Universidade de Brasília
Instituto de Exatas
Departamento de Estatística

Regressão Logística Multinível com Estimação de Equações Generalizadas (GEE): Uma Aplicação em Dados Odontológicos

Ludmilla Lorrany Mattos Silva

Orientador: Professor Dr. Eduardo Freitas da Silva

Brasília

2017

Ludmilla Lorrany Mattos Silva

Regressão Logística Multinível com Estimação de Equações Generalizadas (GEE): Uma Aplicação em Dados Odontológicos

Relatório final apresentado à disciplina Trabalho de Conclusão de Curso de graduação em Estatística apresentado ao Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Orientador: Professor Dr. Eduardo Freitas da Silva

Universidade de Brasília – UnB

Instituto de Exatas

Departamento de Estatística

Trabalho de Conclusão de Curso de Graduação

Brasília

2017

Aos meus pais.

Resumo

Quando várias medidas são observadas repetidamente em uma mesma unidade observacional, podemos ter o que se chama de estrutura multinível. Nesta configuração de dados, também conhecida como hierárquica, os indivíduos são independentes ente si, porém existe uma correlação intra-indivíduos. A estimação de equações generalizadas (*Generalized Estimating Equations - GEE*) é uma técnica capaz de estimar a matriz de correlação intra-indivíduos e assim, produz estimativas não enviesadas para os parâmetros dos modelos de regressão generalizados (*Generalized Regression Models - GLM*).

Neste trabalho foi realizado um estudo com crianças diagnosticadas com a doença Hipomineralização Molar Incisivo. Foi investigada a associação entre a prevalência de dentes quebrados e a presença de opacidades de coloração escura no esmalte. A análise foi realizada utilizando um modelo logístico multinível com estimação de equações generalizadas. Os resultados obtidos revelam que no nível intra-divíduo, dentes de coloração escura são 7,74 vezes mais suscetíveis a sofrerem fratura, confirmando as expectativas dos especialistas.

Palavras-chave: Modelos Lineares Generalizados. Modelos de regressão multinível. Estimação de Equações Generalizadas. Hipomineralização Molar Incisivo

Sumário

Introdução **13**

1 REFERENCIAL TEÓRICO **15**
1.1 Modelos Lineares Generalizados **15**
1.1.1 Regressão Logística 15
1.2 Estimação de equações generalizadas (GEEs) **17**
1.2.1 Estimação 17
1.2.2 Matriz de correlação de trabalho 20

2 APLICAÇÃO **23**
2.1 Introdução **23**
2.2 Metodologia **24**
2.3 Análise descritiva **26**
2.3.1 Análise bivariada no nível do dente 26
2.3.2 Análise bivariada no nível do indivíduo 27
2.4 Modelagem **29**

3 CONCLUSÃO **31**

REFERÊNCIAS **33**

Lista de ilustrações

Figura 1 – Relação hierárquica 13
Figura 2 – Opacidade de coloração branca 23

Lista de tabelas

Tabela 1 – Matrizes de correlação de trabalho 20
Tabela 2 – Frequências no nível do dente 26
Tabela 3 – Molelos log-lineares 27
Tabela 4 – Frequências no nível da criança 27
Tabela 5 – Testes qui-quadrado 28
Tabela 6 – Estimativas dos parâmetros 29

Introdução

Em estudos na área da pesquisa odontológica, é usual lidar com complexas estruturas de dados. Complicações como medidas correlacionadas devido aninhamento ou cluster, violam as suposições de independência dos modelos tradicionais de regressão e análise de variância.

As variáveis que representam os fatores de risco, geralmente são observadas em múltiplos dentes no mesmo paciente. Sendo assim, necessariamente as unidades observacionais são correlacionadas, pois estão agrupadas (aninhadas) em cada indivíduo. Isso gera uma estrutura hierárquica em dois níveis: indivíduo e dente.

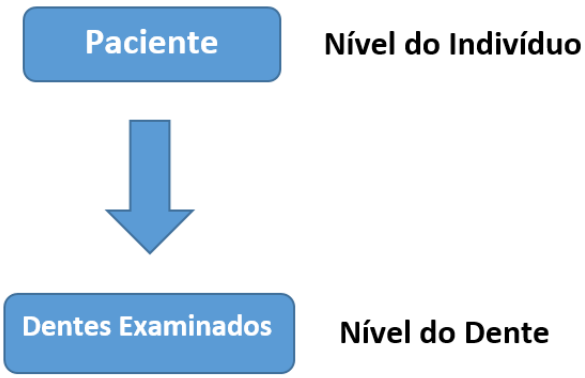

Figura 1 – Relação hierárquica

Trata-se de uma correlação que ocorre de forma natural. Afinal, cada grupo de dentes possui características semelhantes, inerentes ao indivíduo a quem pertencem. Por exemplo, são submetidos à mesma dieta, frequência de escovação, exposição à nicotina ou não. Cada nível dessa hierarquia pode apresentar informações estatísticas diferentes sobre os dados. Faz-se assim necessária uma decomposição das variáveis capaz de captar as relações entre-indivíduos e intra-indivíduos (*between-subject e within-subject).* Estudos que ignoram esta particularidade dos dados tendem a apresentar resultados enviesados, principalmente quando as informações em cada nível divergem.

