valor da j-ésima variável explanatória para o i-ésimo indivíduo, $j = 1,2,\ldots,p$. O Modelo Linear Generalizado (MLG) tem como objetivo descrever a dependência da resposta média $\mu_i = E(y_i)$ sobre as variáveis explanatórias. Modelos de Regressão Poisson, Logístico e Linear são todos casos especiais de modelos lineares generalizados que dividem as seguintes características: Primeiro, a esperança de y_i, μ_i, é assumida estar relacionada ao vetor de p covariáveis $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ através de uma função de ligação, monótona e diferenciável, isto é,

$$h(\mu_i) = x_i'\beta. \tag{3.1}$$

Segundo, a variância de y_i é uma função especificada de sua média na forma

$$V(y_i) = v_i = \phi\, v(\mu_i), \tag{3.2}$$

onde $v(.)$ é denominada função de variância e ϕ, fator de escala, é uma constante conhecida para alguns membros da família de modelos lineares generalizados, enquanto para outros é um parâmetro adicional a ser estimado.

Terceiro, cada classe de modelos lineares generalizados corresponde a um membro da família de distribuição exponencial, com uma função de densidade da forma

$$f(y,\theta_i) = exp\{[y\theta_i - a(\theta_i)]/\phi + c(y,\phi)\}. \tag{3.3}$$

O parâmetro θ_i é conhecido como parâmetro natural e está relacionado a μ_i, por $\mu_i = \dfrac{\partial a(\theta_i)}{\partial \theta_i}$ e $V(y_i) = \dfrac{\partial^2 a(\theta_i)}{\partial^2 \theta_i}\phi$.

Em qualquer modelo linear generalizado, os coeficientes de regressão β podem ser

estimados pela solução da mesma equação de estimação dada abaixo

$$U(\beta)=\sum_{i=1}^{k}\left(\frac{\partial \mu_i}{\partial \beta}\right)' v_i^{-1}[y_i-\mu_i(\beta)]=0. \qquad (3.4)$$

A solução $\hat{\beta}$, que é uma estimativa de máxima verossimilhança, pode ser obtida por *mínimos* quadrados reponderados iterativamente. Uma propriedade importante da família de modelos lineares generalizados é que a função escore $S(\beta)$ depende somente da média e variância de y_i. Wedderburn (1974) foi o primeiro a mostrar que a equação de estimação (3.4), equação escore, pode ser usada para estimar coeficientes de regressão para uma função de ligação e variância independentemente de corresponder a um membro particular da família exponencial. Ele denominou $S(\beta)$ de equação quase-escore. Pode ser difícil decidir qual deve ser a distribuição populacional, mas a forma da relação variância - média é muito mais fácil de ser postulada. Em dados longitudinais, um fator relevante é a *estrutura* de correlação existente no vetor de respostas repetidas e uma proposta que tem recebido considerável atenção nos últimos anos, já com diversas extensões, é das equações de estimação generalizadas (EEG) de Liang & Zeger (1986). Na verdade esta metodologia é uma extensão dos modelos de quase-verossimilhança que permite que a correlação dentro do conglomerado (conjunto de observações pertencentes a mesma unidade de análise) seja incorporada na estimação dos parâmetros de um modelo de regressão através do uso de uma matriz de correlação de trabalho.

3.3 Metodologia das EEG

Em muitos estudos longitudinais, a questão de interesse pode ser formulada no contexto de análise de regressão, onde relaciona-se a variável resposta com as variáveis explanatórias sem esquecer, entretanto, de considerar a estrutura de correlação existente das respostas repetidas. Em geral a escolha do modelo deve depender não apenas da natureza da variável e do tipo de desenho, mas também, do objetivo do estudo. Diferentes questões levam a diferentes

ajustes de modelos para o mesmo conjunto de dados. Alguns objetivos são satisfeitos pelo modelo de efeitos aleatórios, por exemplo, no qual os coeficientes de regressão medem a influência direta das variáveis explanatórias sobre as respostas para cada particular indivíduo. Outros objetivos são melhor alcançados pelo modelo transicional ou condicional, nos quais o efeito das covariáveis sobre a mudança individual é modelado através da resposta condicionada a respostas passadas. Já outros são atingidos pelo modelo marginal, que descreve como a resposta média entre indivíduos muda com as covariáveis, ou seja, quando os parâmetros de interesse são taxas ou médias populacionais. Zeger *et al.*(1988) deduziram relações existentes entre os parâmetros do modelo marginal e de efeitos aleatórios.

Baseado na utilidade para estudos epidemiológicos, na vantagem em se ajustar a regressão de Y sobre X, bem como a associação entre observações repetidas, este capítulo descreverá a modelagem marginal apoiando-se no uso das equações de estimação generalizadas (EEG) para estimar coeficientes de regressão, além da metodologia de regressão logística alternada para estimar a associação, através da razão de chances, considerando as medidas repetidas.

A metodologia EEG modela uma função conhecida da esperança marginal da variável dependente como uma função linear de uma ou mais covariáveis, e a correlação entre observações repetidas para um indivíduo é tratada como parâmetro de perturbação. Sob condições fracas de regularidade, o vetor de parâmetros estimados é assintoticamente não-viciado com distribuição normal para qualquer escolha da matriz de correlação de trabalho, e com variância assintoticamente dependendo de ambos padrões de covariância assumido ou verdadeiro. Este método confia na independência entre indivíduos para estimar consistentemente a variância dos estimadores propostos mesmo quando a estrutura de correlação é incorreta.

Para uma melhor ilustração dessa metodologia, considera-se y_{it}, como o valor da variável resposta do i-ésimo indivíduo no t-ésimo tempo, para i=1,2,...,k e t=1,2,...,n_i e $x_{it} = (x_{ij1}, x_{ij2}, \ldots, x_{ijp})'$ um vetor px1 dos valores das variáveis explanatórias (covariáveis) associadas com y_{it}. O primeiro passo no procedimento das EEG é relacionar a média marginal

$\mu_{it} = E(y_{it})$[1] a uma combinação linear das covariáveis x_{it} por uma função de ligação h, como segue:

$$h(\mu_{it}) = x_{it}'\beta, \tag{3.5}$$

onde β é um vetor px1 de parâmetros desconhecidos.

As escolhas mais comuns para as funções de ligação são: logit ou probit para respostas binárias, *log* para contagens, e ligação linear para respostas contínuas. Por exemplo, se a ocorrência ou não de diarréia é a resposta e idade em meses é uma covariável, então o *logit* da frequência de diarréia deve ser assumido como uma função linear da idade. O vetor β, px1, de parâmetros desconhecidos caracteriza como a distribuição da resposta depende das variáveis explanatórias. O segundo passo é descrever a variância de y_{it} como uma função da média, ou seja,

$$Var(y_{it}) = g(\mu_{it})\phi, \tag{3.6}$$

onde g é uma função de variância conhecida e ϕ é um parâmetro de escala desconhecido. Liang & Zeger (1986) introduziram a idéia de "matriz de correlação de trabalho", $R_i(\alpha)$, uma matriz simétrica positiva definida, para cada $y_i = (y_{i1}, y_{i2}, \ldots, y_{it})'$, com objetivo de incorporar a correlação dentro do conglomerado. Assim a matriz variância-covariância de trabalho de y_i, V_i, é decomposta em termos de $R_i(\alpha)$.

Portanto, o terceiro passo é escolher a forma da matriz de correlação $n_i x\, n_i$, $R_i(\alpha)$, levando em consideração cada $y_i = (y_{i1}, y_{i2}, \ldots, y_{it})'$. O elemento (j, j') de $R_i(\alpha)$ é a correlação entre y_{it} e $y_{it'}$ estimada, hipotetizada ou conhecida. $R_i(\alpha)$ é chamada matriz de correlação de trabalho porque quando as respostas são não-normais, a real correlação entre indivíduos pode depender dos valores médios e daí de $x_{it}'\beta$ (Davis, 1991). Assume-se que esta matriz de correlação é conhecida exceto para um vetor sx1, fixo, de parâmetros desconhecidos α, que está associado com o modelo especificado para corr$(y_{it}, y_{it'})$. Embora esta matriz

---

[1] Assume-se que a densidade marginal de y_{it} é:

$$\pi(y, \theta_i, \phi) = exp\left\{\frac{[y\theta_i - a(\theta_i)]}{\phi} + C(y, \phi)\right\}$$

possa diferir de indivíduo para indivíduo, usa-se comumente uma matriz de correlação $R(\alpha)$, que se aproxima da dependência média entre observações repetidas dos indivíduos. Portanto a matriz de covariância para y_i é expressa da seguinte forma:

$$V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2} \phi, \tag{3.7}$$

onde A_i é uma matriz diagonal $n_i x n_i$, com $g(\mu_{ij})$ como j-ésimo elemento da diagonal.

O quarto passo do procedimento das EEG é estimar os parâmetros do vetor β e sua matriz de covariância pela solução da equação de estimação dada abaixo

$$U(\beta) = \sum_{i=1}^{k} D_i' \left[V_i(\alpha)\right]^{-1} S_i = 0 \tag{3.8}$$

onde,

$D_i = \dfrac{\partial \mu_i}{\partial \beta}$ é uma matriz $n_i xp$ de primeira derivada;

$S_i = (y_i - \mu_i)$ é um vetor $n_i x1$ de desvios com $\mu_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{in_i})'$ e

$V_i(\alpha)$ é definida em (3.7).

A equação (3.8) reduz-se à equação de quase-verossimilhança (3.4) quando $n_i = 1$ para todo i. Para cada i, $U_i(\beta,\alpha) = D_i' V_i^{-1} S_i$ é equivalente à função de estimação sugerida por Wedderburn(1974), exceto quando os V_i's são funções de β e também de α. A equação de estimação generalizada pode ser re-expressa como uma função somente de β se α é substituído em (3.7) e (3.8) por um estimador $\sqrt{k}$ -consistente $\hat{\alpha}(y,\beta,\phi)$, quando β e ϕ são conhecidos. Exceto em escolhas particulares de R e $\hat{\alpha}$, o parâmetro de escala permanece na equação. Logo para completar o processo deve-se substituir ϕ por um estimador $\sqrt{k}$ -consistente $\hat{\phi}(y,\beta)$ quando β é conhecido. Assim a forma da equação de estimação é

$$\sum_{i=1}^{K} U_i\{\beta, \hat{\alpha}[\beta, \hat{\phi}(\beta)]\} = 0. \tag{3.9}$$

A obtenção das estimativas de β é feita de modo iterativo. À cada interação, define-se

a variável dependente z_i como :

$z_i = D_i \beta^1 + S_i$, e com uma resolução de regressão linear ponderada com a variável dependente D_i, obtém-se a nova estimativa. A solução (3.8) é dada por $\hat{\beta} = \left(\sum_{i=1}^{K} D_i' V_i^{-1} D_i \right)^{-1} \left(\sum_{i=1}^{K} D_i' V_i^{-1} z_i \right)$. Os pesos dessas regressões são V_i^{-1}e o processo é repetido até chegar à convergência.

O procedimento das EEG produz estimadores consistentes dos parâmetros de regressão quando a função de ligação é corretamente especificada independentemente da escolha correta de $R_i(\alpha)$. Sob condições de regularidade moderada, Liang & Zeger(1986, Teorema 2) mostraram que, quando $k \to \infty$, $\hat{\beta}$ é um estimador consistente de β e $\sqrt{k}(\hat{\beta} - \beta)$ é assintoticamente normal multivariada com valor esperado 0 e matriz de covariância dada por

$$V_{\hat{\beta}} = \lim_{k \to \infty} K(M_0^{-1} M_1 M_0^{-1}) , \qquad (3.10)$$

em que , $M_0 = \sum_{i=1}^{K} D_i' V_i^{-1} D_i$ e $M_1 = \sum_{i=1}^{K} D_i' V_i^{-1} Cov(Y_i) V_i^{-1} D_i$.

O estimador consistente da covariância de $\hat{\beta}$, $V_{\hat{\beta}}$, pode ser obtido substituindo $Cov(Y_i)$ por $S_i S_i'$ [2], e β, φ, α por suas estimativas na expressão (3.10). Se $Cov(Y_i)$ é corretamente especificada, tal que $V_i = Cov(Y_i)$, então a matriz de covariância de $\hat{\beta}$ reduz-se a

$$V_{\hat{\beta}} = \lim_{k \to \infty} K(M_0^{-1}) \qquad (3.11)$$

produzindo assim um estimador mais eficiente (Carr & Chi, 1992).

Importante ressaltar que, embora este método produza estimativas consistentes mesmo quando se utilize uma estrutura incorreta para a matriz de correlação, podem ocorrer estimativas ineficientes ao se utilizar essa estrutura incorreta (Liang *et al.*, 1992 e Fitzmaurice *et al.*, 1993). Logo, esses autores sugerem o uso da estimativa "robusta" da matriz de covariância de $\hat{\beta}$ dada em (3.10). A propriedade de robustez só é garantida quando existe uma

[2] É o que se denomina como estimador "sanduiche".

pequena fração de dados faltantes ou quando estes são completamente aleatórios.

Em resumo, o procedimento das equações de estimação generalizadas tem muitas características atrativas que podem ser *sumarizadas* da seguinte forma: o método produz estimativas consistentes de $\hat{\beta}$, requerendo somente que a estrutura da média tenha sido corretamente especificada. Independentemente da estrutura de correlação de trabalho ser corretamente especificada, estimativas consistentes dos parâmetros de regressão são obtidas. Além disso, são obtidas estimativas de variância robusta de $\hat{\beta}$ mesmo que esta estrutura de correlação de trabalho tenha sido mal especificada.

Os parâmetros α e de escala ϕ podem ser estimados através dos resíduos de Pearson definidos por :

$$\hat{r}_{it} = \frac{(y_{it} - \hat{\mu}_{it})}{\sqrt{[\hat{V}_i^{-1}]_{it}}} \tag{3.12}$$

especificando ϕ como :

$$\hat{\phi}^{-1} = \sum_{i=1}^{K} \sum_{t=1}^{n_i} \hat{r}_{it}^{\,2} / (N - p), \text{ onde } N = \sum n_i . \tag{3.13}$$

O estimador específico de α depende da escolha de $R_i(\alpha)$. De forma geral α pode ser estimado por uma função de

$$\hat{R}_{uv} = \sum_{i=1}^{K} \hat{r}_{iu}\, \hat{r}_{iv} / (N - p) . \tag{3.14}$$

O procedimento das EEG permite que a dependência do tempo possa ser especificada de diferentes formas. Algumas especificações usadas para a $R_i(\alpha)$ são dadas abaixo:

1. $R_{st} = \begin{cases} 1, & se\ s = t \\ 0, & c.c. \end{cases}$ Esta é a estrutura de trabalho de independência na qual considera-se que as observações por indivíduos são independentes.

2. $R_{st} = \begin{cases} 1, & se\ s = t \\ \alpha, & c.c. \end{cases}$ Esta estrutura de correlação é chamada permutável (ou simétrica). Este modelo de trabalho assume que a correlação é constante entre duas observações no tempo. Esta estrutura é usada em modelos de efeitos aleatórios e é

razoável em situações nas quais as medidas repetidas não são obtidas sobre o tempo.

3. $R_{st} = \begin{cases} 1, & se\ s = t \\ \alpha^{|s-t|}, & c.c. \end{cases}$ Esta estrutura é usual em modelos autoregressivos, algumas vezes chamada de correlação multiplicativa. Este modelo de trabalho é equivalente ao modelo 1-dependente.

4. $R_{st} = \begin{cases} 1, & se\ s = t \\ \alpha, & se\ |s-t| = 1 \\ 0, & c.\ c. \end{cases}$ Esta estrutura corresponde ao modelo estacionário de ordem 1, cuja matriz é tridiagonal, ou seja, as observações são correlacionadas apenas com a observação imediatamente anterior e posterior.

5. $R_{st} = \begin{cases} 1, & se\ s = t \\ \alpha_{st}, & se\ 0 < |s-t| \leq g,\ \alpha_{st} = \alpha_{ts} \\ 0, & c.\ c. \end{cases}$ Esta estrutura corresponde ao modelo estacionário de ordem g.

6. $R_{st} = \begin{cases} 1, & se\ s = t \\ \alpha_{st}, & c.c.,\ \alpha_{st} = \alpha_{ts} \end{cases}$ Esta estrutura corresponde a uma matriz não-estruturada ou não especificada. É usada quando a quantidade de observações no tempo é a mesma para todos indivíduos ($n_i = n$)

Como a matriz de correlação de trabalho R é uma função de α, $R(\alpha)$, a estrutura de correlação pode ser:

Estrutura de correlação	Parâmetros
Independente	Nenhum parâmetro a estimar
Permutável	α é um escalar
Autoregressiva	α é um vetor
Estacionária	α é um vetor
Não-estacionária	α é uma matriz
Não-estruturada ou não-especificada	α é uma matriz

Desse modo, os respectivos estimadores específicos para α e, consequentemente, para $R(\alpha)$, segundo Stata (1996), podem ser expressos da seguinte forma de acordo com a estrutura:

Permutável:

$$\alpha = \sum_{i=1}^{m} \left[\frac{\sum_{j=1}^{n_i} \sum_{k=1}^{n_i} e_{i,j} e_{i,k} \; - \; \sum_{j=1}^{n_i} e_{i,j}^2}{n_i (n_i - 1)} \right] \Bigg/ \sum_{i=1}^{m} \frac{\sum_{j=1}^{n_i} e_{i,j}^2}{n_i}$$

Autoregressiva e estacionária:

$$\alpha = \sum_{i=1}^{m} \left[\frac{\sum_{j=1}^{n_i} e_{i,j}^2}{n_i} \; , \; \frac{\sum_{j=1}^{n_i - 1} e_{i,j} e_{i,j+1}}{n_i} \; , \; \ldots\ldots \; , \; \frac{\sum_{j=1}^{n_i - g} e_{i,j} e_{i,j+k}}{n_i} \right] \Bigg/ \sum_{i=1}^{m} \frac{\sum_{j=1}^{n_i} e_{i,j}^2}{n_i}$$

Não-estacionária e não-estruturada

$$\alpha = \sum_{i=1}^{m} m \begin{bmatrix} N_{1,1}^{-1} e_{i,1}^2 & N_{1,2}^{-1} e_{i,1} e_{i,2} & \cdots & N_{1,n}^{-1} e_{i,1} e_{i,n} \\ N_{2,1}^{-1} e_{i,2} e_{i,1} & N_{2,2}^{-1} e_{i,2}^2 & \cdots & N_{2,n}^{-1} e_{i,2} e_{i,n} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ N_{n,1}^{-1} e_{i,n_i} e_{i,1} & N_{n,2}^{-1} e_{i,n_i} e_{i,2} & \cdots & N_{n,n}^{-1} e_{i,n}^2 \end{bmatrix} \Bigg/ \left[\sum_{i=1}^{m} \frac{\sum_{j=1}^{n_i} e_{i,j}^2}{n_i} \right]$$

Essas estruturas de correlação produzem estimadores mais eficientes, porém são úteis somente quando existem poucas observações no tempo. Além disso, essas estruturas podem resultar numa matriz não-positiva definida quando existem dados faltantes e/ou uma variação no número de observações por indivíduo.

