Universidade Estadual de Maringá
Programa de Pós-graduação em Bioestatística

TIAGO PERES DA SILVA SUGUIURA

Modelos Multiníveis: Gaussiano e Multinomial

Dissertação de Mestrado

Maringá - Paraná
2017

TIAGO PERES DA SILVA SUGUIURA

Modelos Multiníveis: Gaussiano e Multinomial

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá como requisito parcial para obtenção do título de Mestre em Bioestatística.
Orientadora: Profª. Drª. Isolde Previdelli

Maringá - Paraná
2017

.

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá – PR., Brasil)

S947m

Suguiura, Tiago Peres da Silva
Modelos multiníveis: gaussiano e multinomial / Tiago Peres da Silva Suguiura. -- Maringá, 2017.
127 f.: il. color., figs., tabs., mapas.

Orientadora: Prof.a. Dr.a. Isolde Terezinha Santos Previdelli
Dissertação (mestrado) – Universidade Estadual de Maringá, Centro de Ciências Exatas, Programa de Pós-graduação em Bioestatística, 2017.

1. Modelos multiníveis. 2. Modelagem estatística. 3. Modelo Linear misto. 4. Modelo linear generalizado misto. 5. Modelo multinominal. 6. Afastamento gengival vertical. 7. Câncer de mama. I. Previdelli, Isolde Terezinha Santos, orient. II. Universidade Estadual de Maringá. Centro de Ciências Exatas. Programa de Pós-Graduação em Bioestatística. III. Título.

CDD 22. ED.570.15195

JLM-001938

TIAGO PERES DA SILVA SUGUIIURA

Modelos Multiníveis: Gaussiano e Multinomial

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá, como requisito parcial para a obtenção do título de Mestre em Bioestatística.

BANCA EXAMINADORA

Profª. Dra. Isolde Previdelli
Universidade Estadual de Maringá - UEM

Profª. PhD. Lurdes Yoshiko Tani Inoue
University of Washington – EUA

Profª. Dra. Rosangela Getirana Santana
Universidade Estadual de Maringá - UEM

Maringá, 28 de março de 2017.

À minha família.

Agradecimentos

A longa jornada que tem como resultado essa dissertação não poderia ter sido completada sem a ajuda, pessoal ou não, de diversas pessoas e instituições.

Primeiramente, agradeço minha família, sem a qual não teria fundação para continuar essa jornada. Agradeço minha mãe, Sandra, por todo seu suporte e incentivo, desde os momentos mais difíceis até os momentos de felicidade, por toda a sua vida se dedicando em me ajudar conquistar meus objetivos, buscando sempre me auxiliar de qualquer maneira possível, seja acadêmica ou pessoal. Juntamente agradeço meu padrasto Maurício, por seu auxílio, convivência e sabedoria em minha criação. E também agradeço minha irmã Isabela pelos momentos de diversão. Agradeço meus avós, assim como minhas tias e tio de Campo Grande, se tornando uma válvula de escape sempre quando necessário.

Agradeço a minha "família de Maringá". Tia Néia, Vagner, Lívia, Monique, meu pai Carlos, Flávia, Laura e o Augusto, por todos os momentos juntos, todas as viagens, todos os ensinamentos e todas as ajudas necessárias durante todos os anos da minha estadia em Maringá. Sem vocês aqui, essa jornada seria muito mais difícil. Meus avós de Dourados, assim como as demais tias, por sempre torcerem por mim.

Agradeço minha orientadora Isolde Previdelli, sem a qual nada disso seria possível. Além de todo conhecimento acadêmico compartilhado, as lições de vida, as duras e conselhos, e por nunca me deixar desistir quando todos os problemas estavam em minha frente. Mesmo com alguns conflitos e decepções, em nenhum momento desistiu de mim e da minha capacidade de concluir esse trabalho.

Agradeço de coração todas as amizades construídas nesse período, seja dentro da UEM como fora. Desde as amizades mais antigas, da época de graduação em especial Eduardo, Dani, Nayara, Helder, Gabriela, Caroline, Arthur, Thales, Jorge, Thiago, Guilherme e Giovana, e as amizades mais recentes e não menos importantes feitas durante o mestrado, Zé, Marcos, Márcia, Omar, Felipe, Ricardo, Diego, Edilenia, Beatriz, Emerson, Aline, entre outros.

Agradeço o Programa de Pós Graduação em Bioestatística da Universidade Estadual de Maringá, seus professores, a coordenação e a secretaria, por contribuírem com a minha formação.