Para a avaliação de dados com esta estrutura de aninhamento, são utilizados os modelos hierárquicos, comumente chamados em várias áreas por modelos multinível. São alternativas flexíveis, que permitem a análise dos efeitos de tratamentos em todos os níveis da hierarquia possibilitando a comparação entre e dentre os indivíduos.

Existem duas abordagens principais para modelar dados hierárquicos: os modelos mistos, que combinam fatores fixos e aleatórios, e também a abordagem marginal dos

modelos com estimação de equações generalizadas *(Generalized Estimating Equations - GEE)*, que apresenta resultados com precisão satisfatória e alta flexibilidade para modelar a correlação presente nos dados. Neste caso, os efeitos modelados são fixos, calculados de forma que leva em consideração uma estrutura de correlação pré-determinada através da chamada Matriz Auxiliar (*Working Matrix*). Todo o procedimento será descrito em maiores detalhes na parte metodológica deste trabalho. A abordagem GEE será utilizada no estudo de caso, aplicado por um prisma analítico de suas particularidades, vantagens e desvantagens.

Ao longo deste trabalho, será realizada uma revisão objetiva dos Modelos Lineares Generalizados, especificamente os casos que permitem a análise de estruturas multinível. Em seguida, será feita uma aplicação em um conjunto de dados odontológicos de crianças com a doença dentária Hipomineralização Molar Incisivo. Trata-se de um estudo de natureza observacional no qual a variável resposta é binária e portanto, será utilizada uma regressão logística para estimar a razão de chances (*Odds Ratio - OR*) para a quebra do dente quando existem opacidades no esmalte.

1 Referencial Teórico

Nesta seção, é revisado o arcabouço teórico das técnicas de modelos lineares generalizados que permitem o estudo de dados que apresentam estrutura de correlação.

1.1 Modelos Lineares Generalizados

A classe dos modelos lineares generalizados (GLMs) é uma extensão dos modelos lineares tradicionais. Os GLMs permitem a modelagem de uma variável resposta que não necessariamente segue uma distribuição normal. A resposta média da população $\mu = E(y)$ é relacionada a um preditor linear através de uma função de ligação não linear $g(\mu)$ citehosmer. Desse modo, a dependente y pode seguir qualquer distribuição de probabilidade pertencente à família exponencial. Esta é uma vantagem essencial em muitas áreas que estudam variáveis dicotômicas ou outra resposta não linear.

De forma mais detalhada, os GLMs possuem três componentes:

- Componente Aleatória: é a especificação da distribuição de probabilidade da variável resposta Y. As observações $(y_1, ..., y_n)$ geralmente são consideradas independentes.
- Componente Fixa: é a combinação linear das variáveis explicativas e os parâmetros desconhecidos, denominada preditor linear, é dado por:

$$\alpha + \beta_1 x_1 + ... + \beta_k x_k. \tag{1.1}$$

- Função de Ligação: função monotônica diferenciável $g(.)$ que relaciona $\mu = E(Y)$ com o preditor linear, isto é, conecta os componentes aleatória e fixa:

$$g(\mu) = \alpha + \beta_1 x_1 + ... + \beta_k x_k. \tag{1.2}$$

  A função de ligação utilizada depende da distribuição de probabilidade da resposta. No caso do estudo abordado neste trabalho, a variável em questão é dicotomizada e segue uma distribuição binomial. A função de ligação (*link function*) utilizada nesta ocasião é a função *l*ogito.

1.1.1 Regressão Logística

A regressão logística é o caso particular dos modelos lineares generalizados em que a variável resposta é categórica com distribuição binomial. Nesse caso, a função de ligação utilizada é a *l*ogito:

Seja um modelo com variável resposta Y com duas categorias e uma variável explicativa X. Então $\pi(x)$ será a probabilidade de sucesso de Y quando $X = x$:

$$\pi(x) = \frac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)}. \tag{1.3}$$

Aplicando a transformação Logito:

$$logito(\pi(x)) = \ln \frac{\pi(x)}{1 - \pi(x)} = \beta_0 + \beta_1 x. \tag{1.4}$$

Ou seja, o logaritmo natural da chance de sucesso de Y assume forma linear.