Outras estruturas de correlação podem ser consideradas desde que $\hat{\beta}$ e $V_{\hat{\beta}}$ sejam estimativas consistentes e assintoticamente normais para a escolha da mesma. Portanto, a estrutura de $R_i(\alpha)$ pode ser expressa mais genericamente como $g(R_i) = z_i \alpha$, onde z_i é um

conjunto de covariáveis específicas do indivíduo, e $g(R_i)$ é alguma função de ligação adequada. Alternativamente, z_i pode representar uma matriz de planejamento comum da dependência do tempo (Fitzmaurice *et al.*, 1993). A utilização desta forma traz vantagens e desvantagens no sentido de aumentar as alternativas e opções nas escolhas dessas matrizes. Porém, pode trazer outras complicações na resolução dos modelos face às restrições que provavelmente surgirão.

A solução das EEG para estimar β é resolvida pela interação entre os métodos de quase-verossimilhança para a estimação de β e um método robusto para estimação de α, usando diferentes estimativas consistentes de ϕ e R a cada interação, como segue abaixo:

1. Dada uma estimativa de $R_i(\alpha)$ e ϕ, calcula-se uma estimativa atualizada de β usando quadrados mínimos re-ponderados iterativamente.

2. Dada a estimativa $\hat{\beta}$ de β, calcula-se os resíduos padronizados $r_{it} = \frac{y_{it} - \hat{\mu}_{it}}{\sqrt{\left[V_i(\hat{\alpha})\right]^{-1}_{tt}}}$.

3. Utiliza-se os resíduos r_{it} para estimar consistentemente α e ϕ.

4. Repete-se os passos 1 a 3 até a convergência.

Segundo Zeger & Liang (1986), o procedimento EEG é aplicável não somente em análise de dados longitudinais com respostas binárias/discretas, como também pode ser estendido para incluir dados ordinais e multinomiais que possuem natureza multivariada. Este método também permite covariáveis dependentes do tempo, e pode ser adaptado para permitir observações faltantes (Ware *et al.*, 1988). Como exemplo, Miller *et al.* (1993) estenderam as equações de estimação para respostas politômicas e compararam esta metodologia com a de mínimos quadrados ponderados. Ainda no mesmo contexto, Lipstz *et al.* (1994) desenvolveram

modelos mais complicados para a estrutura de correlação de respostas multinomiais repetidas no lugar da estrutura de independência, e modelos menos complicados do que o usado por Miller, que são os modelos de correlação não estruturada.

Desde que surgiu a primeira publicação, em 1986, da metodologia EEG, diversas extensões e comparações com outras metodologias têm surgido nesses doze anos muito embora as aplicações sejam recentes. A exemplo de Park (1993) compara as EEG com o procedimento de máxima verossimilhança para um vetor de resposta normal multivariado. O autor mostra que as EEG reduzem-se às equações escore de máxima verossimilhança somente quando os dados não possuem observações faltantes e a matriz de correlação é a não estruturada. Algumas dessas extensões serão abordadas na seção seguinte.

3.4 Extensões da metodologia EEG

3.4.1 EEG2

No esforço de obter estimativas mais eficientes para ambos os parâmetros média e covariância, extensões da metodologia EEG para permitir estimação conjunta dos parâmetros de regressão e correlação têm sido propostas. Prentice (1988) propôs uma extensão do procedimento EEG para respostas binárias, permitindo a estimação conjunta dos parâmetros de ambas probabilidades da resposta marginal e correlações duas a duas através da introdução de uma segunda equação de estimação para o parâmetro de correlação α, da seguinte forma:

$$U(\alpha) = \sum_{i=1}^{k} E_i' \, W_i^{-1} \, [z_i - \rho_i(\alpha)], \qquad (3.15)$$

onde $E_i = \dfrac{\partial(\rho_i(\alpha))}{\partial\alpha}$;

W_i é a estrutura de covariância de trabalho de z_i e $z_i' = (z_{i12}, z_{i13}, \ldots, z_{i23}, \ldots)$ representa o vetor que contém as correlações duas a duas no mesmo conglomerado cujo valor esperado, $E(z_i)$, é $\rho_i(\alpha)$.

O problema é que a correlação entre variáveis binárias, como o próprio autor mostra, é restrita a pertencer ao seguinte intervalo:

$$max\left\{-\left(\frac{p_1p_2}{q_1q_2}\right)^{1/2}, -\left(\frac{q_1q_2}{p_1p_2}\right)^{1/2}\right\} \leq \rho \leq min\left\{\left(\frac{p_1q_2}{p_2q_1}\right)^{1/2}, \left(\frac{p_2q_1}{p_1q_2}\right)^{1/2}\right\}$$

Lipsitz, Laird & Harrington (1991) modificaram as equações de estimação de Prentice (1988) para permitir a modelagem da associação entre medidas repetidas via o uso da razão de chances.

Zhao & Prentice (1990) e Prentice & Zhao (1991) descreveram uma extensão da metodologia EEG para permitir a estimação conjunta dos parâmetros da média e covariância baseada no modelo exponencial quadrático de um vetor de valores da resposta $y_i' = (y_{i1}, y_{i2}, \ldots, y_{in_i})$, com média $\mu_i(\beta) = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{in_i})$ e covariância $\sigma_i(\beta,\alpha) = (\sigma_{i11}, \sigma_{i12}, \ldots, \sigma_{i22}, \ldots)$ com distribuição conjunta dos y_i da seguinte forma :

$$P(y_i, \mu_i, s_i) = \Delta_i^{-1} exp\left\{y_i' \theta_i + W_i' \lambda_i + C_i(y_i)\right\}, \qquad (3.16)$$

onde

$W_i = (y_{i1}^2, y_{i1}y_{i2}, \ldots, y_{i2}^2, y_{i2}y_{i3}, \ldots)$,

Δ_i é uma constante de normalização,

e os parâmetros canônicos $\theta_i' = \theta_i'(\mu_i, \sigma_i) = (\theta_{i1}, \ldots, \theta_{in_i})$ e

$\lambda_i' = \lambda_i'(\mu_i, \sigma_i) = (\lambda_{i11}, \lambda_{i12}, \ldots, \lambda_{i22}, \lambda_{i23} \ldots)$ são expressos como funções (μ_i, σ_i).

Esses autores propuseram modelar a média e a covariância da resposta como uma função das covariáveis por alguma função de ligação especificada. Na verdade como os parâmetros θ_i e λ_i têm interpretação em termos da distribuição condicional, uma transformação um a um de (θ_i, λ_i) em parâmetros marginais (μ_i, σ_i) é realizada de modo a obter equações análogas às de verossimilhança (Fitzmaurice et al,, 1993), deduzidas da seguinte forma :

$$K^{-1/2} \sum_{i=1}^{K} D_i' [V_i]^{-1} f_i = 0, \qquad (3.17)$$

onde, $f_i = \begin{pmatrix} y_i - \mu_i \\ s_i - \sigma_i \end{pmatrix}$ é o vetor de desvios, com $s_i' = (s_{i11}, s_{i12}, \ldots, s_{i_{n_i-1}n_i})$ sendo o vetor de covariâncias empíricas com cada $s_{ist} = (y_{ist} - \mu_{is})(y_{it} - \mu_{it})$ e $\sigma_{ist} = E(s_{ist})$;

$D_i = \begin{pmatrix} \frac{\partial \mu_i}{\partial \beta'} & 0 \\ \frac{\partial \sigma_i}{\partial \beta'} & \frac{\partial \sigma_i}{\partial \alpha'} \end{pmatrix}$ indica a matriz de derivadas com respeito aos parâmetros α e β,

e

$V_i = \begin{pmatrix} V_{i11} & V_{i12} \\ V_{i21} & V_{i22} \end{pmatrix} = \begin{pmatrix} cov(y_i) & cov(y_i, s_i) \\ cov(s_i, y_i) & cov(s_i) \end{pmatrix}$ indica a matriz de covariância particionada.

A expressão (3.17) produz uma classe de equações de estimação escore para os parâmetros da média e covariância, que depende da função forma de (3.16). Em particular, pode-se esperar diferentes combinações lineares de três ou mais elementos de y_i em $C_i(.)$, gerando assim um amplo intervalo de possíveis expressões para o terceiro e quarto momentos em V_i. Assim, Zhao & Prentice (1991) estabeleceram um procedimento de estimação mais conveniente especificando matrizes de covariância de "trabalho" em (3.17), no qual os momentos de 3ª e 4ª ordem em V_i são definidos como funções dos dois primeiros momentos, obtendo, portanto, estimativas de trabalho. Neste mesmo artigo, os autores sugerem diferentes especificações de trabalho para V_i.

O parâmetro de dependência no modelo original das EEG é parametrizado pela correlação dois a dois. Infelizmente o intervalo de variação do parâmetro da correlação é restrito pelas probabilidades marginais se os efeitos dos conglomerados são incluídos no modelo. Por esta razão, muitos autores, incluindo Lipsitz *et al.* (1991) e Liang *et al.* (1992), têm modelado a dependência pela associação entre dados binários usando a razão de chances, isto é,

$$OR(Y_1, Y_2) = \frac{P(Y_1 = 1, Y_2 = 1)\, P(Y_1 = 0, Y_2 = 0)}{P(Y_1 = 1, Y_2 = 0)\, P(Y_1 = 0, Y_2 = 1)}$$

que é ainda de fácil interpretabilidade.

Liang, Zeger & Qaqish (1992) especificaram a dependência do tempo em termos da razão de chances marginal, descrevendo um conjunto equivalente de equações para estimar conjuntamente os parâmetros da média e associação marginal. Essas equações foram denominadas pelos autores de Equação de Estimação Generalizada de Segunda Ordem (EEG2), que foram deduzidas primeiramente por Zhao e Prentice (1990). A principal vantagem desse procedimento é que produz estimativas mais eficientes para os parâmetros α e β desde que o modelo para ambas, média e associação marginal, seja corretamente especificado. Uma desvantagem desse método é que $\hat{\beta}$ pode deixar de ser consistente quando o modelo para associação marginal é mal-especificado, mesmo que o modelo para média tenha sido corretamente especificado.

Definindo $w_i = (y_{i1}y_{i2}, y_{i1}y_{i3}, y_{i1}y_{i4}, \ldots)'$ como um vetor de valores com comprimento $\frac{n_i(n_i-1)}{2}$ de produtos cruzados no qual $\eta_i = E(w_i)$, a equação de estimação generalizada de segunda ordem para α e β (EEG2) pode ser expressa da seguinte forma:

$$U(\delta) = \sum_{i=1}^{K} \begin{pmatrix} \partial\mu_i/\partial\beta & 0 \\ \partial\eta_i/\partial\beta & \partial\eta_i/\partial\alpha \end{pmatrix}' \begin{pmatrix} Cov(y_i) & Cov(y_i, w_i) \\ Cov(w_i, y_i) & Cov(w_i) \end{pmatrix}^{-1} \begin{pmatrix} y_i - \mu_i \\ w_i - \eta_i \end{pmatrix}. \qquad (3.18)$$

Quando o parâmetro de associação (α) for considerado como um parâmetro de perturbação e o número de conglomerados K for grande em relação ao tamanho de cada conglomerado (n_i), Liang *et al.* (1992) recomendam usar EEG1. No entanto, caso existam poucos conglomerados e/ou o parâmetro de associação α for o foco principal, é preferível o uso das EEG2 desde que a estrutura da média e associação sejam corretamente especificadas.

Apesar das EEG2 possuírem as propriedades de eficiência e otimalidade, o esforço computacional requerido no cálculo da matriz de variância-covariância torna o seu uso proibitivo à medida que o tamanho do conglomerado é superior a 15 (Carey, *et al.*, 1993).

3.4.2 Regressão Logística Alternada (RLA)

As EEG1 e EEG2 permitem a estimação dos parâmetros de 1° e 2° momento no modelo de regressão para dados binários multivariados. Quando a associação entre as observações é de interesse científico e é medida em termos da razão de chances marginal, as EEG2 são mais eficientes, porém, os cálculos tornam-se complicados à medida que o tamanho do conglomerado aumenta devido a inversão da matriz de covariância de dimensão dependente deste tamanho.

Isto é, para um conglomerado de tamanho n, a matriz de covariância dada em (3.18), $cov(Y,W)$, tem dimensão $(n + {}_nC_2)$ x $(n + {}_nC_2)$, tornando-se muito pesado, do ponto de vista da carga computacional a inversão dessa matriz à medida que n aumenta.

Seguindo uma sugestão de Firth (1992) e Diggle (1992), Carey *et al.* (1993) apresentam a metodologia de Regressão Logística Alternada (RLA), um tipo de extensão das equações de estimação generalizadas, como opção aos problemas existentes nas EEG2. O procedimento de RLA combina as EEG1 para β, pois as mesmas produzem uma estimativa robusta e razoavelmente eficiente, com uma nova equação de regressão logística para a estimação do parâmetro de associação (α) usando ${}_{n_i}C_2$ eventos condicionais: Y_{it} dado $Y_{ik} = y_{ik}$, para $1 \leq t \leq k \leq n_i$.

Segundo Ziegler *et al.* (1996) a regressão logística alternada corresponde ao procedimento de Prentice (1988), exceto que o log da razão de chances é usado ao invés da correlação. Sendo assim, para o caso mais simples da regressão do log da razão de chances, isto é, Log OR = α, a estimação de α é feita pela regressão de Y_{it} sobre Y_{ik}.

A idéia central é modelar separadamente a esperança marginal da cada variável binária como também a associação entre os pares de respostas em termos das covariáveis. Seja $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})'$ um vetor n_ix1 de variáveis binárias com média $E(Y_i) = \mu_i$ para o conglomerado $i = 1,2,\cdots,K$ e seja ψ_{its} a razão de chances entre as respostas Y_{it} e Y_{is} ($1 \leq t \leq s \leq n_i$). O modelo marginal é especificado da seguinte forma :

1) $h(\mu_{it}) = x_{it}' \beta_t$ onde $h(.)$ é uma função de ligação conhecida, x_{it} é um vetor px1 de covariáveis associadas com Y_{it} e β_t são os coeficientes de regressão a serem estimados;

2) $log\ \psi_{its} = z_{its}' \alpha$, onde z_{its} é um vetor qx1 de parâmetros de associação a ser estimado.

O procedimento de estimação para β e α é razoavelmente eficiente para ambos e evita a carga computacional do método EEG2. No caso mais simples, com $log\ \psi_{its} = \alpha$, o processo é alternado entre dois passos :

- Para um dado α, estima-se β como um parâmetro na regressão logística marginal usando EEG1;
- Para um dado β, estima-se o parâmetro da razão de chances α usando uma regressão logística de Y_{it} sobre cada Y_{is}(s > t) com um offset[3] que envolve μ_{it} e $v_{its} = E(Y_{it} Y_{is})$.

A hipótese é de que usando eventos condicionais ao invés de produtos dois a dois, equações de estimação mais simples podem ser usadas e ainda produzir estimativas altamente eficientes para α e β.

A título de ilustração considere uma tabela 2x2 relacionando duas respostas de n_i elementos no conglomerado $Y_i, i = 1,2,\cdots,K$. Seja π_{irs} a probabilidade de uma casela na tabela 2x2 que relaciona as respostas Y_{it} e Y_{is}, ou seja, $\pi_{irs} = P(Y_{it} = r, Y_{is} = u)$, $r,u \in [0,1]$.

Assim numa tabela 2x2 temos

	Y_{is}	
Y_{it}	0	1
0	π_{i00}	π_{i01}
1	π_{i10}	π_{i11}

[3] Termo usado em modelos lineares generalizados para representar uma variável adicional no modelo cujo coeficiente não é preciso estimar.

Seja $\mu_{it} = E(Y_{it}) = P(Y_{it} = 1) = \pi_{i10} + \pi_{i11}$, $\mu_{is} = E(Y_{is}) = P(Y_{is} = 1) = \pi_{i01} + \pi_{i11}$ e $\upsilon_{its} = \mu_{its} = E(Y_{it}Y_{is}) = P(Y_{it} = 1, Y_{is} = 1) = \pi_{i11}$, as esperanças marginais de primeira e segunda ordem respectivamente. O passo na dedução da RLA é expressar os parâmetros do log da razão de chances duas a duas em termos das chances condicionais duas a duas, ou seja,

$$\frac{P(Y_{it} = 1 / Y_{is} = y_{is})}{P(Y_{it} = 0 / Y_{is} = y_{is})} = \left(\frac{\pi_{i11}}{\pi_{i10}}\right)^{y_{is}} \left(\frac{\pi_{i10}}{\pi_{i00}}\right)^{1-y_{is}}$$

Tomando log em ambos os lados, obtém-se o modelo de regressão logística condicional com offset, ou ainda o modelo geral de regressão logística alternada:

$$logito\, P(Y_{it} = 1 \mid Y_{is} = y_{is}) = \gamma_{its}\, y_{is} + log\left(\frac{\mu_{it} - \upsilon_{its}}{1 - \mu_{it} - \mu_{is} + \upsilon_{its}}\right) \qquad (3.19)$$

onde γ_{its} é o log da razão de chances e o segundo termo do lado direito de (3.19) é usado como offset em que $\mu_{it} = P(Y_{it} = 1)$ e $\upsilon_{its} = \mu_{its} = P(Y_{it} = 1, Y_{is} = 1)$.

Para uma melhor compreensão desta metodologia algumas definições são necessárias. Assim, define-se $w_i = (Y_{i1}Y_{i2}, Y_{i1}Y_{i3}, \ldots, Y_{i2}Y_{i3}, \ldots)$ como o vetor $\binom{n_i}{2}$x1 de produtos cruzados cujo valor esperado é $\upsilon_i = E(w_i) = [\, E(Y_{i1}Y_{i2}),\, E(Y_{i1}Y_{i3}), \ldots E(Y_{i2}Y_{i3}),\, \ldots]$.