Agradeço à CAPES e à Fundação Araucária pelo suporte financeiro.

"There are places I'll remember
All my life
Though some have changed
Some forever
Not for better
Some have gone and some remain
All these places have their moments
With lovers and friends
I still can recall
Some are dead and some are living
In my life
I've loved them all."
- **Lennon/McCartney**

Resumo

A utilização de modelos multiníveis na análise de dados provenientes da odontologia e da saúde pública têm crescido recentemente devido às características dos estudos. Na odontologia, por exemplo, os “lados”/“raízes” dos dentes estarem aninhados em um dente, que por sua vez está aninhado em cada paciente. Já na saúde pública, as doenças, cuja propensão pode ser influenciada pelo espaço geográfico e hábitos culturais. Análises estatísticas assumindo independência das unidades observacionais é inapropriado e, portanto, deve-se adotar metodologias que acomodem estas características, que é o caso da modelagem multinível. O trabalho foi dividido em duas partes, primeiramente para avaliar o afastamento gengival vertical, onde é apresentado uma comparação entre modelos com e sem estruturas hierárquicas, e na segunda parte, fazendo uma revisão da teoria de modelos multiníveis multinomiais em um estudo do tipo ecológico sobre a distribuição das internações por câncer da mama pelo SUS no estado do Paraná entre 2008 a 2016, levando em consideração as 9 topografias do câncer de mama segundo a Classificação Internacional de Doenças. Os resultados obtidos para o primeiro trabalho indicam que a substância analisada não é eficaz para o aumento gengival vertical, e que o modelo multinível se ajusta melhor à estrutura hierárquica do experimento. Para o segundo trabalho notou-se um melhor ajuste do modelo multinível multinomial para analisar as ocorrências do câncer de mama por CID, e apresentou diferenças entre as ocorrências em termos de Regionais de Saúde (Macro Regionais), Raça e Faixa etária.

Palavras-chave: Modelos multiníveis, modelos multinomiais, modelo misto, modelo generalizado misto, afastamento gengival vertical, câncer de mama.

Abstract

The use of multilevel models in data sets from dentistry and public health has grown recently due to the experiment characteristics. In dentistry, for example, “side”/“root” of a tooth nested within a tooth and the tooth nested within each patient. Considering the public health, with diseases, with rates that can be influenced by the geographic space and culture habits, and for this reason, statistical analysis assuming independence of observed units is inappropriate, therefore, methodologies which take in account this characteristic must be adopted, which is the case of multilevel modeling. This work was separated in two parts, first, to evaluate the vertical gingival retraction, with a comparison between models with and without hierarchical structure, and, in the second part, making a literature review on multinomial multilevel model theory with the goal of analyzing an ongoing research about the distribution of visits due to breast cancer in hospitals from SUS in the state of Paraná from 2008 to Oct/2016, considering the 8 sub-topographies of breast cancer according to the International Classification of Diseases. The results from the first study showed that the analyzed substance is not effective to increase the vertical gingival retraction, and the multilevel model had a better adjustment to the hierarchical structure of the experiment. For the second study a better fit of the multinomial multilevel model was verified to analyze the occurrences of breast cancer by ICD, and presented differences between the occurrences in terms of Health Regions (Macro Regions), Race and Age Range.

Key words: Multilevel models, multinomial models, mixed model, generalized mixed model, vertical gingival retraction, breast cancer.

Lista de figuras

Lista de tabelas

Sumário

Capítulo 1

Introdução

A maior parte da análise estatística assume observações independentes, e essa suposição de independência indica que as respostas sobre a unidade experimental não estão correlacionadas entre si. Porém, quando se tratam de experimentos com uma hierarquia natural (pacientes de um mesmo hospital, informações de um mesmo paciente, vários dentes de um mesmo paciente), respostas de pessoas do mesmo grupo/*cluster* tendem a exibir certo grau de relação. A abordagem multinível ajusta-se às necessidades inerentes ao desenho amostral, uma vez que pode existir correlação entre os níveis da hierarquia, a qual, através de uma análise que não leva em consideração tais fatos, não é detectada. Segundo Hancock e Mueller (2010), as vantagens de modelos multiníveis não são somente estatísticos, tal que a análise multinível possui uma versatilidade que nos permite explorar a informação contida em *clusters* para explicar a variabilidade de uma variável resposta tanto “entre os grupos” como “dentro do grupo”. Ou seja, em análise multinível, estimamos e modelamos explicitamente o grau de relação de observações dentro do mesmo grupo, assim estimando corretamente os erros padrão e minimizando o problema de termos quantidades de erros do Tipo I inflacionados.