Seja a razão de chances (*Odds Ratio*), tem-se que o efeito multiplicativo na chance de sucesso para cada acréscimo de uma unidade em x é dada por:

$$OR = \frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}} = \frac{\frac{\frac{e^{\beta_0+\beta 1}}{1+e^{\beta_0+\beta 1}}}{\frac{1}{1+e^{\beta_0+\beta_1}}}}{\frac{\frac{e^{\beta_0}}{1+e^{\beta_0}}}{\frac{1}{1+e^{\beta_0}}}} = e^{\beta_1}. \tag{1.5}$$

1.2 Estimação de equações generalizadas (GEEs)

Uma medida de interesse binária y_{ij} é observada repetidamente: são obtidas $j = 1, ..., n_i$ medidas para o i-ésimo indivíduo $i = 1, ..., k$. Este é um cenário comum no universo da pesquisa. Na odontologia, a necessidade de observar os vários dentes de um indivíduo é o que resulta nesta estrutura de *cluster*.

Em 1986, LIANG e ZEGER (1986) propuseram uma classe de equações de estimação que geram estimativas consistentes para os parâmetros de regressão sob brandas suposições de dependência. As equações estimação generalizadas (GEE), resultam de uma extensão dos modelos lineares generalizados para dados longitudinais.

No GEE, dados correlacionados são modelados utilizando o mesmo preditor linear e função de ligação que seriam utilizados no caso independente. Entretanto, a estrutura de covariância das medidas repetidas também é modelada. As associações intra-indivíduos, são consideradas através das matrizes auxiliares de correlação intra-indivíduo, que incorporam a estrutura de dependência diretamente na estimação dos betas.

O método GEE ajusta um modelo marginal (*population-averaged model*), ou seja, um modelo de médias com efeitos fixos. Isso significa que, assim como no modelo de regressão linear generalizado, a interpretação dos parâmetros é relativa à média da população e não a um indivíduo específico. O valor esperado marginal da resposta, $E(Y_{it}|x_{it}) = \mu_{it}$, é relacionado ao preditor linear através da função de ligação $g(\mu_{it}) = x'_{it}\beta$, onde $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$' é o vetor p x 1 dos parâmetros regressivos.

1.2.1 Estimação

Afim de entender o desenvolvimento que leva às equações de estimação generalizadas, vamos primeiramente definir o caso em que as observações repetidas para um indivíduo são independentes.

Definindo a anotação utilizada, considere o vetor de respostas observadas no i-ésimo indivíduo $Y_i = [y_{i1}, ..., y_{in_i}]'$ e seja $X_{ij} = [x_{ij1}, ..., x_{ijp}]'$ o vetor de p co-variáveis para cada i-ésimo indivíduo $i = 1, ..., k$.

Assume-se que a densidade marginal de y_{it} como sendo a família exponencial com parâmetro de dispersão ϕ:

$$F(y_{it}) = exp[\{y_{it}\theta_{it} - a(\theta_{it}) + b(y_{it})\}\phi], \tag{1.6}$$

sendo que $\theta_{it} = h(\eta_{it})$ e $\eta_{it} = x_{it}\beta$. Assim, podemos definir os primeiros dois momentos de

y_{it} como sendo:

$$E(y_{it}) = a'(\theta_{it}), \tag{1.7}$$

$$Var(y_{it}) = \frac{a''(\theta_{it})}{\phi}. \tag{1.8}$$

A seguir, temos a equação escore, que é a derivada da função de verossimilhança. O estimador $\hat{\beta}$ de β é obtido pela solução do escore.

$$U_I(\beta) = \sum_{i}^{K} X_i' \Delta_i S_i = 0, \tag{1.9}$$

na qual $\Delta_i = diag(d\theta_{it}/d\eta_{it})$ é uma matriz nxn e $S_i = Y_i - a_i'(\theta)$.

A generalização da estimação para o caso em que as respostas para um mesmo indivíduo são correlacionadas, consiste em adicionar uma matriz auxiliar que represente a correlação existente nos dados. A estimativa de β permanece consistente e estimativas consistentes de variância podem ser obtidas supondo que a matriz de correlação estimada converge para a matriz auxiliar.

Seja a estrutura de correlação intra-indivíduos $R_i(\alpha)$ uma matriz n x n de correlação para medidas repetidas ou em estrutura de *cluster*, sendo α o vetor de parâmetros desconhecidos que especifica a matriz.

Então a matriz de covariância de Y_i é dada como:

$$V_i = \phi A_1^{\frac{1}{2}} \hat{R}(\alpha) A_1^{\frac{1}{2}}, \tag{1.10}$$

na qual, A_i é uma matriz com $v(\mu_{ij})$ sendo o j-ésimo elemento diagonal.