Seja $\xi_i = (\xi_{i12}, \xi_{i13}, \ldots, \xi_{i23}, \ldots)$ um outro vetor $\binom{n_i}{2}$x 1, cujos elementos são $\xi_{its} = E(Y_{it} \mid Y_{is} = y_{is}) = \text{logit}^{-1}\left\{\gamma_{its}\, y_{is} + log\left(\frac{\mu_{it} - \upsilon_{its}}{1 - \mu_{it} - \mu_{is} + \upsilon_{its}}\right)\right\}$; e R_i um vetor de resíduos cujos elementos representam $R_{its} = Y_{it} - \xi_{its}$. Então, a estimativa da regressão logística alternada $\delta = (\beta, \alpha)$ é solução simultânea das seguintes equações de estimação :

$$U(\beta) = \sum_{i=1}^{K}\left(\frac{\partial \mu_i}{\partial \beta}\right)' V_i^{-1}\,(Y_i - \mu_i) = 0 \qquad (3.20)$$

$$U(\alpha) = \sum_{i=1}^{K} \left(\frac{\partial \xi_i}{\partial \alpha} \right)' S_i^{-1} (Y_i - \xi_i) = 0 \qquad (3.21)$$

A equação (3.20) corresponde às EEG1 e a esperança condicional ξ_{its} é a base de uma equação de estimação ponderada para estimar α, onde S_i na equação (3.21) é uma matriz diagonal $\binom{n_i}{2} \times \binom{n_i}{2}$, cujos elementos são $\xi_{its}(1-\xi_{its})$ (variância de Y_{it}). A solução $\delta = (\beta, \alpha)$ dessas equações é obtida usando o algoritmo Gauss-Siedel não linear e as estimativas $\hat{\delta}_{RLA}$ obtidas deste algoritmo são consistentes e assintoticamente normais (Carey *et al.*, 1993).

Em resumo, RLA um procedimento para regressão simultânea da resposta sobre as variáveis explanatórias, bem como da associação entre cada par de resposta em termos da razão de chances duas a duas, combinando as equações de estimação de primeira ordem para β com uma nova equação de regressão logística para estimar α.

No artigo citado acima, os autores fazem um estudo comparativo do tempo médio de convergência dos procedimentos de RLA e EEG2, concluindo que o primeiro é mais econômico do ponto de vista de tempo computacional. Eles também estudaram a eficiência de $\hat{\alpha}$ quando estimado pela RLA e pelas EEG2 e observaram uma eficiência 90% melhor na solução da Regressão Logística Alternada à medida que esta associação torna-se mais acentuada.

3.5 Diagnóstico nas EEG

O diagnóstico na regressão para o modelo linear normal é bem desenvolvido na literatura (por exemplo, Chatterjee & Hadi, 1986). Existem várias quantidades estatísticas para decidir se o modelo de regressão ajusta bem ou não os dados. Essas medidas têm sido

propostas, por exemplo, para identificar pontos aberrantes, para detectar efeito temporal nos dados, para medir a influência de uma simples observação no modelo e nos parâmetros estimados.

Por se tratar de um tema recente, somente nos últimos anos têm surgido na literatura trabalhos que enfocam as técnicas de diagnósticos nas EEG para identificar conglomerados ou observações que influenciam as estimativas dos parâmetros do modelo. Através de contatos com alguns desses autores, John Preisser e Andreas Ziegler, foi possível elaborar uma pequena exploração do assunto neste documento.

No desenvolvimento do tema são descritas resumidamente duas propostas, porém a aplicação se concentrará em apenas uma delas. A primeira baseia-se nos trabalhos desenvolvidos por Ziegler *et al.* (1996) e Ziegler & Arminger (1996). Esse último trabalho mostra que o modelo original de Liang & Zeger pode ser obtido pela teoria de estimação de pseudo máxima verossimilhança de Gourieroux & Monfort (1993). Assim, as técnicas de diagnóstico de regressão nas EEG1 apoiam-se na estimação de pseudo máxima verossimilhança. As medidas utilizadas no artigo baseiam-se nos resíduos e nas estatísticas de influência.

Segundo Ziegler & Arminger (1996), se a matriz de covariância apresentada em (3.7) for especificada corretamente e estimada consistentemente pelo método de máxima verossimilhança quase generalizado [$\hat{V}(Y_i) = \hat{\Omega}(X_i, \hat{\beta}, \hat{\alpha})$] , os resíduos padronizados e Studentizado podem ser usados para avaliar a variação sistemática que é causada por um ou mais regressores, ou seja, detectar pontos aberrantes.

Para detectar pontos de alavanca (ou de alto "leverage") a matriz (n_i x n_i) $H_i = \hat{\Omega}_i^{-1/2} \hat{D}_i \, (\hat{D}_i' \, \hat{\Omega}_i^{-1/2} \hat{D}_i)^{-1} \, \hat{D}_i' \, \hat{\Omega}_i^{'-1/2}$ pode ser usada. A matriz (kn_i x kn_i) bloco diagonal H = diag (H_i) é uma matriz de projeção ortogonal de posto p, daí $tr(H) = p$, e consequentemente $tr(H_i) = \frac{p}{k}$, no caso multivariado.

Com o objetivo de medir a influência de uma observação sobre as estimativas $\hat{\beta}$ das EEG; os autores propõem uma estatística de Cook modificada. A idéia é usar a estimativa

robusta da variância de $\hat{\beta}$ dada em (3 10), obtendo assim:

$$\hat{C}_i = n(\hat{\beta} - \hat{\beta}_{(i)})' [M_0^{-1} M_1 M_0^{-1}]^{-1} (\hat{\beta} - \hat{\beta}_{(i)}) \qquad (3.24)$$

onde $\hat{\beta}_{(i)}$ representa o estimador de pseudo máxima verossimilhança ao se omitir a observação do i– ésimo conglomerado.

Em um outro trabalho (aguardando publicação), Ziegler *et al.* utilizam as técnicas de diagnóstico de regressão para identificar conglomerados que influenciam a estrutura da média e da associação do modelo, isto é, dos parâmetros β e α, respectivamente. Essas técnicas são usadas para analisar dados de agregação familiar no sentido de detectar famílias ou grupos de famílias que têm uma predisposição genética para câncer de esôfago.

A segunda proposta refere-se ao trabalho desenvolvido por Preisser & Qaqish (1996). Esses autores fazem uma generalização dos diagnósticos para modelos lineares generalizados produzidos por Pregibon (1981), Williams (1987) e McCullagh & Nelder (1989). Eles propõem o diagnóstico de exclusão de uma observação ou um conjunto de observações (o conglomerado) para detectar pontos de alavanca (alto "leverage") e avaliar os resíduos com a finalidade de medir sua influência sobre as estimativas dos parâmetros de regressão e valores ajustados. Quando se trata de somente uma observação, eles denominam o diagnóstico como "exclusão da observação", e no outro caso de "exclusão do conglomerado" .

As medidas de influência baseadas na exclusão de casos apresentadas por esses autores utilizam uma aproximação a um passo, primeiramente proposta por Pregibon (1981) e posteriormente por Williams (1987), para modelos lineares generalizados. Desse modo, o efeito de excluir uma ou mais observações é medido pela mudança nos valores de $\hat{\beta}$ e sobre os valores estimados do preditor linear. Na verdade Preisser & Qaqish (1996) demonstram uma generalização da aproximação a um passo nas EEG.

Primeiramente os autores apresentam $DBETAC_i$ e $DBETAO_{ti}$, respectivamente, para medir o efeito da exclusão de um conglomerado e de uma observação sobre a estimativa dos parâmetros de regressão. Essas estatísticas podem ser padronizadas utilizando o vetor de erro-padrão de $\hat{\beta}$. Os autores utilizam o vetor de erro-padrão "naive" por achar que o estimador

"sanduiche" da variância robusta pode ser inflacionado por resíduos grandes.

Para medir a influência das observações sobre os valores estimados do preditor linear e, portanto, dos valores ajustados, Preisser & Qaqish (1996) demonstram em seguida, duas outras estatísticas. A primeira delas mede o efeito do i-ésimo conglomerado excluído sobre o ajuste global, $DCLS_i$, e a segunda mede o efeito da influência da t-ésima observação do conglomerado i sobre o ajuste global, $DOBS_{it}$. Os autores também apresentam um estatística que na verdade é uma generalização do DFITS de Belsley *et al.* (1980, p.15), que tem como objetivo medir a influência de um subconjunto de observações sobre $\hat{\beta}$ e sobre a variância de $\hat{\beta}$, simultaneamente $GCLS_i$ (ou $MCLS_i$).

Capítulo IV
Aplicações e Considerações Finais

4.1 Introdução

Este capítulo apresenta aplicações das metodologias tratadas nos capítulos II e III, utilizando os dois conjuntos de dados descritos no início do trabalho e considerações finais.

No que diz respeito a estimação da razão, devido ao interesse em utilizar uma medida epidemiológica como a incidência e também por ela representar uma medida tipo razão no seu sentido mais amplo, a metodologia da razão para dados discretos de uma amostra de conglomerado em um estágio, será aplicada ao estudo de Serrinha como passo inicial. Posteriormente um modelo utilizando a metodologia de mínimos quadrados ponderados será ajustado para verificar a influência simultânea das variáveis significativas de uma análise univariada preliminar. Na pesquisa AISAM, a metodologia da razão será empregada no contexto de uma amostra de conglomerado em dois estágios, e um modelo será ajustado para o vetor de razão estimada, utilizando a metodologia de mínimos quadrados ponderados.

A resposta de interesse para ambos conjuntos de dados será a densidade de incidência de diarréia, expressa pelo número de episódios severos de diarréia para o estudo de Serrinha e episódios moderados e/ou severos de diarréia para o estudo AISAM, por criança-dia.

Para a metodologia de equações de estimação generalizadas, apresentada no capítulo III, será explorada a questão da medida repetida em estudos longitudinais. Neste caso, para ambos os conjuntos de dados, as respostas de interesse utilizadas no modelo estatístico serão respectivamente o número de episódios severos de diarréia ocorridos num determinado período e a presença/ausência de episódio de diarréia para as pesquisas de Serrinha e AISAM.

Por outro lado, devido às estruturas dos dados, o procedimento de regressão logística alternada, um tipo de extensão das EEG, será abordado em apenas um dos conjuntos de dados-AISAM. Ainda com respeito às metodologias apresentadas no capítulo III, será feita uma breve ilustração de diagnóstico nas EEG.

Nas seções seguintes são descritos os programas utilizados na aplicação das metodologias aos dados e nas análises realizadas.

4.2 Programas computacionais

Com o objetivo de incorporar a variação no tamanho da amostra devido à amostragem de conglomerado, a medida ou a estimativa de interesse adotada neste trabalho é uma razão, mais especificamente, um vetor de razões de médias, juntamente com a estimativa das variâncias baseadas em Série de Taylor. Cabe comentar também sobre o uso do peso de amostragem no ajuste da amostras complexas.

Alguns programas já permitem o ajuste do delineamento amostral na estimativa de interesse, incluindo as questões referentes a medidas como razão, plano amostrais complexos e probabilidades de seleção desiguais. São exemplos, a rotina Pstable do software OSIRIS IV (University of Michigan, Survey Research Center Computer Support Group) ou PC CARP (University Iowa, Survey Section of the Statistical Laboratory) e mais recentemente o PROC RATIO do programa SUDAAN (Research Triangle Institute, NC).

Um programa no IML do SAS foi construído para obter a estimativa do vetor de razões de médias e sua correspondente matriz de variância-covariância, ambas ajustadas ao delineamento complexo da amostra. Na análise de regressão usando mínimos quadrados ponderados, o PROC CATMOD do SAS pode ser facilmente utilizado.

No caso da metodologia de equações de estimação generalizadas, por ser uma técnica bastante recente, alguns programas foram desenvolvidos particularmente por pesquisadores envolvidos no tema. Até a finalização deste trabalho sabe-se da existência de 7 programas para análise de dados usando as EEG.

O primeiro deles é um programa em SAS elaborado por Karim & Zeger (1ª versão, 1988), mais tarde atualizado por Dr.Ulrike Groemping em 1994, ambos disponíveis na internet[1]. Esta macro do SAS modela dados longitudinais através das EEG para uma classe geral de variáveis respostas Gaussiana, Poisson, Binomial e Gama. O usuário pode especificar a estrutura de correlação de trabalho, cuja matriz refere-se à correlação entre observações repetidas dentro do conglomerado (conjunto de observações repetidas numa mesma unidade amostral). As opções oferecidas são: a matriz de correlação identidade, estacionária m-dependente, não-estacionária m-dependente, permutável, auto-regressiva de ordem 1 (AR-1) e não especificada.

Posteriormente, um programa em linguagem Pascal foi elaborado pelo Prof. B. Qaqish. (1989, 1990, 1991) para as EEG estendidas (EEG2), com dados binários correlacionados. Este programa permite a escolha entre as EEG1 e EEG2, cuja única função de ligação disponível é o logito.

Outro programa elaborado por Davis (1993) em linguagem Fortran 77 para análise de dados usando EEG é o RMGEE. Ele pode ser utilizado para variáveis respostas com distribuição Normal, Binomial e Poisson e com as opções independente, permutável e não estruturada para a matriz de correlação de trabalho.

[1] htpp://www.statlab.uni-heidelberg.de/statlib/GEE/GEE1/

Um programa em Linguagem S é uma outra alternativa para as EEG, cuja rotina desenvolvida por Dr. Vicent Carey foi implementada no OSWALD (Smith, 1996). Esta macro pode ser instalada no diretório "library" do S-plus e encontra-se disponível também na internet[2]. As opções para a matriz de correlação de trabalho são as mesmas da macro do SAS. Além da rotina das EEG, o OSWALD também possui uma outra chamada Regressão Logística Alternada (RLA) para analisar dados binários longitudinais, que permite modelar a razão de chances de uma série de dados binários.

Em 1997, Kastner et al. desenvolveram o programa MAREG & WINMAREG para modelar simultaneamente a distribuição marginal e a associação. É possível utilizar o procedimento das EEG introduzidos por Liang & Zeger (1986), além da opção do método de máxima verossimilhança proposto por Fitzmaurice & Laird (1993), para o caso de dados binários como também a generalização a dados multicategorizados feita por Heumann (1996). Portanto pode-se modelar dados binários pela função de ligação logito além das funções de ligação logito cumulativo e logito multinomial para dados multicategorizados[3] .

Também já é possível utilizar o PROC GENMOD da versão 6.12 do SAS para obter as estimativas das EEG através do subcomando REPEATED. Outro software que implementou as EEG é o Stata, com a rotina xtgee.

4.3 Estudo de Serrinha

Trata-se de um ensaio clínico duplo-cego, placebo controlado com 1240 crianças entre 6 a 48 meses de idade acompanhadas por um período de um ano. As crianças foram designadas aleatoriamente aos grupos de vitamina e placebo. Uma cápsula de placebo ou vitamina A foi oferecida às crianças a cada quatro meses, durante um ano. Informações sobre a morbidade, especialmente diarréia e infecção respiratória, foram coletadas em dias alternados da semana, durante todo o período de seguimento do estudo.

[2] htpp://www.maths.lancs.ac.uk/Software/Oswald/
[3] http://www.stat.uni-muenchen.de/~andreas/winmareg.html

Antes da exposição dos resultados das análises, é feita uma breve apresentação das características das crianças, nos dois grupos, no início do estudo (Tabela 1), das comparações descritivas encontradas nos dois grupos, como a prevalência média diária de diarréia em função de diferentes definições (Tabela 2) e a incidência de episódios de diarréia de acordo a severidade (Tabela 3).

Esse estudo detectou um efeito protetor da suplementação de vitamina A na redução da incidência dos episódios severos de diarréia para as crianças do grupo suplementado quando comparado com as crianças do grupo placebo. As crianças suplementadas tiveram uma redução de 10%, 20% e 23%, respectivamente, nas prevalências diárias dos episódios leve, moderado e severo de diarréia quando comparadas com as crianças do grupo placebo (ver abaixo). No estudo, o episódio leve de diarréia foi definido pela duração de até dois dias, com 3 ou mais dejeções líquidas ou semilíquidas nas 24 horas; moderado quando durava 3 ou mais dias, com uma média de até 4 dejeções nas 24 horas e, severo quando durava 3 ou mais dias e tinha uma média de 5 ou mais dejeções líquidas ou semilíquidas nas 24 horas (Barreto et al., 1994).

Tabela 1: Características demográfica, nutricional, clínica e sócio-econômica das crianças no início do estudo

	Vitamina A (n = 620)	Placebo (n = 620)
Características demográficas		
Média de idade em dias (DP)	840 (379)	859 (370)
% Masc / Fem	53.3 / 46.7	51.2 / 48.8
Nascimento em hospital (%)	80.5	77.3
Nutrição		
% de aleitamento no peito	13.2	13.8
Mediana da duração da amamentação (meses)	4.0	4.0
Mediana da duração da exclusão da amamentação (meses)	1.0	1.0
Antropometria (%)		
Altura por idade < - 2 Z score	23.6	20.1
Peso por idade < - 2 Z score	12.1	13.1
Peso por altura < - Z score	1.4	1.4
História clínica (%)		
Sarampo	9.9	11.2
Pneumonia	4.6	4.1
Asma	14.6	13.6
Qualquer doença nas 2 últimas semanas	22.2	22.9
Diarréia nas 2 últimas semanas	14.6	13.8
Padrão Sócio-econômico (%)		
Água encanada em casa	96.1	96.4
Eletricidade	93.6	96.4
Refrigerador	39.5	37.7
Televisão	59.1	57.2
Educação primária completa do pai	72.3	71.0
Educação primária completa do mãe	76.8	76.1

Fonte : Barreto et al. (1994)

Tabela 2: Prevalência média diária de diarréia de acordo a diferentes definições

Núm. De dejeções líquidas por dia	Vitamina A	Placebo	Razão Vitamina A / Placebo	p-valor
≥ 3	0.0478	0.0517	0.92	0.074
≥ 4	0.0232	0.0259	0.90	0.049
≥ 5	0.0099	0.0123	0.80	0.005
≥ 6	0.0043	0.0056	0.77	0.006

Fonte : Barreto et al. (1994)

Tabela 3: Incidência de episódios de diarréia de acordo a severidade

Severidade do episódio	Vitamina A		Placebo		Razão de taxas (95% IC)
	Número	Incidência (x 10^{-3} criança-dia)	Número	Incidência (x 10^{-3} criança-dia)	
Leve	2131	10.48	2179	10.80	0.97 (0.91-1.03)
Moderado	1614	7.94	1770	8.77	0.91 (0.85-0.98)
Severo	165	0.81	204	1.01	0.80 (0.65-0.98)
Total	3745	18.42	3949	19.58	0.94 (0.90-0.98)

Fonte : Barreto et al. (1994)

4.3.1 Método da razão e mínimos quadrados ponderados

Segundo Assis, 1997, na atualidade tem-se delineado a suposição de que a redução da mortalidade sob o efeito do suplemento de vitamina A processa-se, principalmente, através da diminuição da severidade dos episódios infecciosos; suposição esta que tem sido amparada por alguns resultados de estudos clínicos como Bhandar et al. (1994), Ghana (1993) e Barreto et al. (1994). Assim, para analisar os dados de Serrinha com a metodologia da razão e mínimos quadrados ponderados, cada criança corresponderá a um conglomerado, e a resposta utilizada será a densidade de incidência de episódios severos de diarréia, representando uma medida do tipo razão, e as covariáveis estudadas são aquelas que além da importância epidemiológica também apresentaram um efeito estatisticamente significativo em análises preliminares.