A ampla abordagem que os modelos multiníveis proporcionam podem estar ligados a estudos de diversas áreas, como na saúde com doenças epidemiológicas; na demografia, em estudos de crescimentos populacionais ou tendências de desenvolvimentos econômicos; na sociologia, em gestão organizacional ou criminologia, para o estudo de comportamentos e atitudes a nível de comunidade e individuais; na zoologia, para análise e previsão de crescimento de colônias e adaptação a diferentes meios; na educação, para estudos em âmbito escolar, entre outros.

Diante do exposto, foi apresentada a metodologia de modelos multiníveis gaussianos a partir do Capítulo 3 através de um caso em odontologia na qual analisamos os dentes aninhados em pacientes.

A partir do Capítulo 4, utilizando uma pesquisa com 24 pacientes de uma clínica particular de Londrina-PR, foi feita uma avaliação se o Cloridrato de Nafazolina (Colírio Legrand), quando utilizado como agente de retração gengival é capaz de aumentar o afastamento gengival em comparação com a substância padrão. Para tal, foi utilizado um modelo multinível gaussiano, cuja variável resposta é o afastamento gengival em milímetros.

Na segunda parte, a partir do Capítulo 6, foi apresentada a metodologia de modelos multiníveis multinomiais com aplicação na saúde pública, investigando a frequência de ocorrência de algum tipo de câncer de mama de acordo com a Regional de Saúde aninhada dentro das Macro Regionais no estado do Paraná.

A partir do Capítulo 7 foi realizado um estudo tipo ecológico utilizando as informações de internação hospitalar por câncer de mama no estado do Paraná, de 2008 a 2016 obtidas por meio do DATASUS, com o intuito de analisar os tipos de câncer de mama (topografias) em relação às 4 Macro Regionais que estão subdivididas em 22 Regionais de Saúde no estado do Paraná. Para esse estudo, será utilizado um modelo multinível multinomial cuja variável resposta será os tipos de câncer de mama (9 no total).

Estas pesquisas confirmam a interdisciplinaridade do Programa de Pós Graduação em Bioestatística com o Programa de Pós Graduação em Odontologia e o Programa de Pós Graduação em Ciências da Saúde.

Parte I

Modelo Multinível Gaussiano na avaliação do afastamento gengival vertical

Capítulo 2

Introdução

Em 1976, Bennett et al. (1976) publicou um importante estudo sobre estilos de ensino com alunos do ensino fundamental na Inglaterra, no qual os resultados sugeriam que métodos formais de ensino eram associados com maior progresso nas habilidades básicas dos alunos, causando controvérsias consideráveis. Os dados foram analisados utilizando modelos de regressão múltipla tradicionais que reconheciam somente as crianças individualmente como as unidades de análise e ignoraram seus agrupamentos com a classe e os professores.

Os resultados obtidos foram estatisticamente significantes, porém, Aitkin et al. (1981) demonstraram que, considerando a análise com as crianças agrupadas em classes, as diferenças significantes desapareciam e as crianças que receberam tal ensino formal não mostraram diferenças com aquelas que não o receberam.

Segundo Goldstein (2011), essa re-análise é o primeiro exemplo importante de uma análise multinível em dados de ciências sociais. Nesse caso em estudo, qualquer uma das crianças, por estarem na mesma sala e serem ensinados juntos, tendem a ser similares em suas performances. Como resultado, eles forneceram menos informação do que teriam fornecido caso o mesmo número de estudantes tivessem sido ensinados separadamente por diferentes professores.

Métodos estatísticos e algoritmos foram desenvolvidos e, em meados de 1986, a base da análise multinível estava estabelecida. Alguns artigos importantes são: Laird e Ware (1982) que compara modelos multivariados com uma estrutura de covariância geral com modelos de efeitos aleatórios de dois estágios para dados desbalanceados , Mason et al. (1983) onde é descrito e proposto o processo de estimação da máxima verossimilhança restrita/Bayes (REML/Bayes) para um modelo multinível, enquanto Goldstein (1986) demonstra um procedimento iterativo de estimação utilizando mínimos quadrados generalizados para dados com hierarquia e o compara com a máxima verossimilhança no caso normal e os

pesquisadores Aitkin e Longford (1986) discutem o uso geral de componentes de variância ou modelos de “parâmetros aleatórios” para a análise de estudos envolvendo observações agrupadas ou *clusters*.