Se $R_i(\alpha)$ é de fato a verdadeira matriz de correlação de Y_i, então V_i é a verdadeira matriz de covariância $cov(Y_i)$. Sendo assim, equação de estimação generalizada para estimar o vetor de parâmetros $\hat{\beta}$ é dada por:

$$S(\beta) = \sum_{i=1}^{K} D_i' V_i^{-1} (Y_i - \mu_i(\beta)) = 0, \tag{1.11}$$

na qual:

- V_i é a matriz de covariância de Y_i
- $D_i = A_i \Delta_i X_i = \frac{d\mu_i}{d\beta}$

- A matriz p x n_i das derivadas parciais com relação aos parâmetros de regressão para o i-ésimo indivíduo é dada por:

$$D_i' = \frac{d\mu_i'}{d\beta} = \begin{bmatrix} \frac{x_{i11}}{g`(\mu_{i1})} & \cdots & \frac{x_{in_i1}}{g`(\mu_{in_i})} \\ \vdots & & \vdots \\ \frac{x_{i1p}}{g`(\mu_{i1})} & \cdots & \frac{x_{in_ip}}{g`(\mu_{in_i})} \end{bmatrix}$$

sendo $g(\mu_{ij}) = x_{ij}'\beta$ e g é a função de ligação.

Temos que $U_i(\beta, \alpha) = D_i^T V_i^{-1}$ é muito similar à abordagem de quasi-verossimilhança, exceto pelo fato de que V_i não é uma função de β apenas, mas de α também. A Equação (1.11) pode ser escrita em termos de β somente, substituindo α por um estimador consistente quando β e ϕ são conhecidos: $\hat{\alpha}(Y, \beta, \phi)$ e ϕ também será estimado por um estimador $\hat{\phi}(Y, \beta)$. Consequentemente, (2.12) terá a forma:

$$\sum_{i=1}^{k} U_i[\beta, \hat{\alpha}\{\beta, \hat{\phi}(\beta)\}] = 0, \tag{1.12}$$

e $\hat{\beta}$ é definido como sendo a solução da Equação 2.13, que é um sistema não linear de β e dos parâmetros $R(\alpha)$ e ϕ. Como o problema não apresenta forma-fechada, a solução é dada pelo procedimento iterativo de estimação

O algorítimo utilizado para ajustar iterativamente o modelo é descrito a seguir:

Algoritmo 1: GEE

início

Determina a estimativa inicial de β assumindo independência.

para *todo $i \in k$ e $j \in n$, até a convergência* **faça**

Estima a matriz de trabalho R com base no β atual e a estrutura R_0

Calcula a estimativa da covariância

$V_i = \phi A_1^{\frac{1}{2}} W_1^{-\frac{1}{2}} \hat{R}(\alpha) W_1^{-\frac{1}{2}} A_1^{\frac{1}{2}}$

Atualiza β

$\beta_{r+1} = \beta_r - \left[\sum_{i=1}^{K} \frac{\delta\mu_i}{\delta\beta}' V_i^{-1} \frac{\delta\mu_i}{\delta\beta} \right]^{-1} \left[\sum_{i=1}^{K} \frac{\delta\mu_i}{\delta\beta}' V_i^{-1} (Y_i - \mu_i) \right]$

fim

fim

1.2.2 Matriz de correlação de trabalho

A matriz de correlação é normalmente desconhecida e precisa ser estimada. Sua estimação é feita pelo processo de ajustamento iterativo usando o valor atual do vetor β para calcular os parâmetros de correlação α com base nos resíduos de Pearson definidos por:

$$\epsilon_{ij} = \frac{y_{ij} - \mu_{if}}{\sqrt{v(\mu_{ij})/w_{ij}}}. \tag{1.13}$$

O estimador utilizado depende da suposição que se faz sobre a estrutura de correlação das observações. Diversas estruturas de matrizes de correlação de trabalho R_0 são possíveis. A Tabela a seguir detalha algumas estruturas comumente utilizadas e seus estimadores.

Tabela 1 – Matrizes de correlação de trabalho

	Estrutura	Estimador
Fixo	$Corr(Y_{ij}, Y_{ik}) = r_{jk}$ onde r_{jk} é constante	Nesse caso, a matriz de correlação não é estimada
Independente	$Corr(Y_{ij}, Y_{ik}) = \begin{cases} 1, & j = k \\ 0, & j \neq k \end{cases}$	Nesse caso, a matriz de correlação não é estimada
M-dependente	$Corr(Y_{ij}, Y_{i+t}) = \begin{cases} 1, & t = 0 \\ \alpha_t, & t = 1, \ldots, m \\ 0, & t = m \end{cases}$	$\hat{\alpha}_t = \frac{1}{(K_t - p)\phi} \sum_{i=1}^{k} \sum_{j \leq n_i} \epsilon_{ij}\epsilon_{i,j+t}$ $K_t = \sum_{i=1}^{k}(n_i - t)$
Permutável	$Corr(Y_{ij}, Y_{ik}) = \begin{cases} 1, & j = k \\ \alpha, & j \neq k \end{cases}$	$\hat{\alpha}_t = \frac{1}{(N* - p)\phi} \sum_{i=1}^{k} \sum_{j<k} \epsilon_{ij}\epsilon_{i,k}$ $N* = 0.5 \sum_{i=1}^{k} n_i(n_i - 1)$
Não estruturada	$Corr(Y_{ij}, Y_{ik}) = \begin{cases} 1, & j = k \\ \alpha_{jk}, & j \neq k \end{cases}$	$\hat{\alpha}_t = \frac{1}{(K - p)\phi} \sum_{i=1}^{k} \epsilon_{ij}\epsilon_{i,k}$
Autoregressiva AR(1)	$Corr(Y_{ij}, Y_{i+t}) = \alpha^t$ para $t = 0, 1, 2, \ldots, n_i - j$	$\hat{\alpha}_t = \frac{1}{(k_1 - p)\phi} \sum_{i=1}^{k} \sum_{j \leq n_i - 1} \epsilon_{ij}\epsilon_{i,j+1}$ $K_1 = \sum_{i=1}^{k}(n_i - 1)$