Numa primeira etapa estima-se o vetor de razões, no caso o vetor de densidades de incidência, com a respectiva matriz de covariância, para os fatores de risco. Como foi mencionado no capítulo II, o vetor de razões pode ser gerado de subgrupos definidos simplesmente pelas variáveis explanatórias ou fatores de risco, como também pela classificação cruzada desses fatores de risco. No entanto, deve-se ter cautela, pois na presença de muitos fatores de risco, o método da razão para a classificação cruzada pode não ser apropriado devido ao tamanho pequeno das frequências das caselas resultantes desta classificação cruzada. Ou seja, com um número grande de variáveis explanatórias, a suposição de normalidade

assintótica para as estimativas específicas dos subgrupos exigida pelo método da razão pode não ser válida.

Como resultado dessa primeira etapa, na Tabela 4 abaixo, encontra-se as densidades de incidência de diarréia severa obtidas pelo número de episódios de severo de diarréia por criança-dia, observadas para os subgrupos definidos por alguns fatores de risco, e a estimativa da variância dessas incidências.

Na Tabela 5, é utilizado um teste para comparar essas densidades de incidência, isto é, um teste para as razões das densidades de incidência baseado na estatística escore e são apresentados intervalos de confiança (a 95%) para a razão dessas densidades de incidência.

Tabela 4. Densidades de incidência de diarréia severa para alguns fatores de risco

Fatores	Níveis	Número de crianças	Número de episódios severo de diarréia	Crianças-dia em risco	R=densidade de incidência (x 10^{-3})	Var(R)[a] (x 10^{-3})
Grupo	Vitamina A	451	157	163833	0.9583	0.000009
	Placebo	448	209	162579	1.2855	0.000015
Sanitário	Presença	653	217	237507	0.9136	0.000006
na casa	Ausência	149	149	88905	1.6759	0.000036
Sexo da	Masculino	466	183	168962	1.0831	0.000011
criança	Feminino	433	183	157450	1.1622	0.000014
Tratamento	Algum	631	227	229123	0.9907	0.000007
da água	Nenhum	268	139	97289	1.4287	0.000031
Tipo de	Outros	823	320	299092	1.0699	0.000006
Piso	Terra	76	46	27320	1.6884	0.000134
Habitantes	0 a 4	395	156	143766	1.0851	0.000013
na casa	5 a 10	407	171	147796	1.157	0.000014
	> 10	97	39	34850	1.1191	0.000053
Escolaridade	2°. grau	87	20	31668	0.6316	0.000033
da	1°. grau	136	69	49474	1.3947	0.000069
mãe	Prim/analf.	676	277	245270	1.1294	0.000007

Haz (altura/	Normal	665	240	241560	0.9935	0.000007
idade) final[b]	Desnutrido	231	133	83759	1.4685	0.000003
Destino	Rede esgot.	310	94	112765	0.8336	0.000011
dos dejetos	Outros	589	272	213647	1.2731	0.000011

[a] Estimativa da variância por série e Taylor
[b] Haz: Indicador antropométrico em z-score[4]

Tabela 5. Estimativa e intervalo de confiança da razão das densidades de incidência, valor da estatística e o nível descritivo do teste

Fatores	Razão das densidades de incidência [IC a 95%]	Estatística escore	valor p
Grupo	0.7454 [0.5694 , 0.9758]	4.57	0.0325
Sanitário	1.8344 [1.3902 , 2.4203]	18.4	0.0000
Sexo da criança	0.9319 [0.7075 , 1.2273]	0.25	0.6171
Tratamento da água	1.4421 [1.0760 , 1.9327]	6.01	0.0142
Tipo de piso	1.5737 [1.0043 , 2.4659]	8.64	0.0033
Habitantes na casa	1.0663 [0.7999 , 1.4213]	0.19	0.6629
	1.0313 [0.6558 , 1.6219]	0.01	0.9203
Escolaridade da mãe	2.2083 [1.1256 , 4.3324]	5.31	0.0212
	1.7882 [0.9991 , 1.8970]	0.09	0.7641
Haz (altura por idade)	1.4780 [1.2339 , 1.7704]	17.99	0.0000
Dejetos	1.5273 [1.1374 , 2.0508]	7.93	0.0049

[4] Esse indicador antropométrico é calculado com base nos parâmetros de uma população de referência do National Center for Health Statistics (NCHS).

Com exceção do fator grupo, todas as razões de densidades de incidência foram calculadas utilizando as densidades de incidência do segundo nível do fator (Tabela 4) pelos demais níveis. Os fatores grupo, sanitário, tratamento da água, escolaridade da mãe, tipo de piso, indicador antropométrico haz e destino dos dejetos são significativos nesta análise univariada. Pelos resultados observa-se que a suplementação com vitamina A exerce um fator protetor contra a diarréia severa demostrado pelo fator grupo (redução de 26%). O fator sanitário revela um risco de 1.8 vezes maior de apresentar episódios severos de diarréia, por criança-dia, na ausência de sanitário na casa. Já o fator destino dos dejetos, por exemplo, apresenta um risco de 1.5 vezes de apresentar episódios severos de diarréia, por criança-dia, na ausência de rede de esgoto.

A segunda etapa envolve a estimativa das razões e sua matriz de covariância para a classificação cruzada dos fatores de risco significativos na análise univariada (Tabela 5) e posteriormente o ajuste de modelo estatístico através da metodologia de mínimos quadrados ponderados dessas densidades de incidência. É importante lembrar que não foi possível utilizar todos os fatores significativos na classificação cruzada devido a ocorrência de frequências pequenas em algumas caselas.

Tendo em vista uma melhor interpretabilidade para os coeficientes estimados, optou-se pelo uso de um modelo linear para o log das razões. Assim, o modelo abaixo será ajustado para o vetor de log das densidades de incidência utilizando a metodologia de mínimos quadrados ponderados.

$$E(Log\,R) = \beta_0 + \beta_1\, GRUPO1 + \beta_2\, TRATAGUA + \beta_3\, HAZ$$

onde GRUPO1 é variável indicadora assumindo o valor 0 para placebo e 1 para vitamina A ; TRATAGUA também é uma variável indicadora assumindo o valor 0 para algum tipo de tratamento da água e o valor 1 para nenhum tratamento da água e a variável indicadora HAZ representa um indicador antropométrico altura/idade em z-escore, assumindo 0 para as crianças classificadas como normais e 1 para desnutridas. Ajustes preliminares mostraram que nenhuma interação foi significativa, resultando assim no modelo aditivo descrito acima.

A Tabela 6 apresenta as estimativas e erros padrões dos parâmetros do modelo adotado para a regressão da razão de médias.

Tabela 6. Estimativa e erro-padrão dos efeitos dos fatores pelo método de regressão da razão

Efeito	Estimativas ($\hat{\beta}$)	Erro-padrão	valor-p	RDI [a]	IC a 95% [b]
Intercepto	-6.8879	0.1139	0.0000		
Grupo	-0.2823	0.1383	0.0413	0.754	0.575-0.989
Tratagua	0.2964	0.1474	0.0444	1.345	1.008-1.796
Haz	0.3803	0.1466	0.0095	1.463	1.097-1.949
χ^2 de bondade de ajuste (4 gl)	1.07		p=0.8992		

[a] Razão das densidades de incidência = risco relativo
[b] Intervalo de confiança para o risco relativo

Observa-se nos resultados da Tabela 6 que todos os fatores são significativos no modelo. A estatística χ^2 indica que o modelo representa adequadamente os dados. Os resultados da última coluna da tabela são as estimativas do risco relativo, oriundas da exponenciação dos coeficientes, juntamente com os intervalos de confiança. Assim, há uma redução de 25% na incidência de diarréia severa para crianças suplementadas; um risco de 1.3 vezes maior das crianças apresentarem episódios severos de diarréia, por crianças-dia, na ausência de algum tratamento na água de beber e um risco de 1.5 vezes maior das crianças apresentarem episódios severos de diarréia, por crianças-dia, quando são desnutridas. Do ponto de vista epidemiológico, é importante comentar que após o ajuste do modelo simultâneo desses fatores, seus respectivos efeitos não se modificam em relação às análises univariadas, indicando assim o que os epidemiologistas definem como fatores não confundidores de efeito.

A tabela com as razões observadas e preditas pelo modelo é apresentada na próxima página.

Tabela 7. Razões (Densidades de Incidência) e erros-padrões observados e preditos pelo modelo definidos pela classificação cruzada simultânea de grupo, tratagua e haz

Grupo	Tratagua	Haz	R=DI observadas ($x\ 10^{-3}$) *	Desvio Padrão(R) ($x\ 10^{-3}$)	$\hat{R}$ preditas pelo modelo	DP($\hat{R}$)
Vitamina A	Algum	Normal	0.7935	0.1210	0.7692	0.0936
		Desnutrido	1.1517	0.2356	1.1251	0.1657
	Nenhum	Normal	0.9607	0.2231	1.0345	0.1616
		Desnutrido	1.4633	0.3696	1.5132	0.2379
Palcebo	Algum	Normal	0.9626	0.1298	1.0201	0.1162
		Desnutrido	1.6042	0.3799	1.4921	0.2275
	Nenhum	Normal	1.6067	0.3556	1.3720	0.2042
		Desnutrido	1.8515	0.4566	2.0069	0.3227

* Incidência de episódios severo de diarréia por criança-dia

Observa-se na Tabela 7 que os erros-padrões preditos pelo modelo são menores que os observados.

4.3.2 As equações de estimação generalizadas e extensões

Dos programas mencionados, dois deles foram utilizados para análise dos dados: a segunda versão da macro do SAS para análise com as EEG e a macro em linguagem S implementada no OSWALD para aplicar a metodologia de Regressão Logística Alternada .

Para a metodologia das EEG, os dados de entrada devem estar num formato SAS, contendo a variável resposta e as covariáveis, sendo que este arquivo deve conter um registro (ou linha) para cada medida repetida por indivíduo avaliado. Apesar desta macro processar conglomerados de tamanhos diferentes, optou-se pelo uso do mesmo número de observações repetidas, para a resposta de interesse, das crianças no estudo. Ou seja, somente crianças com um ano completo de acompanhamento contribuíram com informação para a aplicação da metodologia, resultando assim num total de 899 crianças. Ao trabalhar com esta restrição foi possível utilizar a matriz de correlação não estruturada (ou não especificada), que representa uma estimativa da correlação amostral. O arquivo de saída das EEG inclui as estimativas

"naive" e robustas dos parâmetros do modelo com os vetores de erro-padrão, a estatística z-escore e valor-p. Vale ressaltar que na última versão desta macro (2.03) é possível calcular a razão de chances ("odds ratio") e seus respectivos intervalos de confiança para os parâmetros do modelo quando a função de ligação escolhida for logito ou log.

Para essa análise, optou-se pela variável resposta indicador da severidade da doença, isto é, o número de episódios severos de diarréia ocorridos no intervalo entre duas suplementações consecutivas[5] . As variáveis independentes utilizadas no modelo multivariado foram: grupo de tratamento (GRUPO), sexo da criança (SEXO), idade da criança em meses (IDADE), tipo de piso da casa (PISO), número de habitantes na casa (HABCASA), escolaridade da mãe (ESCMAE), existência de sanitário na casa (SANIT), como a água de beber é tratada (TRATAGUA), destino dos dejetos (DEJETOS), incremento do indicador antropométrico altura/idade (INCHAZ) e incremento do indicador antropométrico peso/idade (INCWAZ). Como nesta metodologia é possível introduzir variáveis dependentes do tempo, e também porque na metodologia do estudo foi realizada uma avaliação nutricional a cada quatro meses nas crianças, é possível trabalhar com diferentes valores para idade, incremento dos indicadores altura/idade e peso/idade. Portanto, a idade considerada para análise refere-se àquela no início de cada intervalo ou "round"[4] e as variáveis referentes aos incrementos dos indicadores altura /idade e peso/idade são definidos da seguinte forma :

INCHAZ = valor do indicador HAZ no final do intervalo - valor do indicador HAZ no início do intervalo

INCWAZ = valor do indicador WAZ no final do intervalo - valor do indicador WAZ no início do intervalo

Esses indicadores revelam a situação das privações do ambiente sócio-econômico e o efeito das infecções repetidas sobre o estado nutricional. O resultado final das diferenças mencionadas acima são categorizados para esta análise, devido a dificuldade em se definir pontos de cortes adequados à idade no incremento dos indicadores altura/idade e peso/idade.

[5] Neste estudo as crianças foram suplementadas com vitamina A ou placebo a cada 4 meses durante 1 anos, originado assim, 3 intervalos ou rounds (na linguagem epidemiológica).

Assim, um incremento de valor positivo no indicador altura/idade representa um ganho na altura em relação a idade, ou seja, um impacto positivo, mesmo que a criança continue desnutrida segundo o critério do NCHS (< -2 z-score).

Apesar de se ter a idade na sua escala original, a mesma é categorizada por não atender o pressuposto de linearidade no log da variável resposta de interesse. As demais variáveis são independentes do tempo, além de categorizadas.

O modelo utilizado para este conjunto de dados é :

$$\begin{aligned} Log\,E(y_{it}) = {} & intercepto + b_1 SEXO_{it} + b_2 GRUPO_{it} + b_3 SANIT_{it} + b_4 PISO_{it} \\ & + b_5 TRATAGUA_{it} + b_6 DEJETOS_{it} + b_7 INCHAZ_{it} + b_8 INCWAZ_{it} \\ & + b_9 IDADE1_{it} + b_{10} IDADE2_{it} + b_{11} HABCASA1_{it} \\ & + b_{12} HABCASA2_{it} + b_{13} ESCMAE1_{it} + b_{14} ESCMAE2_{it} \end{aligned}$$

sendo y_{it} a resposta do i-ésima criança no j-ésimo intervalo (ou tempo), no caso, o número de episódios severos de diarréia para a i-ésima criança ocorrido no j-ésimo intervalo; SEXO uma covariável categorizada assumindo valor 0 para o sexo masculino e 1 para o sexo feminino; GRUPO uma covariável categorizada assumindo 0 para placebo e 1 para vitamina A; SANIT assumindo 0 se a casa possui sanitário e 1 caso contrário; PISO assumindo valor 0 para piso do tipo cimento, cerâmica, madeira e outros e 1 para piso de terra; TRATAGUA assumindo valor 0 para algum tratamento na água de beber e 1 para nenhum tratamento; DEJETOS assumindo valor 0 para rede de esgoto e 1 para fossa/riacho/superfície; INCHAZ assumindo o valor 1 para um incremento negativo no indicador altura/idade e 0 para um incremento positivo; INCWAZ assumindo o valor 1 para um incremento negativo no indicador peso/idade e 0 para um incremento positivo; IDADE1 e IDADE2 variáveis indicadoras definidas da seguinte forma :

IDADE1	IDADE2	
0	0	criança com 36 ou mais meses de idade
1	0	criança com menos de 24 meses de idade
0	1	criança de 24 à 36 meses de idade

HABCASA1 e HABCASA2 variáveis indicadoras assumindo os seguintes valores :

HABCASA1	HABCASA2	
0	0	se há até 4 moradores na casa
1	0	se há de 5 a 10 moradores na casa
0	1	se há mais de 10 moradores na casa

ESCMAE1, ESCMAE2 variáveis indicadoras definidas da seguinte forma :

ESCMAE1	ESCMAE2	
0	0	mães com segundo grau
1	0	mães com primeiro grau
0	1	mães com primeiro grau incompleto ou nenhuma escolaridade

Inicialmente a metodologia das EEG foi utilizada com três matrizes de correlação distintas de trabalho: Independente, Permutável e Não Estruturada. Os resultados são apresentados na Tabela 8.

Tabela 8. Estimativa e erro-padrão usando as EEG para os dados de Serrinha com três matrizes de correlação de trabalho distintas para o log do número médio de episódios severos de diarréia ocorridos em um intervalo de tempo

	Estrutura de Correlação		
Covariável	Independente	Permutável	Não Estruturada
Intercepto	-3.4326	-3.4682	-3.7464
	(0.372)	(0.374)	(0.407)
Sexo	0.0679	0.0693	0.0332
	(0.134)	(0.134)	(0.142)
Grupo	-0.2584	-0.2649	-0.2047
	(0.134)	(0.133)	(0.145)
Sanit	0.4318	0.4075	0.5429
	(0.155)	(0.154)	(0.163)
Piso	0.1109	0.1228	0.0681
	(0.220)	(0.220)	(0.238)
Tratagua	0.1513	0.1635	0.1289
	(0.147)	(0.147)	(0.158)
Dejetos	0.1189	0.1182	0.1392
	(0.168)	(0.168)	(0.178)
Idade1	1.1763	1.1623	1.1889
	(0.164)	(0.164)	(0.177)
Idade2	0.5709	0.5610	0.5929
	(0.188)	(0.185)	(0.201)
Inchaz	0.2843	0.3244	0.2616
	(0.127)	(0.125)	(0.135)
Incwaz	0.0654	0.0993	0.0976
	(0.120)	(0.117)	(0.131)
Habcasa1	0.0898	0.0932	0.1071
	(0.145)	(0.145)	(0.154)
Habcasa2	-0.0219	-0.0188	0.0617
	(0.215)	(0.215)	(0.250)
Escmae1	0.5577	0.5573	0.6153
	(0.334)	(0.336)	(0.345)
Escmae2	0.3179	0.3268	0.4456
	(0.304)	(0.305)	(0.314)
-log do |matriz de covar.|	57.538933	57.680764	55.921467

Segundo Qu et al., 1995, quando não existe uma correlação muito forte entre as observações, os valores das estimativas dos coeficientes do modelo proposto para as diferentes matrizes de correlação de trabalho não diferem muito. Pode-se ainda, usar um resultado existente na saída das EEG - log do determinante da covariância estimada- para comparar essas estruturas de covariância dentro do mesmo modelo de acordo com Thall & Vail (1990), ou seja, quanto maior este valor melhor é a estrutura de covariância adotada para o modelo .