Para a odontologia, é comum em dados clínicos de doenças periodontais (conjunto de condições inflamatórias e de caráter crônico e origem bacteriana que afetam a gengiva) entre outras, os “lados” dos dentes estarem aninhados em um dente, que por sua vez estão aninhados em cada paciente e dividem características comuns de observações correlacionadas. Devido a sua estrutura hierárquica, análises assumindo independência das observações é inapropriado e, portanto, metodologias que tentam agrupar os dados em nível de paciente resulta em perda de informação valiosa e podem não refletir a associação específica. O uso de efeitos aleatórios em modelagem multinível é uma maneira comum e conveniente de modelar tal estrutura de grupos. É, portanto, além de apropriado, necessário considerar “lado do dente” e “dente dentro do paciente” como fatores aleatórios (MDALA et al., 2012).

Segundo Gilthorpe Mark S (2000), seu estudo demonstra o benefício de modelos multiníveis sobre técnicas convencionais em estudos odontológicos. Como um resultado de desenvolvimento computacional através de velocidade de processamento e memória, a regressão multinível pode ser utilizada com mais facilidade e, consequentemente, pesquisadores estão melhor equipados para analisar estruturas de dados mais complexos, particularmente dentro da odontologia onde dados multiníveis e multivariados são comuns.

Capítulo 3

Modelo Multinível Gaussiano

Um modelo que apresenta somente fatores de efeitos fixos além do erro experimental, que é sempre aleatório, é denominado modelo fixo. Os modelos que apresentam apenas fatores de efeitos aleatórios exceto a constante relativa à média, que é sempre fixa, é denominado modelo aleatório. Um modelo misto é aquele que apresenta tanto fatores de efeitos fixos como aleatórios além do erro experimental e da média. Nesses casos, os efeitos fixos estão estimando os coeficientes relativos à população, enquanto os efeitos aleatórios estão levando em conta as diferenças entre os indivíduos em resposta ao tratamento em questão.

Modelos de efeitos mistos são primariamente usados para descrever a relação entre a variável resposta e alguma covariável nos dados que pode estar agrupada de acordo com um ou mais fatores de classificação.

Exemplos de planejamentos estatísticos que utilizam dados agrupados são abrangentes entre: *cross-over*, o qual é um estudo que compara dois ou mais tratamentos ou intervenções nos quais os pacientes, após terminado o curso de um tratamento, são ligados a outro; *Split-plot*, no qual os blocos que são divididos são novamente analizados; e até em estudos temporais como *dados longitudinais*, no qual a mesma unidade experimental é analisada várias vezes no decorrer do tempo, surgindo o interesse de verificar o comportamento individual do sujeito que gerou a medida repetida (PINHEIRO; BATES, 2006).

Para experimentos com uma certa estrutura hierárquica, análises estatísticas tradicionais que assumem independência produzirão erros padrão incorretos. Em tal cenário, as estimativas dos erros padrão são subestimados e, portanto, os erros do Tipo I são inflacionados por todos os testes estatísticos que utilizam a suposição de independência (HANCOCK; MUELLER, 2010).

As análises deste trabalho foram realizadas com o *software* estatístico `R` (versão 3.2.1) (R Core Team, 2015) com o uso do pacote `nlme` (PINHEIRO et al., 2016).

3.1 Modelos Mistos

Dados correlacionados surgem frequentemente em análises estatísticas, sejam por meio de agrupamento de sujeitos, ou através de medidas repetidas na mesma unidade experimental, seja ao longo do tempo ou sem considerar o tempo, como por exemplo, um sujeito receber varios tipos de tratamentos (*cross-over*). As análises de modelos mistos fornecem uma abordagem geral e flexível nessas situações, pois permitem uma grande variedade de estruturas de correlação para serem modeladas. Estudos sobre modelos mistos estão disponíveis na literatura de maneira abrangente e ao longo dos anos vários autores se dedicaram a estudar esse tipo de modelagem, e novas abordagens surgiram. Nesse trabalho, seguiremos as notações de acordo com Singer e Andrade (1986), no qual apresentam dados provenientes de estudos com medidas repetidas por meio de modelos mistos da seguinte forma