O método GEE é capaz de proporcionar excelentes resultados com estimativas consistentes para os parâmetros de regressão. Entretanto, o desafio da técnica reside na escolha da matriz auxiliar, afinal não é possível saber de antemão qual estrutura é a correta. A escolha apropriada da matriz é de suma importância, pois pode resultar em considerável redução ou eliminação do viés de estimação. A seguir, estão elencadas as características das principais matrizes e os casos em que geralmente são utilizadas.

- A estrutura **Independente** é a matriz identidade, assim nenhum coeficiente de correlação é estimado. É usada em ocasiões em que é razoável supor que as múltiplas

medidas na mesma unidade observacional são independentes entre si.

- A matriz **Permutável** assume que qualquer par de medidas em um indivíduo possui a mesma correlação. Adequado para problemas de *cluster* e aninhamento, como crianças de uma mesma escola ou dentes de um indivíduo. É uma opção interessante para pequenas amostras por ser uma estrutura parcimoniosa com apenas um parâmetro α à ser estimado.

- Em problemas de medidas repetidas, coletadas ao longo do tempo, a correlação intra indivíduo pode ser caracterizada como **Autoregressiva**, pois tende a diminuir com o passar do tempo. Em alguns estudos longitudinais a redução no nível de correlação pode ser abrupto, assim a matriz **AR-1**, que é uma autoregressiva de ordem um, se mostra como uma alternativa.

- A correlação **Não Estruturada** não possui um padrão, é a opção mais flexível, porém seu número de parâmetros cresce rapidamente e a estimação pode se tornar instável. Para amostras pequenas e desbalanceadas é desejável estruturas mais parcimoniosas.

2 Aplicação

Nesta seção, é apresentada uma típica aplicação do modelo logístico para dados hierárquicos, utilizando estimação de equações generalizadas. Consiste em um estudo de caráter observacional com interesse em estimar razões de chances da quebra do dente dado um fator de risco, em portadores da doença Hipomineralização Molar Incisivo (HMI).

2.1 Introdução

A Hipomineralização Molar Incisivo (HMI) é um defeito do esmalte que tem origem na fase de desenvolvimento do dente e que atinge os primeiros molares e incisivos permanentes. O MHI é uma patologia que causa porosidade do esmalte, deixando-o com aspecto de giz. Como consequência, o dente fica hipersensibilizado, mais suscetível à careação e a impactos mecânicos, podendo quebrar-se até mesmo com a própria mastigação. A doença pode ser identificada visualmente pela presença de opacidades em pelo menos um dos primeiros molares. As alterações na translucidez do esmalte podem variar de uma coloração esbranquiçada até castanho-amarelada ou marrom, que tende a representar maior severidade do problema. Na Figura a seguir, observamos um exemplo de MHI.

Figura 2 – Opacidade de coloração branca

O fator de risco analisado neste caso é a coloração do esmalte. Fatos estilizados sugerem que a tonalidade mais escurecida do dente está associada a maior chance de quebra em relação aos dentes que possuem cor mais esbranquiçada. O objetivo deste estudo será testar a hipótese de independência entre a frequência de dentes quebrados e as variáveis que indicam a tonalidade da cor, tipo de dente (molar ou incisivo) e sua localização (superior ou inferior).

Os dados utilizados neste estudo são uma amostra de 170 estudantes do Distrito Federal, todos portadores da doença HMI, foi extraída de um estudo de prevalência de cárie e MIH realizado em 2013 com crianças de 7 a 12 anos de escolas públicas do Paranoá/DF, realizado para uma pesquisa de doutorado em odontologia.

2.2 Metodologia

Inicialmente, é conduzida uma análise descritiva para melhor caracterizar as observações. Para conhecer o nível de associação entre as variáveis, testes bivariados são realizados em ambos os níveis da hierarquia. A análise no nível do indivíduo é feita através de testes qui-quadrado. Ao nível do dente, por sua vez, a análise exige que a correlação intra-indivíduo seja considerada. Para isso, será utilizado o modelo log-linear (CARVALHO et al., 2011).

As observações do banco de dados são da forma $\{Y_{ij}, X_{ij}\}$, sendo $j = 1, 2...J$ o j-ésimo dente para o i-ésimo indivíduo $i = 1, 2...I$. A variável resposta Y_{ij} é binária, assim para cada observação ij assume-se $Y_{ij} = 1$ ou $Y_{ij} = 0$ e X_{ij} é a co-variável ou fator de risco observado no nível do dente.