Diante dos resultados obtidos observou-se uma pequena diferença nas estimativas e erros-padrões para cada estrutura de correlação adotada, revelando assim uma correlação pequena ou moderada. Para esta análise, decidiu-se pelo uso da estrutura de correlação permutável não somente pelo emprego do critério de Thall & Vail, como também por motivos operacionais, já que até o momento só é possível realizar diagnóstico nas EEG quando esta estrutura de correlação for permutável ou independente. Os valores das estimativas, a estatística z-escore e o valor-p são mostrados na Tabela 9.

Tabela 9. Estimativa, erro-padrão, estatística do teste e nível descritivo (robusto) para o log do número médio de episódios severos de diarréia usando a estrutura de correlação permutável

Covariável	Estimativa	Erro-padrão	Z-escore	valor-p
Intercepto	-3.4682	0.374	-9.27	0.0000
Sexo	0.0693	0.134	0.52	0.6056
Grupo	-0.2649	0.133	-1.98	0.0472
Sanit	0.4075	0.154	2.64	0.0082
Piso	0.1228	0.220	0.56	0.5762
Tratagua	0.1635	0.147	1.11	0.2649
Dejetos	0.1182	0.168	0.70	0.4813
Idade1	1.1623	0.164	7.09	0.0000
Idade2	0.5610	0.185	3.04	0.0024
Inchaz	0.3244	0.125	2.59	0.0095
Incwaz	0.0993	0.117	0.85	0.3961
Habcasa1	0.0932	0.145	0.64	0.5205
Habcasa2	-0.0188	0.215	-0.09	0.9301
Escmae1	0.5573	0.336	1.66	0.0968
Escmae2	0.3268	0.305	1.07	0.2847

É observada, portanto a influência de algumas variáveis como: GRUPO, SANIT, IDADE1, IDADE2 e INCHAZ no log do número médio de episódios severos de diarréia. O valor estimado $\beta_2 = -0.2649$ indica que há uma redução no log do número médio de episódios severos de diarréia para o grupo de crianças suplementadas com vitamina A, considerando as outras variáveis fixas (dado que todas as outras variáveis estão no modelo); $\beta_9 = 0.3244$ indica um aumento de 32% no log do número médio de episódios severos de diarréia para crianças com incremento negativo no indicador altura/idade, considerando as outras variáveis fixas. Já $\beta_3 = 0.4075$ revela um aumento no log do número médio de episódios severos de diarréia na ausência de sanitário na casa, considerando as outras variáveis independentes fixas.

Pelo fato das outras variáveis não apresentarem influência significativa no log do número médio de episódios severos de diarréia foi construído o seguinte modelo com finalidade exploratória :

$$Log\, E(y_{it}) = intercepto + b_1\, GRUPO_{it} + b_2\, SANIT_{it} + b_3\, IDADE1_{it} + b_4\, IDADE2_{it} + b_5\, INCHAZ_{it}$$

Os valores das estimativas e do nível descritivo do teste são mostrados na Tabela 10 para a estrutura de correlação de trabalho permutável.

Tabela 10. Estimativa, erro-padrão, estatística do teste e nível descritivo (robusto) para o log do número médio de episódios severos de diarréia usando a estrutura de correlação permutável

Covariável	Estimativa	Erro-padrão	Z-escore	valor-p
Intercepto	-2.9029	0.162	-17.89	0.0000
Grupo	-0.2872	0.135	-2.13	0.0331
Sanit	0.5449	0.140	3.90	0.0001
Idade1	1.1789	0.166	7.12	0.0000
Idade2	0.5689	0.185	3.08	0.0021
Inchaz	0.3424	0.125	2.74	0.0061

Para uma melhor interpretação dos parâmetros do ponto de vista epidemiológico e pelo fato da resposta representar dados de contagem (Poisson), é necessário exponenciar as

estimativas dos parâmetros afim de obter o risco relativo para comparar os níveis das covariáveis significativas no modelo. Na Tabela 11, abaixo, encontram-se as estimativas do risco relativo.

Tabela 11. Estimativa do Risco Relativo e seu intervalo de confiança a 95% para as variáveis do modelo em questão

Covariável	Risco Relativo*	Intervalo de confiança (95%)
Intercepto	-	-
Grupo	0.7504	0.5765-0.9766
Sanit	1.7244	1.3207-2.2513
Idade1	3.2509	2.3312-4.5338
Idade2	1.7664	1.2372-2.5218
Inchaz	1.4084	1.1126-1.7828

*Aproximado pela razão de chances (odds ratio)

Examinando os valores das estimativas do risco relativo, na Tabela 11, pode-se concluir que há um efeito protetor do grupo suplementado em relação ao placebo, ou seja, uma redução de 25% no número médio de episódios severos de diarréia para as crianças suplementadas, supondo as outras variáveis independentes fixas. Para a variável sanitário observa-se um risco de 1.72 vezes maior de apresentar episódios severos de diarréia quando a casa não possui sanitário quando comparado àquelas com presença de sanitário. Para a criança com menos de 24 meses de idade observa-se um risco 3.25 vezes maior de ocorrer episódios severos de diarréia quando comparado com crianças maiores de 36 meses de idade. Já com crianças de 24 a 36 meses esse risco diminui para 1.77, isto é, crianças de 24 a 36 meses tem um risco 1.77 vezes maior de apresentar episódios severos de diarréia comparado àquelas com idade superior a 36 meses. Com relação ao incremento do indicador, há um risco 1.41 vezes maior da criança de ocorrer episódios severos de diarréia quando a mesma apresenta um incremento negativo no indicador altura/idade comparado àquelas com incremento positivo.

4.3.3 Diagnóstico nas EEG

Como mencionado no capítulo III, é apresentada uma breve ilustração das recentes técnicas de diagnósticos nas EEG. Os resultados baseiam-se na proposta de Preisser & Qaqish (1996), cujas medidas de influência apoiam-se na generalização da aproximação a um passo de modelos lineares generalizados.

Com o objetivo exploratório, alguns gráficos são construídos para observar pontos de alavanca (alto "leverage") e medir a influência de uma criança (conglomerado) sobre as estimativas dos parâmetros de regressão e sobre o ajuste global. São utilizadas as estatísticas $DBETAC_i$, $DCLS_i$, a matriz de projeção H_i e $GCLS_i$.

É importante lembrar que o diagnóstico apresentado aqui refere-se a exclusão do conglomerado, no caso, a criança e a matriz de correlação assumida é a permutável. Portanto, k=899 crianças e $corr(Y_{it}, Y_{it'}) = \rho$, para i=1,2,....,k e $t \neq t'$=1,2,....n_i são considerados.

O programa, ainda na versão de teste, utilizado para esta análise foi gentilmente cedido pelo pesquisador. Trata-se de uma macro do SAS bem similar a macro de Karim & Zeger para as equações de estimação generalizadas. Com esta nova macro é possível ajustar modelos pelas EEG e também obter as estatísticas para o diagnóstico caso a estrutura de correlação adotada seja a permutável ou independente.

O efeito do i-ésimo conglomerado (criança) sobre o ajuste geral pode ser visto no Gráfico 1 da distância de Cook generalizada ($DCLS_i$) versus identificação da criança. Observa-se que as crianças de número 51, 65, 347, 968 e 1212 exercem uma grande influência no ajuste global. Os demais pontos rotuladas têm uma moderada influência.

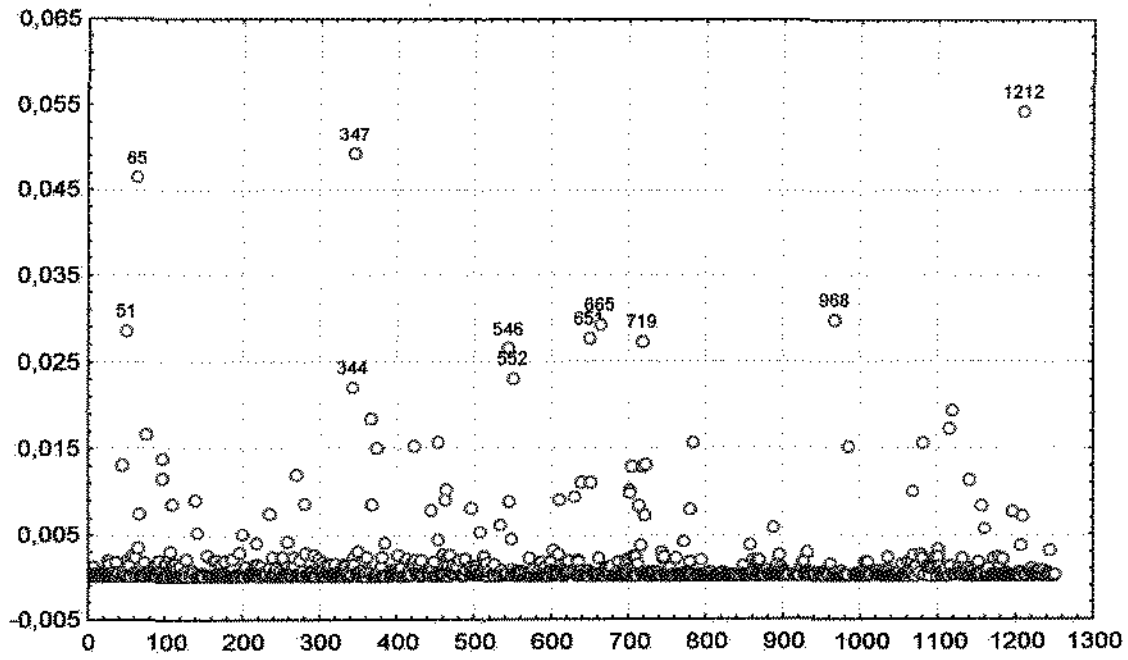


Gráfico 1: Distância de Cook generalizada por criança (conglomerado)

No gráfico 2, observa-se que 3 conglomerados (crianças) têm grande influência, sendo que a criança 1212 (grupo placebo) possui maior leverage. Os demais têm uma moderada influência. Desses pontos, 55% são crianças pertencentes ao grupo placebo.

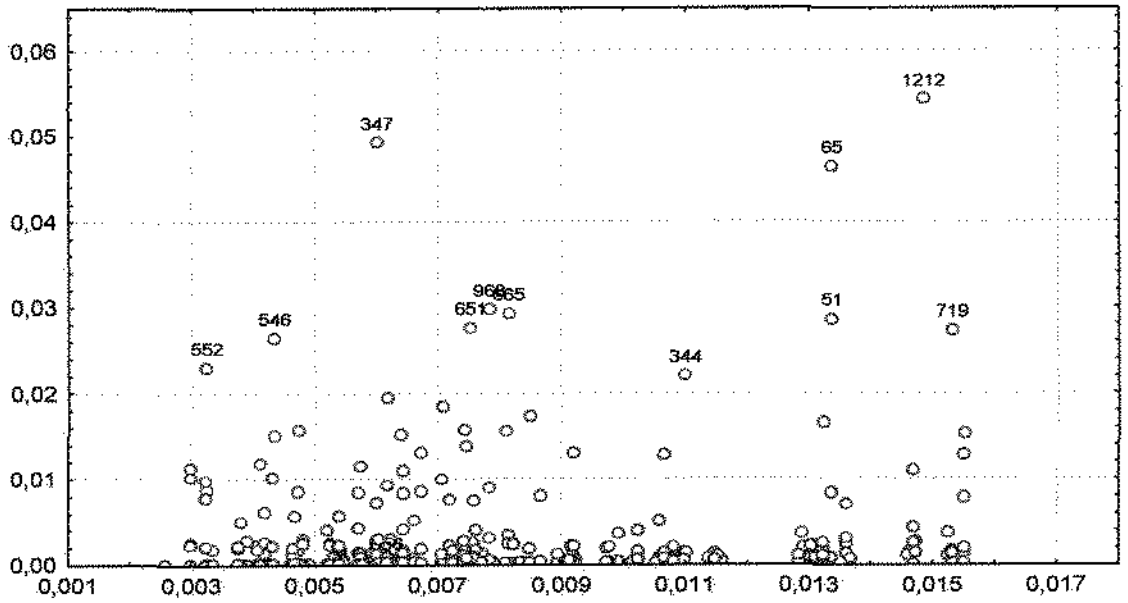


Gráfico 2: Distância de Cook versus o leverage da criança (conglomerado)

O leverage versus a identificação da criança é apresentado no gráfico 3. Não foi detectado pontos específicos com altos leverage e sim uma faixa de valores mais acentuada do restante dos pontos.

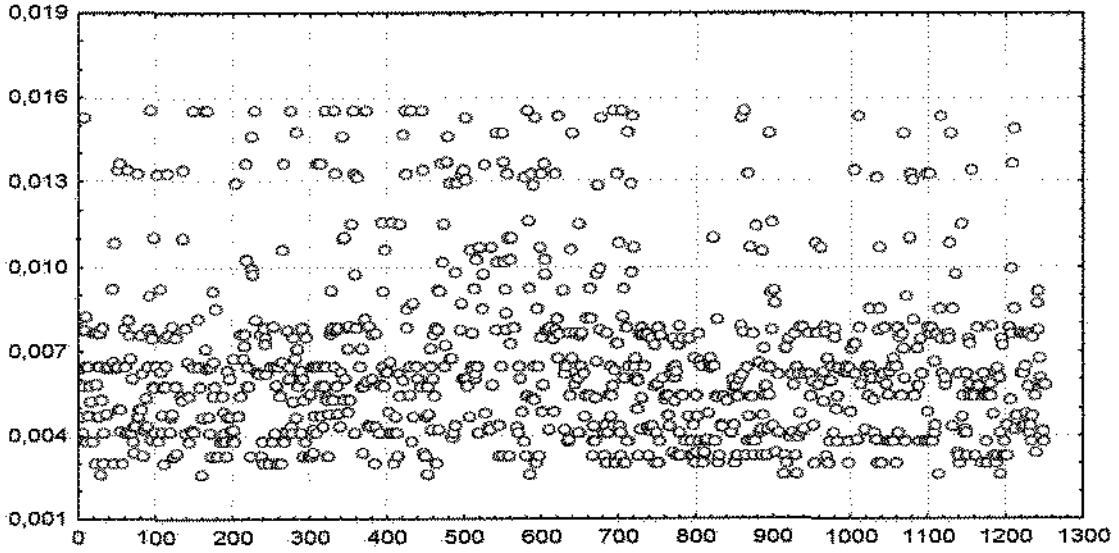


Gráfico 3: Pontos de leverage para cada criança

A influência das observações sobre os parâmetros e variâncias dos mesmos, simultaneamente é apresentada no gráfico 4. Observa-se que as criança 51e 719 (placebo) têm uma influência mais acentuada em relação ao restante. Dos pontos marcados 90% referem-se a crianças do grupo placebo.

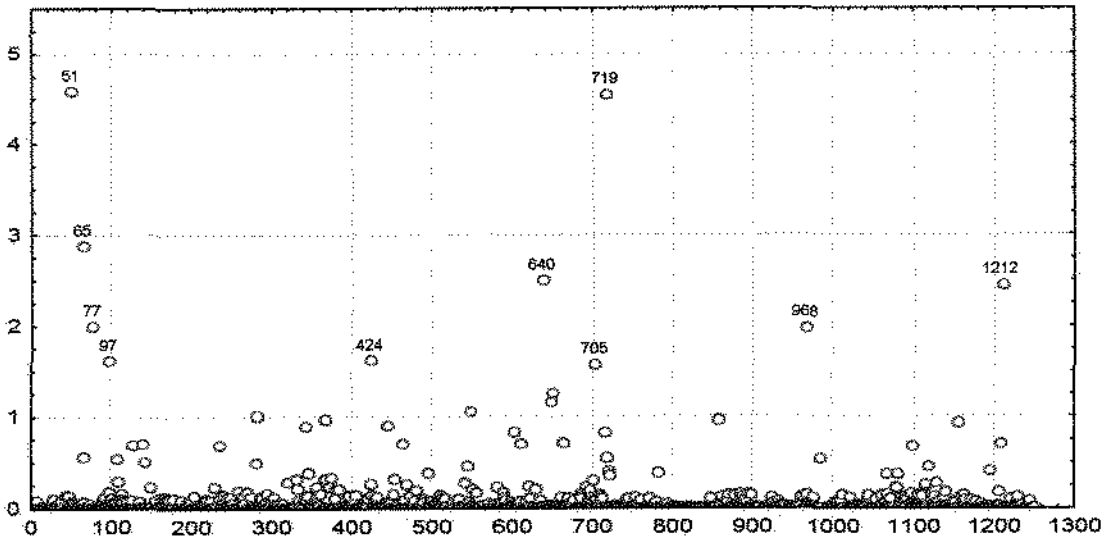


Gráfico 4: Generalização do DFITS versus a criança (conglomerado)

A Tabela 13 mostra o impacto, em termos percentuais, nas estimativas dos coeficientes ao se excluir alguns conglomerados (crianças) identificados como influentes nos gráficos acima.

Tabela 13. Estimativas e (mudança em percentual) dos coeficientes referentes ao modelo contendo todos os dados e ao modelo obtido excluindo algum conglomerado para o log do número médio de episódios severos de diarréia

Covariável	Todos os dados	Exclusão do conglomerado: 51	65	347	719	1212
Intercepto	-2.9029	-2.9029 (-0.37%)	-2.9015 (-0.05%)	-2.9478 (1.54%)	-2.8919 (-0.38%)	-2.9151 (0.42%)
Grupo	-0.2872	-0.2711 (-5.60%)	-0.2703 (-5.87%)	-0.3116 (8.48%)	-0.2685 (-6.51%)	-0.2588 (-9.90%)
Sanit	0.5449	0.5218 (-4.23%)	0.5213 (-4.33%)	0.5164 (-5.24%)	0.5188 (-4.79%)	0.5033 (-7.64%)
Idade1	1.1789	1.1558 (-1.96%)	1.1884 (0.79%)	1.2273 (4.10%)	1.1598 (-1.62%)	1.1575 (-1.82%)
Idade2	0.5689	0.5783 (-1.65%)	0.5252 (-7.69%)	0.6159 (8.26%)	0.5688 (-0.02%)	0.5546 (-2.51%)
Inchaz	0.3424	0.3296 (-3.76%)	0.3372 (-1.52%)	0.3695 (7.89%)	0.3276 (-4.33%)	0.3647 (6.50%)

O impacto nos coeficientes estimados do modelo em questão é pequeno; no geral em torno de 10%. Observa-se ainda na Tabela 13 que os conglomerados (crianças) 347 e 1212 possuem em média impacto maior nos coeficientes estimados. Buscando maiores detalhes sobre essas duas crianças observa-se que elas pertencem ao grupo placebo, as casas onde residem não possuem sanitário, o número total de episódios ocorridos no ano para cada criança foi respectivamente 10 e 13 sendo que 4 severos para a criança 347 e 7 severos para a criança 1212. As idades das mesmas no início do estudo eram 41 e 16 meses. É interessante observar que os efeitos dessas duas crianças (347 e 1212) agem de maneiras opostas em relação ao grupo que pertencem (placebo), de forma que há uma compensação. A Tabela 14, apresenta a

estimativa do risco relativo para o modelo contendo todas as observações e para o modelo exluindo os casos influentes.