$$\mathrm{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon_i} \tag{3.1}$$

para $i = 1, 2, \cdots, N$ onde N é o número total de unidades experimentais, $\mathrm{y}_i = (y_{i1}, \cdots, y_{in_i})^T$ com dimensão $(n_i \times 1)$ é o perfil de respostas da i-ésima unidade experimental, $\boldsymbol{\beta}$ é um vetor com dimensão $(p_i \times 1)$ de parâmetros (efeitos fixos) desconhecidos, $\mathbf{X}$ é uma matriz de especificação dos efeitos fixos com dimensão $(n_i \times p)$, conhecida e de posto completo, $\mathbf{b}_i$ é um vetor com dimensão $(q_i \times 1)$ de parâmetros para os efeitos aleatórios, $\mathbf{Z}_i$ é uma matriz de especificação dos efeitos aleatórios com dimensão $(n_i \times q)$ conhecida e de posto completo e $\boldsymbol{\varepsilon}_i$ é um vetor de erros aleatórios com dimensão $(n_i \times 1)$.

Para esse trabalho assumiremos que $\mathbf{b}_i \sim \mathbf{N}_q(\mathbf{0}, \mathbf{G})$ e $\varepsilon_i \sim \mathbf{N}_{n_i}(\mathbf{0}, \mathbf{R}_i)$, em que $\mathbf{G}$, com dimensão $(q \times q)$ e $\mathbf{R}_i$ com dimensão $(n_i \times n_i)$ são matrizes simétricas definidas positivas e além disso, $\mathbf{b}_i$ e $\boldsymbol{\varepsilon}_i$ são variáveis aleatórias independentes.

Quando $\mathbf{R}_i = \sigma^2\mathbf{I}_{p_i}$, o modelo é chamado de **modelo de independência condicional homocedástico**.

Sob esse modelo, o vetor de respostas associado à i-ésima unidade amostral tem distribuição normal multivariada com vetor de média e matriz de covariâncias dados, respectivamente, por

$$\mathbb{E}(\mathrm{y}_i) = \mathbf{X}_i\boldsymbol{\beta} \tag{3.2}$$

e

$$\mathbb{V}(\mathrm{y}_i) = \mathbf{V}_i = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i^T + \mathbf{R}_i\,. \tag{3.3}$$

Consequentemente, temos que o modelo marginal é dado por:

$$\mathrm{y}_i \sim \mathbf{N}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i^T + \mathbf{R}_i)\,. \tag{3.4}$$

Assumimos que os q efeitos aleatórios do vetor $\mathbf{b}_i$ seguem uma distribuição normal multivariada, com média **0** e matriz de covariância $\mathbf{G}$ (WEST et al., 2014).

Os elementos sobre a diagonal principal da matriz $\mathbf{G}$ representam as variança de cada efeito aleatório em $\mathbf{b}_i$, e os elementos das outras diagonais representam as covariâncias entre dois efeitos aleatórios correspondentes. Se o vetor $\mathbf{b}_i$ possuir q efeitos aleatórios associados ao modelo, temos que $\mathbf{G}$ é a seguinte matriz simétrica definida positiva

$$\mathbf{G} = Var(\mathbf{b}_i) = \begin{bmatrix} Var(b_{1i}) & cov(b_{1i}, b_{2i}) & \cdots & cov(b_{1i}, b_{qi}) \\ cov(b_{1i}, b_{2i}) & Var(b_{2i}) & \cdots & cov(b_{2i}, b_{qi}) \\ \vdots & \vdots & \ddots & \vdots \\ cov(b_{1i}, b_{qi}) & cov(b_{2i}, b_{qi}) & \cdots & Var(b_{qi}) \end{bmatrix}$$

Da mesma forma temos que os n_i resíduos no vetor $\boldsymbol{\varepsilon}_i$ são variáveis aleatórias que seguem uma distribuição normal multivariada com média **0** e uma matriz de covariância definida positiva e simétrica dada por $\mathbf{R}_i$, definida da seguinte maneira

$$\mathbf{R}_i = Var(\boldsymbol{\varepsilon}_i) = \begin{bmatrix} Var(\varepsilon_{1i}) & cov(\varepsilon_{1i}, \varepsilon_{2i}) & \cdots & cov(\varepsilon_{1i}, \varepsilon_{n_i i}) \\ cov(\varepsilon_{1i}, \varepsilon_{2i}) & Var(\varepsilon_{2i}) & \cdots & cov(\varepsilon_{2i}, \varepsilon_{n_i i}) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\varepsilon_{1i}, \varepsilon_{n_i i}) & cov(\varepsilon_{2i}, \varepsilon_{n_i i}) & \cdots & Var(\varepsilon_{n_i i}) \end{bmatrix}$$