Os efeitos entre e intra-indivíduos podem ser obtidos separadamente através do modelo logístico multinível. Para isso, cada covariável observada no nível do dente (x_{ij}) é decomposta em duas componentes. Assim é possível distinguir a informação contida no nível da criança como sendo a média da covariável para os indivíduos ($\bar{x}_i$). E a informação no nível do dente será ($x_{ij} - \bar{x}_i$) (HOLT; SCOTT, 1982).

Esse tipo de modelagem se faz necessária uma vez que os dados apresentam estrutura de agrupamento por indivíduo, ou seja, observações pertencentes a uma mesma criança, tendem a apresentar características semelhantes. A estimação por equações generalizadas (GEE) será utilizada para que o padrão de correlação contido nos dados possa ser considerado no modelo.

- Na implementação da GEE, o modelo de regressão para a média marginal $\mu_{ij} = E[y_{ij}|x_{ij}, \bar{x}_i]$ tem a seguinte forma:

$$g(\mu_{ij}) = \alpha + \beta_{entre}\bar{x}_i + \beta_{intra}(x_{ij} - \bar{x}_i), \tag{2.1}$$

  na qual g é a função de ligação logito.

Note que os coeficientes (α, β_{entre}, β_{intra}) representam o efeito marginal do fator de risco na média da população. O coeficiente de $\bar{x}_i$ mede o efeito da covariável quando a média para os indivíduos difere em uma unidade. Já o coeficiente de ($x_{ij} - \bar{x}_i$) reflete o

efeito baseado na comparação dentro do *cluster*, quando a diferença entre os dentes é de uma unidade (MANCL; LEROUX; DEROUENL, 2000).

A escolha da matriz auxiliar de correlação intra indivíduos que melhor se ajusta aos dados envolve comparar critérios de informação para modelos com diferentes matrizes. Também é importante a ponderação racional de qual alternativa se adéqua melhor dada a natureza do problema de forma parcimoniosa. Ou seja, a matriz auxiliar deve ser aquela que não exija muitos parâmetros a serem estimados e que também não apresente uma estrutura muito simples. A matriz Permutável satisfaz a essas condições e portanto será a alternativa adotada neste trabalho.

Foi utilizado o software SAS (release 9.4) para manipulação do banco de dados, bem como para a aplicação dos modelos descritos utilizando o procedimento PROC GENMOD.

2.3 Análise descritiva

Nesta seção apresentamos um panorama geral de como se distribuem as variáveis e como elas se comportam. Foi realizada uma análise descritiva uni e bivariada em ambos os níveis da hierarquia.

A base de dados em questão, possui dados de 170 crianças, todas acometidas pela doença HMI. Para cada uma delas foram registradas informações de, em média, 3 dentes nos quais foram identificadas alterações na opacidade. Ao todo, 437 dentes foram analisados.

Além do sexo da criança, foram obtidas medidas categóricas no nível do dente: a variável Cor indica se a opacidade presente no dente possui coloração branca ou não branca; Posição classifica o dente como superior ou inferior, enquanto Tipo identifica molares e incisivos.

A Tabela 2 apresenta as proporções e porcentagens em que as variáveis ocorrem.

Tabela 2 – Frequências no nível do dente

		Frequência	Porcentagem
Quebra	Não Quebra	397	90,85
	Quebra	40	9,15
Tipo	Incisivo	137	31,35
	Molar	300	68,65
Posição	Inferior	182	41,65
	Superior	255	58,35
Cor	Branco	302	69,11
	Não Branco	135	30,89

Observa-se que apenas 40 dos 437 dentes analisados sofreram quebra (9,15%), trata-se assim de um evento pouco observado. Quanto a coloração, existe um balanceamento maior entre brancos (69%) e não brancos (30,89%). As variáveis de controle Tipo e Posição também sem mostram bem distribuídas. Foram observados 255 (58%) dentes superiores e 182 (41%) inferiores. Quanto ao tipo, 68% são molares e 31% incisivos.

2.3.1 Análise bivariada no nível do dente

A análise bivariada permite verificar de forma exploratória quais variáveis se mostram mais importantes para o modelo.

Nesta etapa queremos investigar preliminarmente a relação entre as variáveis independentes e resposta. No nível do dente, a correlação intra-indivíduo deve ser considerada e para isso, é utilizado o modelo log-linear. Utilizando o modelo linear generalizado com distribuição de Poisson e variância robusta, é possível identificar a repetição do indivíduo

e a correlação entre as observações pertencentes a uma única criança. Desse modo, determinamos o quanto a contagem de dentes quebrados depende dos níveis de cada variável independente.