Tabela 14. Estimativa do risco relativo (I.C. a 95%) referentes aos modelos contendo todos os dados e ao modelo excluindo algum conglomerado para o log do número médio de episódios severos de diarréia

Covariável	Todos os dados	Exclusão do conglomerado:				
		51	65	347	719	1212
Grupo	0.7504	0.763	0.763	0.732	0.765	0.772
	0.5765-0.9766	0.585-0.993	0.586-0.993	0.563-0.953	0.587-0.996	0.594-0.993
Sanit	1.7244	1.685	1.684	1.676	1.680	1.654
	1.3207-2.2513	1.289-2.204	1.288-2.202	1.276-2.201	1.276-2.212	1.264-2.165
Idade1	3.2509	3.177	3.282	3.412	3.189	3.182
	2.3312-4.5338	2.278-4.430	2.356-4.572	2.470-4.713	2.304-4.414	2.300-4.401
Idade2	1.7664	1.783	1.691	1.851	1.766	1.741
	1.2372-2.5218	1.251-2.541	1.181-2.421	1.289-1.659	1.229-2.538	1.209-2.508
Inchaz	1.4084	1.390	1.401	1.447	1.388	1.440
	1.1126-1.7828	1.089-1.776	1.106-1.774	1.135-1.845	1.086-1.773	1.125-1.843

Pode-se concluir que o modelo adotado descreve bem a ocorrência de episódios severos de diarréia em função de algumas covariáveis, além de discriminar bem as crianças suplementadas com vitamina A das crianças que receberam placebo.

4.4 Estudo AISAM

Segundo Moraes (1996), devido às condições precárias de moradia das encostas e vales como falta de esgotos sanitários, drenagem e contenção das encostas, habitadas principalmente pela população de baixa renda, a Prefeitura Municipal de Salvador em 1979 iniciou estudos visando definir soluções para alguns desses problemas. A bacia do Camurujipe foi indicada como a área a reclamar uma ação mais urgente. Esta área é habitada por uma população de cerca de 800 mil habitantes de baixa renda distribuídos em 34 agrupamentos. O

projeto de saneamento básico instalado nessa área foi caracterizado por um sistema de rampas e escadarias drenantes, sistema não convencional de drenagem de águas pluviais usadas também para o escoamento dos esgotos e a implementação, em alguns agrupamentos, de uma rede simplificada de esgotamento sanitário.

A pesquisa AISAM foi desenvolvida em três comunidades que possuíam facilidades de saneamento através de escadarias e rampas drenantes (Grupo 2), três que tinham a rede simplificada de esgotamento sanitário, além da solução anterior (Grupo 3), e três que ainda não tinham sofrido intervenção de saneamento (Grupo 1) selecionadas ao acaso. A Tabela 15, contém a relação das comunidades, do número de casas e crianças selecionadas para a pesquisa.

O desenho é do tipo longitudinal, com 1162 crianças menores de 5 anos para o estudo da morbidade por diarréia. Com o objetivo de coletar as histórias de diarréia das crianças, um sistema de registro diário através de um calendário quinzenal foi implementado. Também foram levantadas informações da causa, dos sintomas de diarréia e tratamento aplicado pelos entrevistadores de campo. A diarréia foi definida pela mãe em face a sua percepção ao longo da vida, de quando seus filhos estavam com diarréia. Um episódio de diarréia foi definido como um ou mais dias com diarréia separado de qualquer outro episódio por pelo menos 2 dias livres do sintoma de diarréia, utilizando a definição em Morris *et al.*, 1994. Uma indicação da severidade de diarréia foi obtida questionando o número máximo de evacuações num intervalo de 24 horas. Tomando como base essa última informação, foi adotado um critério do grau de severidade do episódio de diarréia, isto é, um episódio leve de diarréia foi definido pela duração de até dois dias; moderado quando durava 3 ou mais dias, e com um número máximo de 4 evacuações num dia e severo quando durava 3 ou mais dias e com um número mínimo de 5 ou mais evacuações num dia.

Tabela 15. Relação das comunidades, do número de casas e crianças selecionadas para o estudo AISAM

Grupo de estudo	Comunidades	Nº de casas selecionadas	Nº de crianças selecionadas	
			< 5 anos	5–14anos
Grupo 1 (controle)	Arraial de Baixo	108	126	210
	Baixa Camarujipe	99	132	211
	Nova Divinéia	115	138	210
Total	3	322	388	631
Grupo 2 (intermediário)	Antônio Balbino	116	125	210
	Bom Juá	105	126	210
	Santa Mônica	124	133	211
Total	3	345	384	631
Grupo 3 (padrão)	B. Vista S. Caetano	106	126	211
	Jardim Caiçara	118	134	210
	Sertanejo	114	130	210
Sub-total	3	338	390	631
Total geral	9	1005	1162	1893

Fonte : Moraes (1996)

Pelas análises realizadas por Moraes (1996), nenhuma diferença significativa foi observada nos grupos em relação à distribuição da idade e sexo das crianças. Com relação aos indicadores sócio-econômico, demográfico, cultural e ambiental algumas diferenças foram observadas como por exemplo : o tempo médio de escolaridade do cabeça da família no grupo 1 foi menor em relação aos outros dois grupos; o mesmo para a escolaridade das mães, cobertura da coleta de lixo, pavimentação da rua e escoamento do esgoto, como era esperado. Por outro lado, nenhuma diferença foi encontrada nos três grupos em relação ao acesso à eletricidade, número de quartos, tamanho da casa, presença de animais na casa e tipo de abastecimento de água. A renda média per capita foi menor no grupo 1. Uma indicação da experiência sócio-cultural foi avaliada pela origem do cabeça da família e da mãe, nesse caso um percentual maior dessas pessoas oriundas da área rural foi encontrada no grupo 1.

Antes de apresentar os resultados obtidos das análises usando as metodologias de razão e as EEG, são resumidos na Tabela 16 os resultados encontrados em Moraes (1996). A

medida da morbidade de diarréia usada foi incidência, isto é, a taxa de incidência expressa como os episódios por crianças-ano, calculadas da razão entre o número de novos episódios ocorridos e o número de crianças-ano de observação obtida dos calendários.

Tabela 16. Morbidade por diarréia em crianças menores de 5 anos por grupo de estudo, Nov/89-Nov/90

Criança-dia de observação	
Grupo 3	114.305
Grupo 2	111.721
Grupo 1	118.217
Incidência geral (episódios/crianças-ano)	
Grupo 3	1.73
Grupo 2	3.32
Grupo 1	5.55
Razão das densidades de incidência	
G3/G1	0.31
G2/G1	0.60
G3/G2	0.52
% de dias com diarréia	
Grupo 3	1.9
Grupo 2	3.5
Grupo 1	4.9

Fonte : Moraes (1996)

Estes resultados mostram principalmente que crianças residentes nas comunidades com um nível melhor de saneamento (grupo 3) tiveram 69% menos episódios de diarréia do que aquelas residentes nas comunidades sem nenhum tipo de intervenção sanitária (grupo 1). Observou-se também uma diferença significativa na proporção média de dias com diarréia na comparação dos grupos.

4.4.1 Método da razão e mínimos quadrados ponderados

O estudo AISAM tem uma estrutura de delineamento complexa que se enquadra na proposta da metodologia da razão. As comunidades do vale Camurujipe foram estratificadas com relação ao tipo de condição de saneamento (no caso o grupo). Posteriormente, uma seleção aleatória das comunidades (conglomerado) com reposição no primeiro estágio foi efetuada e em seguida domicílios foram selecionados ao acaso dentro dessas comunidades, no segundo estágio. Todas as crianças menores de 5 anos residentes nesses domicílios foram acompanhadas para o estudo da morbidade por diarréia.

Para esse delineamento amostral, será utilizada uma estimativa do tipo razão, mais especificamente, a densidade de incidência de diarréia. Na verdade será construído um vetor dessa razão, onde cada elemento desse vetor representa uma incidência de diarréia moderada e/ou severa referente a uma classificação cruzada dos níveis da característica do delineamento.

O esquema de amostragem do estudo (Moraes, 1996) se enquadra num planejamento alto-ponderado pois as comunidades foram selecionadas proporcionalmente ao tamanho e além de um número aproximadamente constante de domicílios escolhidos ao acaso de cada comunidade (cerca de 120).

Numa primeira etapa um vetor de razão será construído referente a classificação cruzada das características do delineamento, ou seja, das características referentes às unidades primárias de seleção (comunidades), as unidades secundárias (domicílios) e a unidade amostral de análise (criança) simultaneamente; no caso o vetor de incidência de episódios moderados e/ ou severos de diarréia, por crianças-dia. Posteriormente, um modelo log-linear para a regressão da razão será adotado utilizando a estimação de mímimos quadrados ponderados.

Utilizou-se as características com resultados significativos encontrados em análises preliminares e também em Moraes (1996). Assim a característica concernente às unidades primárias de seleção, isto é, comunidades, diz respeito à condição de saneamento a que as mesmas se enquadram, ou seja, aos grupos 1, 2 e 3. As características concernentes às unidades secundárias, domicílios, escolhidas foram : número de cômodos na casa, se a casa tinha ou não sanitário, se a cozinha era utilizada só para preparar os alimentos e o tipo de acondicionamento

do lixo. Essas variáveis formam um conjunto de características referentes aos domicílios. Já para as características relativas às crianças foram selecionadas: idade da criança, sexo da criança, escolaridade da mãe da criança. Em cada um dos conjuntos das características do delineamento citadas acima pode haver uma combinação dos níveis das variáveis que compõem os mesmos. Entretanto deve-se ter cautela em relação ao uso dessas combinações por causa das frequências resultantes das caselas na classificação cruzada dessas características.

Como resultado dessa primeira etapa, na Tabela 17 abaixo, são apresentadas as densidades de incidência de episódios moderados e/ou severos de diarréia, por crianças-dia, observadas. Foi construído um vetor das incidências de diarréia moderada e/ou severa, correspondente a classificação cruzada de grupo, uso da cozinha e sexo da criança.

Como passo seguinte, o modelo de regressão para o log das razões, isto é, o log da densidades de incidência, é ajustado utilizando mínimos quadrados ponderados :

$$E(Log\,R) = \beta_0 + \beta_1 GRUPO1 + \beta_2 GRUPO2 + \beta_3 COZINHA + \beta_4 SEXO$$

onde GRUPO1 e GRUPO2 são variáveis indicadoras para o grupo 1 (controle) e grupo 2 (intermediário) respectivamente; COZINHA é uma variável indicadora assumindo o valor 0 quando o uso da cozinha é só para cozinhar alimentos e 1 para outros tipos de uso; SEXO é uma outra variável indicadora assumindo o valor 0 quando a criança for do sexo masculino e 1 quando for do sexo feminino.

Tabela 17. Densidades de incidência (razões) e erros-padrões, para episódios moderados e/ou severos de diarréia, observados para a classificação cruzada simultânea de grupo, uso da cozinha e sexo da criança

Grupo	Uso da cozinha	Sexo da criança	R=DI observadas por crianças-dia ($x\ 10^{-3}$)	Desvio Padrão(R) ($x\ 10^{-3}$) *
Grupo 1 (controle)	Só para cozinhar	Masculino	7.6523	2.4197
		Feminino	6.0367	2.2361
	Outros usos	Masculino	9.5446	2.9432
		Feminino	6.9759	3.0492
Grupo 2 (intermed.)	Só para cozinhar	Masculino	3.5344	0.7436
		Feminino	3.0334	0.8298
	Outros usos	Masculino	7.5335	0.9924
		Feminino	5.6421	1.4504
Grupo 3 (padrão)	Só para cozinhar	Masculino	2.9991	1.0723
		Feminino	2.4095	0.5266
	Outros usos	Masculino	3.7781	0.6194
		Feminino	3.3799	0.7782

*Calculados usando expansão em Série de Taylor

Na Tabela 18, é apresentado um quadro de análise de variância e avaliação da bondade de ajuste do modelo. Pelos resultados, observa-se que a variável sexo não é significante, ao nível de 5%, no modelo adotado.

Tabela 18. Tabela de Análise de Variância para o modelo de regressão

	Grau de liberdade	Estatística χ^2	p-valor
Intercepto	1	4648.26	0.0000
Grupo	2	23.14	0.0000
Cozinha	1	10.86	0.0010
Sexo	1	1.70	0.1920
Residual	7	3.11	0.8751

Um novo modelo é ajustado sem a variável que representa o sexo da criança,

$$E(Log\ R) = \beta_0 + \beta_1 GRUPO1 + \beta_2 GRUPO2 + \beta_3 COZINHA$$

O quadro da Análise de Variância do modelo reduzido é apresentado na Tabela 19 abaixo. A estatística de bondade de ajuste indica que o modelo representa adequadamente os dados.

Tabela 19. Tabela de Análise de Variância para o modelo de regressão reduzido

	Grau de liberdade	Estatística χ^2	p-valor
Intercepto	1	4799.01	0.0000
Grupo	2	25.73	0.0000
Cozinha	1	13.81	0.0002
χ^2 de bondade de ajuste (8 gl)	8	4.81	0.7779

Os resultados das estimativas dos coeficientes do modelo reduzido encontram-se na Tabela 20. Observa-se um risco 2.5 vezes maior de ocorrer episódios moderados e/ou severos de diarréia quando se está no grupo 1 (grupo controle- sem intervenção) comparado com o grupo 3. Pode-se concluir também que o risco de apresentar episódios moderados e/ou severos de diarréia é 1.68 vezes maior quando a cozinha é utilizada para outros fins comparado àquelas que só utilizam para cozinhar alimentos.

Tabela 20. Estimativa e erro-padrão dos efeitos dos fatores pelo método de regressão da razão

Efeito	Estimativas ($\hat{\beta}$)	Erro-padrão	valor-p	RDI [a]	IC a 95% [b]
Intercepto	-6.0785	0.1425	0.0000		
Grupo 1	0.9618	0.2067	0.0000	2.614	1.745-3.923
Grupo 2	0.5336	0.1449	0.0002	1.705	1.284-2.265
Cozinha (Outros)	0.5188	0.1396	0.0002	1.680	1.278-2.209

[a] Razão das densidades de incidência = risco relativo
[b] Intervalo de confiança para o risco relativo

Na Tabela 21, encontram-se as razões e os erros-padrões preditos pelo modelo final adotado.

Tabela 21. Densidades de incidência (razões) e erros-padrões preditos pelo modelo de regressão reduzido

Grupo	Uso da cozinha	$\hat{R}$ preditas (x 10^{-3})	DP($\hat{R}$) (x 10^{-3})
Grupo 1 (controle)	Só para cozinhar	5.9956	1.1168
	Outros usos	10.0731	1.9001
Grupo 2 (intermed.)	Só para cozinhar	3.9071	0.5212
	Outros usos	6.5642	0.6987
Grupo 3 (padrão)	Só para cozinhar	2.2915	0.3265
	Outros usos	3.8500	0.4560

Como mencionado no capítulo II, combinações do níveis das variáveis que compõem uma característica do delineamento são possíveis de serem construídas. Assim, a título de exploração dos dados, uma nova classificação cruzada das características do delineamento é montada, Tabela 22, com um vetor das incidências de episódios moderados e/ou severos de diarréia, correspondente a classificação cruzada simultânea de grupo, se trata a água de beber e uma combinação de sexo e idade das crianças.

Tabela 22. Densidades de Incidência (razões) e erros-padrões, para episódios moderado e/ou severo de diarréia, observados para a classificação cruzada simultânea de grupo, tratamento da água de beber e sexo com idade

Grupo	Tratamento da água de beber	Combinação de sexo e idade da criança	R=DI observadas ($x\ 10^{-3}$)	Desvio Padrão(R) ($x\ 10^{-3}$)
Grupo 1 (controle)	Nenhum	Masculino de 24 a 59 meses	9.5177	4.3128
		Masculino de 0 a 23 meses	11.1594	2.8968
		Feminino de 24 a 59 meses	6.8647	3.4351
		Feminino de 0 a 23 meses	8.6000	1.9555
	Algum	Masculino de 24 a 59 meses	6.8231	2.3104
		Masculino de 0 a 23 meses	8.4024	1.1499
		Feminino de 24 a 59 meses	4.9030	1.8547
		Feminino de 0 a 23 meses	4.2998	1.6062
Grupo 2 (interm.)	Nenhum	Masculino de 24 a 59 meses	5.9941	0.6865
		Masculino de 0 a 23 meses	11.3095	3.5916
		Feminino de 24 a 59 meses	4.4409	1.7086
		Feminino de 0 a 23 meses	6.5699	3.7014
	Algum	Masculino de 24 a 59 meses	3.6644	0.7614
		Masculino de 0 a 23 meses	6.0060	0.7934
		Feminino de 24 a 59 meses	2.9959	0.7345
		Feminino de 0 a 23 meses	5.6914	1.6668
Grupo 3 (padrão)	Nenhum	Masculino de 24 a 59 meses	1.7699	0.9490
		Masculino de 0 a 23 meses	7.2819	4.4609
		Feminino de 24 a 59 meses	2.2248	1.0862
		Feminino de 0 a 23 meses	6.3364	1.9287
	Algum	Masculino de 24 a 59 meses	1.9001	0.4087
		Masculino de 0 a 23 meses	4.1075	1.4389
		Feminino de 24 a 59 meses	1.7029	0.3724
		Feminino de 0 a 23 meses	3.6587	0.7823

Como passo seguinte, o modelo de regressão abaixo para o log das densidades de incidência, é ajustado utilizando mínimos quadrados ponderados :

$$E(Log\ R) = \beta_0 + \beta_1\ GRUPO1 + \beta_2\ GRUPO2 + \beta_3\ TRATAGUA + \beta_4\ CRIANÇA2$$
$$\beta_5\ CRIANÇA3 + \beta_6\ CRIANÇA4$$

onde GRUPO1 e GRUPO2 são variáveis indicadoras para o grupo 1 (controle) e grupo 2 (intermediário) respectivamente; TRATAGUA é uma variável indicadora assumindo o valor 0 quando não se faz nenhum tratamento na água e 1 quando se faz algum tratamento e CRIANÇA2, CRIANÇA3, CRIANÇA4 são variáveis indicadoras para sexo masculino de 0 à 23 meses, sexo feminino de 24 à 59 meses e sexo feminino de 0 à 23 meses respectivamente.