A matriz $\mathbf{X}_i$ de ordem $(n_i \times p)$ é a especificação dos efeitos fixos, ou seja, representa os valores conhecidos das t covariáveis, e é definida como

$$\mathbf{X}_i = \begin{bmatrix} X_{1i}^{(1)} & X_{1i}^{(2)} & \cdots & X_{1i}^{(p)} \\ X_{2i}^{(1)} & X_{2i}^{(2)} & \cdots & X_{2i}^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n_i i}^{(1)} & X_{n_i i}^{(2)} & \cdots & X_{n_i i}^{(p)} \end{bmatrix}$$

A matriz $\mathbf{Z}_i$ de ordem $(n_i \times q)$ é a especificação dos efeitos aleatórios, ou seja, representa os valores conhecidos das q covariáveis. A matriz $\mathbf{Z}_i$ dada por

$$\mathbf{Z}_i = \begin{bmatrix} Z_{1i}^{(1)} & Z_{1i}^{(2)} & \cdots & Z_{1i}^{(q)} \\ Z_{2i}^{(1)} & Z_{2i}^{(2)} & \cdots & Z_{2i}^{(q)} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n_i i}^{(1)} & Z_{n_i i}^{(2)} & \cdots & Z_{n_i i}^{(q)} \end{bmatrix},$$

será estruturada de acordo com a disposição dos dados.

Segundo West et al. (2014), em muitos casos, variáveis preditoras com efeitos que variam aleatoriamente entre indivíduos são representados em ambas as matrizes $\mathbf{X}_i$ e $\mathbf{Z}_i$.

Por exemplo, em um modelo linear misto no qual somente os interceptos são aleatórios, a matriz $\mathbf{Z}_i$ será simplesmente composta por uma coluna de 1's.

Para esse estudo foi utilizado como notação para o efeito aleatório representando o intercepto do i-ésimo paciente, o termo $\mathbf{b}_{0i}$.

3.2 Acréscimo do Efeito Multinível

O acréscimo do efeito multinível não afeta a estimativa dos efeitos fixos, somente a parte dos efeitos aleatórios. Com isso, o vetor resposta ainda possui uma distribuição normal multivariada com vetor de média dado por

$$\mathbb{E}(\mathrm{y}_i) = \mathbf{X}_i\boldsymbol{\beta}, \tag{3.5}$$

já a matriz de covariância $\mathbb{V}(\mathrm{y}_i)$ sofrerá alterações.

Assim como os elementos de $\mathbf{b}_{0i}$ seguem uma distribuição normal multivariada com média **0** e matriz de covariância $\mathbf{G}_1$, o mesmo acontece com o efeito multinível acrescentado, ou seja, $\mathbf{b}_{0j|i}$ seguirá uma distribuição normal multivariada com média $\mathbf{0}$ e uma matriz de covariância $\mathbf{G}_2$.

Da mesma maneira que existe a matriz $\mathbf{Z}_{1i}$ que representa a especificação dos efeitos aleatórios em nível de paciente $\mathbf{b}_{0i}$, existe uma matriz $\mathbf{Z}_{2i}$ que representa a especificação dos efeitos aleatórios multiníveis acrescentados (em nivel de dente).

Portanto, o vetor resposta possui uma distribuição normal multivariada com matriz de covariância dada por

$$\mathbb{V}(\mathrm{y}_i) = \mathbf{Z}_{1i}\mathbf{G}_1\mathbf{Z}_{1i}^T + \mathbf{Z}_{2i}\mathbf{G}_2\mathbf{Z}_{2i}^T + \mathbf{R}_i\,. \tag{3.6}$$

3.3 Métodos de Estimação

Para estimar os parâmetros do modelo multinível existem varias metodologias disponíveis na literatura, o método bayesiano detalhado em Searle et al. (2009), os métodos de Máxima Verossimilhança e Máxima Verossimilhança Restrita discutidos em Harville (1977) e Robinson (1991), o método de Mínimos Quadrados Generalizados e Equações de Estimação Generalizadas (*Generalized Estimating Equations*- GEE) em Draper e Smith (2014) entre outros. Para auxílio, existem os pacotes `nlme, multilevel` e `lme4` para o *software* `R (3.2.1)` e os procedimentos `PROC MIXED` e `PROC GLIMMIX` para o *software SAS (9.3)*.