Tabela 3 – Molelos log-lineares

		Porcentagem de quebrados	Razão de Chances	P-valor
Cor	Não Branco	19,26	4,26	0,0001
	Branco	4,64		
Tipo	Molar	11,0	2.44	0,0252
	Incisivo	5,11		
Posição	Superior	10,59	1,49	0,3322
	Inferior	7,14		

A análise bivariada no nível do dente revela que a prevalência de dentes quebrados é associada à coloração (p=0,0001). O dente que apresenta opacidade não branca é 2,94 vezes mais suscetível a sofrer quebra do que os dentes brancos. Segundo o teste, as variáveis Tipo e Posição não apresentam dependência significante ao nível de 0,005%.

Embora as variáveis controle Tipo e Posição não tenham apresentado significância quando analisadas no nível do dente, uma análise mais profunda, que inclua a decomposição entre e intra-indivíduos, pode vir a revelar resultados diferentes e portanto, foram incluídas no modelo logístico.

2.3.2 Análise bivariada no nível do indivíduo

Enquanto no nível intra-indivíduos, as características eram observadas no nível do dente, no caso do nível da criança, observamos se há a ocorrência de pelo menos um dente quebrado e se há pelo menos um dente com opacidade não branca. Seguindo esse critério, a Tabela 4 apresenta as frequências de crianças que apresentaram dentes danificados, com opacidade escura e também o sexo.

Tabela 4 – Frequências no nível da criança

	Cor	Frequência	Porcentagem
Quebra	Quebrados	48	28,24
	Não quebrados	122	71,76
Cor	Branco	68	40
	Não Branco	102	60
Sexo	Masculino	86	50,59
	Feminino	84	49,41

Observa-se que 28,24% (48) das crianças avaliadas apresentaram dentes quebrados, contra 71,76% (122) que não possuíam nenhum dente danificado. A maior parte das

crianças possuem pelo menos um dente não branco (60%). A amostra é bem distribuída entre meninos e meninas, 50,59% das crianças são do sexo masculino.

No nível da criança, foram realizados testes qui-quadrado de associação entre a ocorrência de dentes quebrados e a coloração. A Tabela 5 apresenta os resultados.

Tabela 5 – Testes qui-quadrado

		Porcentagem de quebrados	χ^2	P-valor
Cor	Não Branco	36,27	8,1333	0,0043
	Branco	16,18		
Sexo	Masculino	27,91	0,0093	0,9233
	Feminino	28,57		

O teste qui-quadrado de associação entre a Quebra e Cor foi significante ao nível de 0,005. Indicando que crianças que apresentaram pelos menos um dente não branco, possui chance 2,94 vezes maior de apresentar pelo menos um dente quebrado. Já variável Sexo, conforme esperado, não se mostrou significantemente relacionada com a prevalecia de dentes quebrados.

Assim como as variáveis de controle Tipo e Posição, a covariável Sexo também é incluída no modelo final logístico multivariado. Não foi realizada nenhuma técnica de seleção de variáveis. Todas as características utilizadas neste estudo, foram selecionadas por motivações clínicas e a inclusão no modelo, independe da significância estatística aferida na análise exploratória.

2.4 Modelagem

Levando em consideração todas as peculiaridades do conjunto de dados, a modelagem proposta foi realizada utilizando o modelo logístico multinível. A indicadora de quebra do dente foi utilizada como variável dependente. A coloração do dente é o fator explicativo, mantendo controladas o sexo da criança e as características no nível do dente: posição e tipo. O modelo foi ajustado utilizando o estimação de equações generalizadas (GEE) com matriz de correlação intra indivíduo permutável. Todos os testes foram conduzidos utilizando o procedimento GENMOD no software SAS, com p-valores maiores que 0,05 sendo considerados estatisticamente significantes.

O modelo final com as variáveis explicativas e seus coeficientes estimados estão representados a seguir.

$$logito(\pi_{ij}) = \beta_0 + \beta_1 Cor(entre) + \beta_2 Cor(intra) + \beta_3 Posic(entre) \\ +\beta_4 Posic(intra) + \beta_5 Tipo(entre) + \beta_6 Tipo(intra) + \beta_7 Sexo$$

Através deste modelo, foi testada a hipótese nula de que a quebra do dente e a coloração são independentes. A Tabela 6 apresenta as estimativas obtidas para os parâmetros entre e intra-indivíduos, bem como o erro padrão, a razão de chances e o p-valor. Entre parênteses, está indicado qual é a categoria de referência utilizada em cada variável.