Na Tabela 23, é apresentado o quadro de análise de variância e a bondade de ajuste do modelo. Pelos resultados, observa-se que todas as variáveis são significativas e que o modelo se ajusta bem aos dados (p=0.9015). Portanto, na Tabela 24, são apresentadas as estimativas e erros-padrões das variáveis do modelo, juntamente com os riscos relativos e respectivos intervalos de confiança.

Tabela 23. Tabela de Análise de Variância para o modelo de regressão

	Grau de liberdade	Estatística χ^2	p-valor
Intercepto	1	8917.89	0.0000
Grupo	2	26.49	0.0000
Tratagua	1	14.73	0.0001
Criança	3	27.34	0.0000
Residual	17	8.84	0.9451

Tabela 24. Estimativa e erro-padrão dos efeitos dos fatores baseado no método de regressão da razão

Efeito	Estimativas ($\hat{\beta}$)	Erro-padrão	valor-p	RDI [a]	IC a 95% [b]
Intercepto	-5.6702	0.1521	0.0000	-	
Grupo 1	0.7678	0.1507	0.0000	2.155	1.604-2.896
Grupo 2	0.5216	0.1374	0.0001	1.685	1.287-2.205
Tratagua	-0.4295	0.1119	0.0001	0.651	0.523-0.810
Criança2	0.5170	0.1361	0.0001	1.677	1.284-2.189
Criança3	-0.2269	0.1629	0.1638	0.797	0.579-1.097
Criança4	0.3301	0.1546	0.0327	1.391	1.027-1.883

[a] Razão das densidades de incidência = risco relativo

[b] Intervalo de confiança para o risco relativo

Observa-se nos resultados da Tabela 24 que os fatores Grupo1, Grupo2, Tratagua, Criança2 e Criança4 são significativos no modelo. Como os resultados da última coluna da tabela representam as estimativas do risco relativo oriundas da exponenciação dos coeficientes, percebe-se que há um risco 2.2 vezes maior de ocorrer episódios moderados e/ou severos de diarréia quando se está no grupo sem nenhum tipo de intervenção de saneamento comparado com o grupo padrão (grupo 3). O mesmo acontece no grupo de condição intermediário, ou seja, um risco 1.7 vezes maior de apresentar episódios moderados e/ou severos de diarréia quando comparado com o grupo 3. Há uma redução de 35% na incidência de diarréia moderada e/ou severa quando se utiliza a água de beber com algum tratamento. Quando se tem crianças do sexo masculino na faixa etária de 0 à 23 meses há um risco 1.8 vezes maior de ocorrer episódios moderados e/ou severos de diarréia comparado com as de mesmo sexo na faixa etária de 24 à 59 meses. O mesmo acontece com crianças do sexo feminino na faixa etária de 0 à 23 meses.

Na tabela 25 encontram-se as estimativas das razões preditas pelo modelo da classificação cruzada de grupo, tratamento de água e a combinação de sexo e idade da criança. Observa-se que os erros-padrões preditos pelo modelo são bem menores do que os observados.

Tabela 25. Densidade de incidência (razão) e erro-padrão, para episódios moderados e/ou severos de diarréia, preditos pelo modelo para a classificação cruzada simultânea de grupo, tratamento da água de beber e sexo com idade

Grupo	Tratamento da água de beber	Combinação de sexo e idade da criança	R=DI preditas pelo modelo ($x 10^{-3}$)	Desvio Padrão(R) ($x 10^{-3}$)
Grupo 1 (controle)	Nenhum	Masculino de 24 a 59 meses	7.7283	1.0424
		Masculino de 0 a 23 meses	12.4575	1.6978
		Feminino de 24 a 59 meses	5.9205	1.0440
		Feminino de 0 a 23 meses	10.3334	1.5429
	Algum	Masculino de 24 a 59 meses	4.8346	0.7488
		Masculino de 0 a 23 meses	8.1077	0.8655
		Feminino de 24 a 59 meses	3.8533	0.6441
		Feminino de 0 a 23 meses	6.7256	1.0461
Grupo 2 (interm.)	Nenhum	Masculino de 24 a 59 meses	5.8072	0.5699
		Masculino de 0 a 23 meses	9.7388	1.3288
		Feminino de 24 a 59 meses	4.6284	0.7536
		Feminino de 0 a 23 meses	8.0786	1.2297
	Algum	Masculino de 24 a 59 meses	3.7795	0.4416
		Masculino de 0 a 23 meses	6.3383	0.6695
		Feminino de 24 a 59 meses	3.0123	0.4576
		Feminino de 0 a 23 meses	5.2578	0.8277
Grupo 3 (padrão)	Nenhum	Masculino de 24 a 59 meses	3.4470	0.5242
		Masculino de 0 a 23 meses	5.7808	1.0791
		Feminino de 24 a 59 meses	2.7474	0.4944
		Feminino de 0 a 23 meses	4.7952	0.7580
	Algum	Masculino de 24 a 59 meses	2.2434	0.3205
		Masculino de 0 a 23 meses	3.7623	0.5407
		Feminino de 24 a 59 meses	1.7881	0.2664
		Feminino de 0 a 23 meses	3.1209	0.4399

4.4.2 As equações de estimação generalizadas e extensões

Katz *et al.* (1993) utilizaram a metodologia de RLA para estimar o grau de associação da diarréia entre crianças de uma mesma família e/ou mesma vila. Por se tratar de um estudo longitudinal, a aplicação desta metodologia ao conjunto de dados AISAM, será feita explorando a questão da medida repetida na estimação dos parâmetros de um modelo de regressão. A medida de morbidade a ser utilizada nesta análise será a ocorrência de pelo menos um episódio de diarréia em cada uma das 26 quinzenas de acompanhamento. Devido à estrutura dos dados e das limitações com relação ao tamanho do conglomerado (conjunto de observações repetidas por unidade de análise) existente nas EEG1 e EGG2, optou-se pelo uso da metodologia de Regressão Logística Alternada (RLA) para estimar simultaneamente os parâmetros de um modelo logístico e um parâmetro de associação comum entre todos os pares de resposta das crianças, representado pela razão de chances marginal.

O modelo utilizado para esta metodologia neste conjunto de dados é :

$$\begin{aligned} Logito\ P(Y_{it}=1) = {} & intercepto + b_1 SEXO_{it} + b_2 GRUPO2_{it} + b_3 GRUPO3_{it} + \\ & b_4 COMODOS_{it} + b_5 TRATAGUA_{it} + b_6 ONDE1_{it} + b_7 ONDE2_{it} \\ & b_8 IDADE1_{it} + b_9 IDADE2 \end{aligned}$$

onde Y_{it} é a resposta da i-ésima criança na t-ésima quinzena, isto é, $Y_{it}=1$ se a criança apresenta pelo menos um episódio de diarréia na quinzena t e $Y_{it}=0$ caso contrário; SEXO uma covariável categorizada assumindo valor 0 para o sexo masculino e 1 para o sexo feminino; GRUPO2 e GRUPO3 são variáveis indicadoras para o grupo 2 (intermediário) e grupo 3 (padrão) respectivamente; COMODOS variável categorizada assumindo 0 quando a casa tem mais de 3 cômodos e 1 quando a casa tem até 3 cômodos; TRATAGUA assumindo o valor 0 para nenhum tratamento da água de beber e 1 para algum tratamento; ONDE1 e ONDE2 variáveis indicadoras representando médio e nenhum acondicionamento do lixo da casa respectivamente e IDADE1 e IDADE2 variáveis indicadoras para as idades de 0 à 11 meses e de 12 à 23 meses respectivamente.

Para a estrutura de associação, assume-se o modelo mais simples para o log da razão de chances entre pares de observação, ou seja, uma estrutura de dependência permutável que representa uma razão de chances constante entre todos os pares de observações dentro do conglomerado e igual para todos os indivíduos. A estrutura de associação pode ser ajustada como uma função de algumas variáveis explanatórias, obtendo-se assim um modelo mais genérico para o log da razão de chances.

Na Tabela 26 são apresentadas as estimativas e erros-padrões das variáveis do modelo, juntamente com o valor da estatística z.

Tabela 26. Estimativa, erro-padrão e estatística do teste (robusta) para o modelo de Regressão Logística Alternada

Covariáveis	Estimativas ($\hat{\beta}$)	Erro-padrão (robusto)	Estatística z
Intercepto	-2.0096958	0.1558926	-12.891542
Sexo (F)	-0.2416558	0.1001837	-2.412127
Grupo2	-0.3647601	0.1271348	-2.869081
Grupo3	-0.9446783	0.1243253	-7.598437
Comodos	0.3594711	0.1005667	3.574453
Tratagua	-0.3889396	0.1230805	-3.160043
Onde1	0.2695910	0.1118571	2.410137
Onde2	0.6976098	0.1746578	3.994152
Idade1	0.5021383	0.1405925	3.571587
Idade2	0.4613339	0.1130163	4.082012
Log OR	1.019966	0.07686889	13.26891

Observa-se nos resultados da Tabela 26 que todas as covariáveis do modelo representam fatores de risco significantes para a presença de diarréia. A probabilidade da criança apresentar diarréia é menor no sexo feminino, isto é, a chance de uma criança do sexo feminino ter diarréia é 21% menor, considerando as outras covariáveis fixas. Já com relação ao grupo, a razão de chances da criança apresentar diarréia é 1.44 (1/exp(-0.3647)) vezes maior

quando a mesma pertence ao grupo 1 comparado com uma criança do grupo 2 e ainda 2.53 (1/exp(-0.9447)) vezes maior quando a criança é do grupo 1 em relação ao grupo 3, ou seja, quanto pior é o nível de saneamento das comunidades maior é a chance da criança apresentar diarréia. A razão de chances da criança apresentar diarréia quando a água de beber recebe algum tratamento diminuí cerca de 32% em relação a nenhum tratamento da água, considerando todas as outras covariáveis fixas. A razão de chances da criança ter diarréia é 2.01 vezes maior para casas sem nenhum acondicionamento de lixo em relação às casas com um bom acondicionamento de lixo. Crianças de 0 à 11 meses e de 12 à 23 meses possuem um risco 1.65 e 1.59 maior de apresentar diarréia com relação às crianças de 24 à 59 meses, considerando todas as outras covariáveis fixas.

Com relação ao parâmetro de associação, observa-se uma forte associação da diarréia entre todas as possíveis respostas das crianças, exp(1.019)= 2.77, ou seja, crianças com diarréia numa quinzena tem uma chance de apresentar diarréia numa outra quinzena 2.8 vezes maior do que àquelas que não apresentaram diarréia nessa quinzena.

4.5 Considerações finais

O problema de analisar dados categorizados dicotômicos ou politômicos com estruturas complexas é muito frequente nas ciências biomédicas. Seja em estudos longitudinais ou em situações de amostragem por conglomerado em um ou mais estágios existe uma necessidade cada vez mais crescente e mais abrangente de caracterizar melhor as estruturas desses dados.

Na análise de dados de uma amostra por conglomerado deve-se levar em conta a correlação entre as observações no mesmo conglomerado. Além disso, quando o número de observações por conglomerado é desbalanceado, a contribuição devido a variabilidade no tamanho do conglomerado também deve ser considerada. Um modo de incluir ambas considerações é usar razão de médias (Snyder, 1993). Na estimação de parâmetros de interesse

como médias, totais e razões, em geral, são considerados os pesos e o plano amostral utilizado na obtenção da amostra. Menos comum, porém, é a incorporação dos pesos e plano amostral na construção e ajustes de modelos estatísticos.

Nos estudos longitudinais, a dependência de cada resposta categorizada em relação à variáveis explanatórias levando em consideração também uma possível associação entre as respostas, vem estimulando o desenvolvimento de metodologias flexíveis e eficientes bem como a publicação de programas para execução dessas metodologias em pacotes estatísticos.

Três metodologias, regressão de razão de médias com mínimos quadrados ponderados, equações de estimação generalizadas e regressão logística alternada foram apresentadas e discutidas com o propósito de permitir análises que considerem aspectos não triviais dos dados. O objetivo foi gerar um entendimento maior desses assuntos em termos conceituais e técnicos visando colocar ao alcance dos pesquisadores uma ferramenta de análise mais abrangente, aproveitando melhor as estruturas dos dados das pesquisas.

A intenção foi explorar habilidades específicas de cada uma das metodologias procurando assim contribuir com técnicas modernas de análise para avaliações de possíveis fatores de risco associados a questões de saúde que dizem respeito diretamente as crianças avaliadas nas pesquisas.

A motivação foi explorar temas atuais, metodologias recentes na área de estatística, tendo como ilustração uma contribuição real de análise de dados, devido a importância de estudos que procuram avaliar a associação entre indicadores de morbidade e mortalidade principalmente para doenças infecto-contagiosas e a crescente produção de pesquisas nessa área nos países em desenvolvimento.

O fato de agora estarem disponíveis programas computacionais para análise desses tipos de dados mais complexos, faz com que o uso dessas metodologias ganhe cada vez mais espaço tanto no contexto do delineamento quanto no contexto de medidas repetidas.

Sobre os resultados das análises cabe comentar que apesar da proposta não ser a comparação de metodologias, em geral, nas análises feitas os resultados observados apontam alguns dos mesmos indicadores sócio-ambientais como elementos privadores de um melhor estado de saúde infantil. Por exemplo: a ausência de sanitário domiciliar revelou-se um forte

fator de risco associado a ocorrência de episódios severos de diarréia nas metodologias de razão de médias e das EEG para os dados de Serrinha. O suplemento com vitamina A também revelou-se em ambas metodologias fator protetor contra a diarréia severa. Os diagnósticos, mesmo abordados de maneira informal, também contribuíram para uma melhor avaliação das crianças identificadas como influentes no ajuste de modelos. Já nos dados do AISAM, as diferentes condições de saneamento nas comunidades estudadas e a falta de tratamento na água de beber, mostraram-se fortes fatores de risco para a diarréia infantil nas diferentes metodologias empregadas.

A existência de limitações com relação a essas metodologias, por exemplo, no caso da regressão ponderada, a questão de frequências pequenas nas caselas por causa da classificação cruzada, ou nas EEG, a introdução de uma matriz de covariância arbitrária nas equações escores (Lindesey & Lambert, 1998), tem fomentado discussões na literatura estatística gerando oportunidades de mais publicações, o que vem beneficiar os pesquisadores de outras áreas com seus dados cada vez mais complexos.

No que diz respeito a pesquisas futuras, temas importantes encontram-se em aberto, como por exemplo, dados faltantes, enquanto que as técnicas de diagnósticos aqui exploradas de forma elementar, também apontam para novos e continuados trabalhos pois não há até o momento, por exemplo nas EEG, testes específicos para ajuste de modelo. Com relação a regressão logística alternada é necessário estudos de simulação que incorporem uma estrutura de série de tempo para o modelo do log da razão de chances dois a dois, principalmente no que diz respeito a matrix de covariância.

A apresentação e discussão dessas metodologias e as análises apresentadas, naturalmente não esgotam todos os diferentes aspectos possíveis de serem abordados dos pontos de vista teórico e aplicado, como também não são as únicas metodologias para tratar dados complexos. Por exemplo, os modelos lineares hierárquicos têm despontado nos últimos anos como um ferramental capaz de lidar com medidas repetidas e também com vários aspectos do plano amostral.

Espera-se com esse trabalho ter trazido à discussão alguns temas avançados que estão disponíveis na literatura estatística e que suas implementações tenham permitido avanços no

aproveitamento dos aspectos relativos a não trivialidade dos dados, produzindo assim, resultados mais fidedignos às inferências que se desejava realizar.

Referências Bibliográficas

ALMEIDA, N.F., ROUQUAYROL, M.Z. (1992). *Introdução à Epidemiologia Moderna*. 2ª edição. Belo Horizonte: COOPMED, 186 p.

ASSIS, A.M.O. (1996). Suplementação com vitamina A e o crescimento pondo-estatural infantil. Tese de Doutorado, Instituto de Saúde Coletiva da Universidade Federal da Bahia.

BARRETO, M.L., SANTOS, L.M.P., ASSIS, A.M.O., ARAÚJO, M.P.N. FARENZENA, G. G., SANTOS, P.A.G., FIACCONE, R.L. (1994) Effect of vitamin A suplementation on diarrhoea and acute lower respiration infection in young children in Brazil. *Lancet, 344*, 228-231.

BAQUI, A.H., BLACK, R.E., YUNUS, M.D. (1991). Methodological issues in diarrhoeal disease epidemiology: definition of diarrhoeal episodes. *Int. J. Epidemiol., 20*, 1057-1063 .

BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex survey. *Inter. Statist. Rew.*, 51, 279-292.

BRIER, S.S. (1980). Analysis of contingency tables under cluster sampling. *Biometrika, 67*, 591-596.

BRESLOW, N.E., DAY, N.E. (1987). *Statistical Methods in Cancer Research*. Vol II: The Design and Analysis of Cohort Studies. Lyon: IARC.

CAREY, V., ZEGER, S.L., DIGGLE, P. (1993) Modeling multivariate binary data with alternating logistic regressions. *Biometrika, 80*, 517-26.

CARR, G.J., CHI, E.M. (1992). Analysis of variance for repeated measures data: a generalized estimating equations approach. *Statist. Med., 11*, 1033-1040.

CHATTERJEE, S., HADI, A.S. (1986). Influential observations, high leverage points, and outliers in linear regression. *Statist. Sci., 1(3)*, 379-416.

CONNOLLY, M.A. LIANG, K.Y. (1988). Conditional logistic regression models for correlated binary data. *Biometrika, 75*, 501-506.

CORDEIRO, G. (1992). Introdução à teoria de verossimilhança. *In: 10º Simpósio Nacional de Probabilidade e Estatística*, Anais, Rio de janeiro, 174 p.

DAVIES, G.M. (1994). *Applications of sample survey methodology to repeated measures data structures in dentistry*. Ph.D. thesis, Department of Biostatistics University of North Carolina, Chapel Hill.