Como um dos objetivos desse estudo era compreender as características das matrizes de um modelo misto, foi utilizado, como método de estimação, as Equações dos Modelos Mistos de Henderson, e será apresentada na Seção 4.2.

Capítulo 4

Modelo Multinível Gaussiano na Avaliação do Afastamento Gengival Vertical

O estudo no qual foi utilizado a metodologida de modelos multiníveis exibida no capítulo anterior é um ensaio clínico desenvolvido pelo Programa de Pós Graduação de Odontologia da Universidade Estadual de Maringá com o objetivo de avaliar o efeito do uso do Cloridrato de Nafazolina, em comparação com o Cloreto de Alumínio (elemento de referência), no afastamento gengival vertical. Para tal, foram utilizados 24 pacientes de um consultório particular da cidade de Londrina-PR. Mediante declaração por escrito, com comprovação do Comitê de Ética da Universidade Estadual de Maringá de acordo com o parecer de número 1.515.263 (vide Apêndice A).

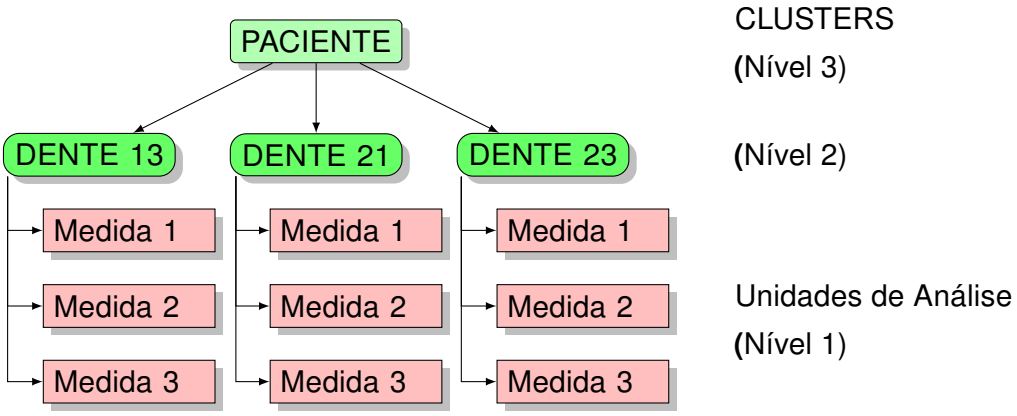
Foram utilizados como critérios de inclusão: boa condição sistêmica geral, saúde periodontal e biotipo gengival espesso. E como critérios de exclusão: fumantes, dentes caninos ou incisivos superiores com lesões de cárie, abrasão, erosão, pilares protéticos ou restaurações insatisfatórias, portadores de doença periodontal, assim como a utilização de medicamento contínuo.

Os procedimentos foram realizados em duas sessões clínicas: primeiramente para avaliação do afastamento gengival vertical e depois para avaliação do fluxo de líquidos no interior do sulco gengival antes e após o afastamento da gengiva.

Cada um dos 24 pacientes teve três dentes avaliados (dentes 13, 21 e 23). Todos os dentes avaliados receberam fios afastadores e foram separados em 3 grupos de forma randomizada para receberem tais fios. Cada um dos dentes recebeu um fio afastador com substâncias químicas diferentes ou sem substância. De cada dente foram retirados três medidas, as quais foram utilizadas como unidades de análise.

No diagrama abaixo representamos uma estrutura hierárquica com medidas repeti-

das, a qual representa a estrutura dos dados do estudo em questão.



Note que as unidades de análise (nível 1) são as 3 medidas retiradas de cada dente, e os grupos/*clusters* hierárquicos são os dentes de cada paciente.

Inicialmente foi realizado o isolamento relativo com roletes de algodão nas áreas correspondentes aos dentes a serem avaliados. Após a limpeza com fio dental e bolinha de algodão embebida em *clorexidine* a 2% foi feito o enxágue e a secagem dos dentes. Uma camada de protetor de gengiva fotopolimerizável *Top dam*® foi aplicada à superfície dental dos elementos 13, 21 e 23 ao nível da margem cervical do sulco gengival, para registrar a posição inicial da mesma conforme a Figura 1.