Tabela 6 – Estimativas dos parâmetros

Parâmetro	Coeficiente	Erro padrão	Razão de Chances	P-valor
Cor-entre (Não branco x Branco)	0,958	0,547	2,607	0,0798
Cor-intra (Não branco x Branco)	2,046	0,558	7,740	0,0002
Posição-entre (Superior x Inferior)	0,391	0,588	1,478	0,5067
Posição-intra (Superior x Inferior)	0,255	0,507	1,291	0,6148
Tipo-entre (Molar x Incisivo)	-0,110	0,642	0,896	0,8640
Tipo-intra (Molar x Incisivo)	1,090	0,439	2,975	0,0129
Sexo (Masculino x Feminino)	0,222	0,394	1,249	0,5721

No nível do dente, ou seja, intra-indivíduos, a variável Cor apresentou p-valor significativo (p=0,0002), rejeitando a hipótese nula de independência. Assim, temos evidências para concluir que existe associação entre a cor da opacidade e a chance de quebra. O coeficiente estimado apresenta valoração positiva indicando aumento na chance de quebra associado ao escurecimento do dente. Em termos de razão de chances, o dente que apresenta coloração não-branca tem em média, 7,74 vezes mais chance de se quebrar do que um dente branco. Entretanto, a análise no nível do indivíduo apresenta p-valor não significante, indicando que entre os indivíduos, não se identifica efeito da cor na variável

resposta. Nesse ponto, podemos perceber a importância da análise que considera todos os níveis da hierarquia, de outro modo, estimativas equivocadas podem ser obtidas.

A variável Tipo do dente apresentou significância ao nível do dente, revelando que dentes molares são mais suscetíveis à quebra, entretanto a diferença é de apenas 0,09%. O sexo da criança e a posição do dente apresentaram apresentaram p-valores não significantes em ambos os níveis da hierarquia.

Os resultados obtidos pela modelagem do problema confirmam as hipóteses dos pesquisadores.

3 Conclusão

Na pesquisa odontológica e também em outras áreas, uma variedade de trabalhos tratam de dados hierárquicos, lançando mão de diversas técnicas que permitem o estudo de dados correlacionados. Entretanto, é comum nos depararmos com estudos que implícita ou explicitamente, assumem que as fontes de informação entre e intra-indivíduos produzem a mesma estimativa para os efeitos dos fatores de risco ou tratamento. Geralmente, cada nível da hierarquia revela uma informação sobre os dados, podendo inclusive apresentar resultados divergentes. Nesses casos, negligenciar a decomposição das variáveis em fatores intra e entre-indivíduos pode resultar em estimativas enviesadas, de interpretação incoerente.

A utilização da abordagem da estimação de equações generalizadas (GEE), juntamente com a decomposição das covariáveis em fontes de informação nos níveis do dente e do indivíduo, mostrou resultados satisfatórios, corroborando com as expectativas clínicas dos pesquisadores. A análise bivariada em ambos os níveis apontaram associação entre a coloração da opacidade no esmalte e a quebra do dente. Entretanto, a análise multivariada através da regressão logística revelou que essa associação se dá principalmente no nível do dente. O efeito da coloração não branca aumenta, em média, 6,74% a fragilidade do dente. Nenhum efeito se mostrou significante ao nível da criança, confirmando a importância da análise em todos os níveis da hierarquia.

Conforme ilustrado neste estudo, covariáveis podem causar efeitos diferentes em cada nível da hierarquia. A pergunta seguinte é por que isso ocorre. Efeitos intra e entre-indivíduos podem diferir por viés causado por erros de medida ou confundimento causado por cofatores não observados (PALTA; YAO, 1991). Próximos estudos podem procurar investigar essa questão e ainda trazer uma comparação de métodos.

Referências

AGRESTI, A. *An Introduction to Categorical Data Analysis*. England: John Wiley and Sons INC, 2007. Nenhuma citação no texto.

BROWN, H.; PRESCOTT, R. *Applied Mixed Models in Medicine*. United Kingdom: John Wiley and Sons INC, 2006. Nenhuma citação no texto.

CARVALHO, J. et al. Impact of enamel defects on early caries development in preschool children. *Caries Res*, v. 45, n. 1, p. 353–360, 2011. Citado na página 24.

HOLT, A. J.; SCOTT, D. The effect of two-stage sampling on ordinary least square methods. *J Am Stat Assoc*, v. 77, p. 848–854, 1982. Citado na página 24.

HOSMER, D. W.; LEMESHOW, S. *Applied logistic regression*. Canada: John Wiley and Sons, 2000. ISBN 0471356328, 9780471356325. Nenhuma citação no texto.

JOHNSON, R.; WICHERN, D. *Applied Multivariate Statistical Analysis. Applied Multivariate Statistical Analysis*. New Jersey: Pearson Prentice Hall, 2007. Nenhuma citação no texto.

LIANG, K.-Y.; ZEGER, S. L. Longitudinal data analysis using generalized linear models. *Biometrika*, v. 73, p. 13–22, 1986. ISSN 2050-1439. Citado na página 17.

MANCL, L.; LEROUX, B.; DEROUENL, T. Between-subject and within-subject statistical information in dental research. *Journal of Dental Research*, n. 79, p. 1778–1781, 2000. Citado na página 25.

PALTA, M.; YAO, T. J. Analysis of longitudinal data with unmeasured confounders. *Biometrics*, v. 47, n. 2, p. 1355–1369, 1991. ISSN 1573-0565. Citado na página 31.