DAVIS, C. (1991). Semi-parametric and non-parametric methods for the analysis of repeated measurements with applications to clinical trials. *Statist. Med., 10*, 1959-1980.

DAVIS, C.S. (1993). Computer program for regression analysis of repeated measures using generalized estimating equations. *Comput. Methods Progr. Biomed., 40*, 15-31.

DIGGLE, P.J., LIANG, K., ZEGER, S.L. (1994). *Analysis of longitudinal data*. Oxford, Science Publication.

DONNER, A., DONALD, A. (1988). The statistical analysis of multiple binary measurements. *J. Clin. Epidemiol., 41*, 899-905.

DONNER, A. (1989). Statistical methods in opthalmology: An adjusted chi-square approach. *Biometrics, 45*, 605-611.

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

FITZMAURICE, G.M., LAIRD, N.M., ROTNITZKY, A.G. (1993). Regression models for discrete longitudinal response. *Statist. Sci., 8(3)*, 284-309.

FITZMAURICE, G.M., LIPSITZ, S.R. (1995). A model for binary time series data with serial odds ratio patterns. *Appl. Statist., 44*, 51-61.

FREEMAN, D., FREEMAN, J.L. BROCAR. D., KOCH, G.G. (1976). Strategies in the multivariate analysis of data from complex surveys II. An application to the United States National Health Interview Survey. *Inter. Statist. Rev., 44*, 317-330.

FREEMAN, D.H., BROCK, D.B. (1977) The role of covariance matrix estimation in the analysis of complex survey data. *In: Survey Sampling and Measurement*. ed. K. N. Namboodiri, p. 121-40. New York: Academic.

GHANA VAST Study Team. (1993). Vitamin A supplementation in northern Ghana: effects on clinic attendances, hospital admissions, and child mortality. *Lancet. 342*, 7-12.

GODAMBE, V.P. (1960) An optimun property of regular maximum likelihood estimation. *Annals of mathematical, Statist, 31*, 1208-1211.

---. (1991) *Estimating Functions*. Oxford. Oxford University Press. (Oxford Statistical Science series - 7).

GRIZZLE, J.E., STARMER, C.F., KOCH, G.G. (1969). Analysis of categorical data by linear models. *Biometrics, 25*, 459-504.

HANSEN, M.H., HURWITZ, W.N., MADOW, W.G. (1953). Sample Survey Methods and Theory. Vol. II. John Wiley & Sons.331p.

HEAGERTY, P.J., ZEGER, S.L. (1996). Marginal regression models for clustered ordinal meassurements. *J. Am. Statist. Assoc., 91*, 1024-1036.

HEUMANN, C. (1996). Marginal regression modeling of correlated multicategorical response. A likelihood approach. *SFB386 - Discussion Paper 19*, Universität München.

KASTNER, C., ZIEGLER,A., HEUMANN, C. (1997). MAREG and WinMAREG - A tool for marginal regression models. *Comput. Statist. and Data Anal., 24*, 237-241.

KATZ, J., CAREY, V.J., ZEGER, S.L., SOMMER, A. (1993). Estimation of design effects and diarrhea clustering within households and villages. *Am. J. Epidemiol., 138*, 994 -1006.

KOCH, G., FREEMAN, D. H., FREEMAN, J. L. (1975) Strategies in the multivariate analysis of data from complex surveys. *Inter. Statist. Rev., 43*, 59-78.

KOCH, G., LANDIS, R.J., FREEMAN, D.H., FREEMAN, J.L. and LEHNEN, R. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics, 33*, 133-158.

KOCH, G.G., GILLINGS, D.B., STOKES, M.F. (1980). Biostatistical implications of design, sampling and measurement to the analysis of health science data. *Ann. Rev. Pub. Health 1*, 163-225.

KOCH, G.G., IMREY, P.B., SINGER, J.M., ATKINSON, S.S. and STOKES, M.E. (1985) Analysis of categorical data. *In: Colletion Seminaire de Mathematiques Superieures 96*, G. Sabidussi (ed.) Montreal Les Presses de L'Université de Montreal.

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

KOCH, G.G., SINGER, J.M., STOKES, M. (1992). Some aspects of weighted least squares analysis for longitudinal categorical data. *In: Statistical Models for Longitudinal Studies of Health*. Ed. Dwyer, J., Feinleib, M., Lippert, P., Hoffmeister, H., New York, Oxford University Press.

LANDIS, J.R., LEPKOWSKI, J.M., STEPHEN, A.E., STEHOUWER, S. (1982). A statistical methodology for analyzing data from a complex survey: The first national health and nutrition examination survey. *Vital Health Statistic, 2(92)*, 1-52.

LANDIS,J. R., LEPKOWSKI, J.M., DAVIS, C.S., MILLER, M.E. (1987). Cumulative logit models for weight data from complex sample surveys. *Proceeding of the Social Statistics Section of the American Statistical Association*, 165-170.

LANDIS, J.R., MILLER, M.E., DAVIS, C.S., KOCH, G.G. (1988). Some general methods for the analysis of categorical data in longitudinal studies. *Statist. Med., 7*, 109-137.

LAVANGE, L.M., KEYS, L.L., KOCH, G.G., MARGOLIS, P.A. (1994). Application of sample survey methods for modelling ratios to incidence densities. *Statist. Med., 13*, 343-355.

LIANG, K.Y., ZEGER, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*,13-22.

--- (1995). Inference based on estimating functions in the presence of nuisance parameters. *Statist. Sci., 10*, 158-199 (with discussion).

LIANG, K.Y., ZEGER, S.L., QAQISH, B. (1992). Multivariate regression analyses for categorical data. *J. Roy. Statist. Soc., 54*, 3-40. (with discussion).

LINDSEY, J.K., LAMBERT, P. (1998). Marginal models for repeated measurements. *Statits. Med., 17*, 447-469.

LIPSITZ, S.R., LAIRD, N.M., HARRINGTON, D.P. (1991). Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika, 78*, 156-160.

LIPSITZ, S.R., KIM, K., ZHAO, L. (1994). Analysis of repeated categorical data using generalized estimating equations. *Statist. Med., 13*, 1149-1163.

MILLER, M.E., DAVIS, C.S., LANDIS, J.R. (1993). The analysis of longitudinal polytomous data: generalized estimating equations and connections with weighted least squares. *Biometrics, 49*, 1033-1044.

McCULLAGH, P., NELDER, J.A (1989). *Generalized linear models*. London, Chapman and Hall.

MORAES, L.R.S. (1996). *Health impact of drainage and sewerage in poor urban areas in Salvador, Brasil*. Ph.D. thesis, Department of Epidemiology and Population Sciences, London School of Hygiene and Tropical Medicine.

MORRIS, S.S., COUSENS, S.N., LANEDA, C.F., KIRKWOOD, B.R. (1994). Diarrhoea-Defining the Episode. *Inter. J. Epidemiol., 23(3)*, 617-623.

PARK, T. (1993). A comparison of the generalized estimating equation approach with the maximum likelihood approach for repeated measurements. *Statist. Med., 12*, 1723-1732.

PREGIBON, D. (1981). Logistic regression diagnostic. *Ann. Statist. 9*, 705-724.

PREISSER, J., QAQISH, B.F. (1996). Deletion diagnostic for generalized estimating equations. *Biometrika, 83(3)*, 551-562.

PRENTICE, R.L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics, 44*, 1033-1048.

PRENTICE, R.L., ZHAO, P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continous responses. *Biometrics, 47*, 825-839.

RAO, J. N.K., SCOTT, A.J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables. *J. Am. Statist. Assoc., 76*, 220-230.

--- (1984). On chi-squared tets for multi-way contingency tables with cell proportions estimated from survey data. *The Annals of Statist., 12*, 46-60.

ROSNER, B. (1984). Multivariate methods in opthalmology with application to other paired-data situations. *Biometrics, 40*, 1025-1035.

--- (1989). Multivariate methods for clustered binary data with more than one level of nesting. *J. Am. Statist. Assoc., 84*, 373-380.

SINGER, P.M., ANDRADE, D.F. (1986). Análise de dados longitudinais. *In: VII Simpósio Nacional de Probabilidade e Estatística*, Anais. Campinas, 106p.

SMITH, D.M., ROBERTSON, B., DIGGLE, P. (1996). Oswald : Object-oriented software for the analysis of longitudinal data in S. *Technical Report MA 96/192*, Department of Mathematics and Statistics, University of Lancaster, LA1 4YF, United Kingdom.

SNYDER, E.S. (1993). *The analysis of binary data with large, unbalanced and incomplete clusters using ratio mean and weighted regression methods.* Doctoral thesis, Department of Biostatistics University of North Carolina, Chapel Hill.

SOMMER, A, TARWOTJO, J., DJUNAEDI, E. (1986) Impact of vitamin A supplementation on childhood mortality: a randomised controlled community trial. *Lancet, ii*:1169-1173.

STANISH, W.M., GILLINGS, G.G., KOCH, G.G. (1978). An application of multivariate ratio methods for the analysis of a longitudinal clinical trial with missing data. *Biometrics, 34*, 305-317.

STATA (1997). *Stata User's Guide, Release 5*, College Station, Texas: Stata Press.

STRAM, D.O., WEI, L.J., WARE, J.H. (1988). Analysis of repeated ordered categorical outcomes with possibly missing observatons and time-dependent covariates. *J. Am. Statist. Assoc., 83*, 631-637.

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

WEDDERBURN, R.W.M. (1974). Quase-likelihood functions generalized linear models and the Gauss-Newton methods. *Biometrika, 61*, 439-466.

WEI, L.J., STRAM, D.O. (1988). Analysing repeated measurements with possibly missing observations by modelling marginal distributions. *Statist. Med., 7*, 139-148.

WEST, K.P., POLHEREL, R.P., KATZ, S. (1991). Efficacy of vitamin A in reducing child mortality in Nepal. *Lancet, 338*, 67-71.

WILLIAMS, D.A. (1987). Generalized linear model diagnostics using the deviance and single case deletions. *Appl. Statist.,36*, 181-191.

WOODRUFF, R.S A. (1971). Simple method for approximating the variance of a complicated estimate. *J. Am. Statist. Assoc., 66*, 411-414.

ZEGER, S.L., LIANG, K.Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics, 42*, 121-130.

ZEGER, L. (1988). Commentary. *Statist. Med.*, 7, 161-168.

ZEGER, S.L., LIANG, K.Y., ALBERT, P.A. (1988). Methods for longitudinal data: a generalized estimating equation approach. *Biometrics, 44*, 1049-1060.

ZIEGLER, A., KASTNER,C., BLETTNER,M., GROMPING,U. (1996). The generalized estimating equations in the past ten years: an overview and a biomedical application. *SFB 388 Discussion Paper no. 24.*

ZIEGLER, A., ARMINGER, G. (1996). Parameter estimation and regression diagnostics using generalized estimating equations.*In Faulbaum F., Bandilla W. (eds.): "SoftStat'95 - advances in statistical software 5."* Stuttgart: Lucius & Lucius, 229-237.

ZIEGLER, A., BLETTNER, M., KASTNER, C., CHANG-CLAUDE, J. (a publicar). Identifying influential families using regression diagnostics for generalized estimating equations. *Genetic Epidemiology*.

ZHAO, L.P., PRENTICE, R.L. (1990). Correlated binary regression using a quadratic exponencial model. *Biometrika*, 77, 642-648.

APÊNDICE A

Amostragem de Conglomerado em 2 Estágios

1. Notação

Admite-se que N conglomerados tenham sidos selecionados ao acaso com reposição de uma população com Γ conglomerados. Sejam φ_M, a probabilidade de selecionar o conglomerado M, $M = 1,2,\ldots,\Gamma$ e n_M os elementos selecionados do conglomerado M por amostragem aleatória simples com reposição. O j-ésimo elemento na amostra do i-ésimo conglomerado amostrado é representado por y_{ij}, onde $i = 1,2,3,\ldots,N$ e $j = 1,2,3,\ldots,n_i$. Note que $n_i = n_M$ quando o conglomerado M é selecionado. A média amostral para o conglomerado i é dada por

$$\bar{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}.$$

O valor esperado condicional de $\bar{y}_i$ dada a seleção do conglomerado M é a média por conglomerado M, isto é ,

$$E\{\bar{y}_i / i = M\} = E\{\frac{\sum_{j=1}^{n_i} y_{ij}}{n_i} / i = M\}$$

$$= \frac{1}{n_M}\sum_{j=1}^{n_i} E\{y_{ij} / i = M\} = \frac{\sum_{\xi=1}^{\nu_M} \bar{Y}_{M\xi}}{\nu_M} = \bar{Y}_M$$

O valor esperado não condicional de $\bar{y}_i$ é :

$$E\{\bar{y}_i\} = E\{E[\bar{y}_i / i = M]\} = E\{\bar{Y}_M\} = \sum_{M=1}^{\Gamma} \varphi_M \bar{Y}_M .$$

Um outro estimador para a média por conglomerado e para a média geral é dado por

$$\bar{z}_i = \frac{\bar{y}_i \, \nu_i}{\varphi_i \, \nu_+},$$

onde ν_i é o número total de elementos no i-ésimo conglomerado amostrado, φ_i é a probabilidade de selecionar o conglomerado i, e ν_+ é o número total de elementos na população. Ver Sukatme para uma melhor compreensão do estimador acima.

O valor esperado condicional de $\bar{z}_i$, dada a seleção do conglomerado M é

$$E\left\{\bar{z}_i \;/\; i = M\right\} = E\left\{\frac{\bar{y}_i \, \nu_i}{\varphi_i \, \nu_+} \;/\; i = M\right\} = \frac{\nu_M}{\varphi_M \, \nu_+} E\left\{\bar{y}_i \;/\; i = M\right\}$$

$$= \frac{\nu_M}{\varphi_M \, \nu_+} \bar{Y}_M .$$

A esperança de $\bar{z}_i$ é

$$E\left\{\bar{z}_i\right\} = E\left[E\left(\bar{z}_i \;/\; i = M\right)\right] = E\left\{\frac{Y_M}{\varphi_M \, \nu_+}\right\} = \bar{Y} .$$

Então $\bar{z}_i = \frac{\bar{y}_i \, \nu_i}{\varphi_i \, \nu_+}$ é um estimador viciado para a média por conglomerado $\bar{Y}_M$ e também um estimador não-viciado para a média geral $\bar{Y}$. Quando $\frac{\nu_i}{\varphi_i \, \nu_+} = 1$, $\bar{y}_i = \bar{z}_i$ são ambos estimadores não-viciados para a média por conglomerado e média geral.

Assim, a estatística $\bar{z}_i = \frac{\bar{y}_i \, \nu_i}{\varphi_i \, \nu_+}$ pode ser escrita como uma média de elementos amostrais ponderados, ou seja, $\bar{z}_i = w_i \, \bar{y}_i$, onde $w_i = \frac{\nu_i}{\varphi_i \, \nu_+}$.

Os $\bar{z}_i$ são independentes e identicamente distribuídos se os conglomerados são selecionados com reposição e à seleção dentro dos conglomerados é independente uma das outras. A variância para a média por conglomerado, isto é, a variância de $\bar{z}_i$ condicional a i=M é

$$V(\bar{z}_i \;/\; i = M) = V\left[w_i \bar{y}_i \;/\; i = M\right] = w_M^2 \, V\left(\bar{y}_i \;/\; i = M\right)$$

$$= w_M^2 \, (1 - f_M) \frac{1}{n_M} \sigma_M$$

onde $\sigma_M = \frac{1}{n_M - 1} \sum_{\xi=1}^{\nu_M} (Y_{M\xi} - \bar{Y}_M)^2$, $f_M = \frac{n_M}{\nu_M}$ e $(1 - f_M)$ é a correção para população finita (Kish, 1965).

Agora, a variância para a média global, ou seja, a variância de $\bar{z}_i$ é

$$V(\bar{z}_i) = \sigma_b + \sigma_a = \sum_{M=1}^{\Gamma} \left\{ (w_M \, \bar{Y}_M - \bar{Y})^2 \varphi_M + w_M^2 \, \frac{1}{n_M} (1 - f_M) \, \varphi_M \sigma_M \right\}.$$

Portanto, a média simples $\bar{z}_i$ é

$\bar{z} = \frac{1}{N} \sum_{i=1}^{N} \bar{z}_i$, cujo valor esperado é $E(\bar{z}_i) = \bar{Y}$ e a variância é $V(\bar{z}) = \frac{1}{N} (\sigma_b + \sigma_w)$. Um estimador não-viciado para $V(\bar{z})$ é dado por :

$$\hat{v}(\bar{z}_i) = \frac{1}{N(N-1)} \sum_{i=1}^{N} (\bar{z}_i - \bar{z})^2 .$$

APÊNDICE B

Exemplos de matrizes de variância-covariância e de correlação

1. Mínimos quadrados ponderados

- Valores das variâncias de R (x 10^{-5}) para o modelo da pág 101 - Estudo de Serrinha

0.0014641	0	0	0	0	0	0	0
0	0.005551	0	0	0	0	0	0
0	0	0.0049774	0	0	0	0	0
0	0	0	0.0136604	0	0	0	0
0	0	0	0	0.0016848	0	0	0
0	0	0	0	0	0.0144324	0	0
0	0	0	0	0	0	0.0126451	0
0	0	0	0	0	0	0	0.0208483

- Valores das variâncias-covariâncias de R (x 10^{-6}) para o modelo da pág 119 - Estudo AISAM

5.8550	4.0532	6.9349	7.3205	0	0	0	0	0	0	0	0
4.0532	5.0001	3.8086	4.5045	0	0	0	0	0	0	0	0
6.9349	3.8086	8.6626	8.9254	0	0	0	0	0	0	0	0
7.3205	4.5045	8.9254	9.2975	0	0	0	0	0	0	0	0
0	0	0	0	0.5532	0.0430	0.6289	0.9448	0	0	0	0
0	0	0	0	0.0430	0.6887	-0.3811	0.6531	0	0	0	0
0	0	0	0	0.6289	-0.3811	0.9848	0.7104	0	0	0	0
0	0	0	0	0.9448	0.6531	0.7104	2.1037	0	0	0	0
0	0	0	0	0	0	0	0	1.1499	0.5644	0.4276	-0.2736
0	0	0	0	0	0	0	0	0.5644	0.2773	0.2024	-0.1459
0	0	0	0	0	0	0	0	0.4276	0.2024	0.3837	0.2468
0	0	0	0	0	0	0	0	-0.2736	-0.1459	0.2468	0.6056

2. Equações de estimação generalizadas

- Valores das correlações para o modelo da pág 109

1.000	0.139	0.139	matriz de correlação permutável
0.139	1.000	0.139	
0.139	0.139	1.000	