Figura 1 – Aplicação do protetor gengival fotopolimerizável *Top dam*®

Após ser removido todo o excesso do contorno gengival com o auxílio de uma sonda,

o gel foi polimerizado. Em seguida, utilizando a técnica do duplo fio, foram posicionados os fios retratores conforme a Figura 2. Inicialmente foi posicionado o fio retrator (*ultrapak*) nº 000 no interior do sulco gengival da face vestibular de cada um dos dentes. Em seguida, o fio retrator (*Ultrapak*®) nº1 foi posicionado de forma randomizada sobre o primeiro fio. Após a colocação do fio 000, o fio retrator (*Ultrapak*®) nº 1 foi instalado no primeiro dente embebido em cloridrato de nafazolina (*Legrand*® - Grupo 1), sobre o segundo dente, embebido em cloreto de alumínio (*Hemostop*® - Grupo 2), e sobre o terceiro dente, o fio foi posicionado sem qualquer tipo de substância (Contole - Grupo 3).

Figura 2 – Técnica de aplicação do fio retrator.

Os fios ficaram embebidos nas respectivas soluções por 7 minutos antes de serem aplicados nos dentes. Após um período de quatro minutos, os fios retratores foram retirados do sulco gengival, a área foi seca com jatos de ar e a moldagem foi realizada utilizando silicona de adição (*polivinil siloxana*) (3D - *Angelus*, Londrina - Brasil).

Após a tomada de presa da silicona de adição a moldeira foi removida da boca. Decorridas duas horas desta moldagem, o molde foi vazado em gesso especial tipo IV, os modelos foram então recortados em pequenos blocos, e a partir destes, 72 imagens dos modelos (uma imagem por dente que sofreu afastamento gengival) foram capturadas por uma câmera, acoplada a uma lupa (Olympus SZ-STS) como mostra a Figura 3. As imagens foram analisadas através do programa *Image pro-plus* (versão 4.5) para medir a distância entre o protetor gengival *Top dam*® (que marca a posição inicial da gengiva) até o nível gengival.

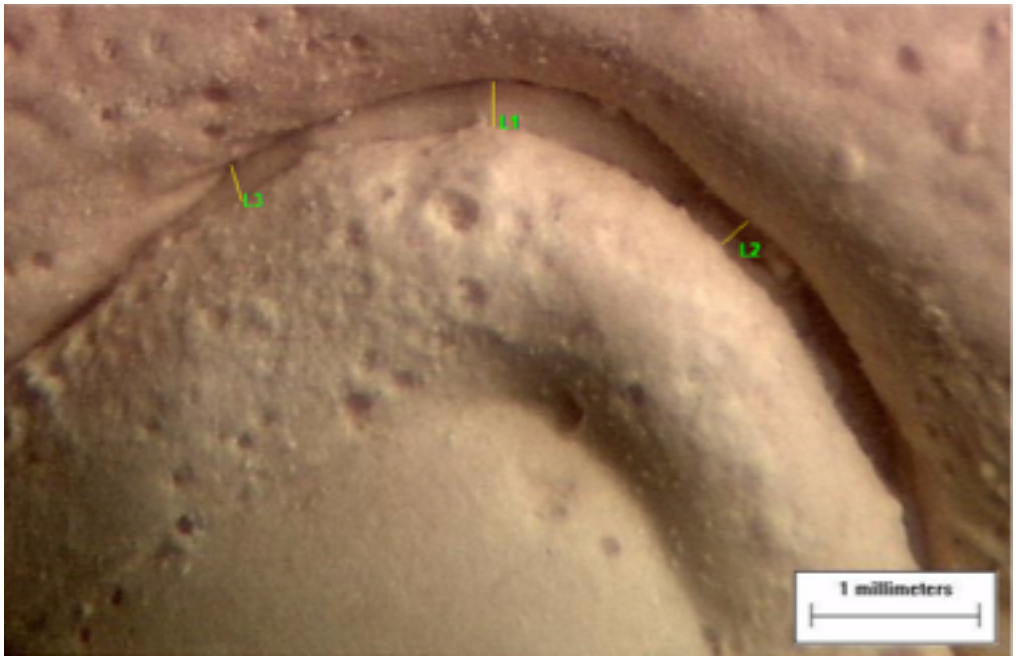


Figura 3 – Molde em gesso.

Essas medidas foram realizadas por um examinador cego aos procedimentos. Todas as imagens foram obtidas entre 24 e 96 horas após o vazamento do gesso.

4.1 Descrição dos Dados

As informações que o banco de dados podem prover é de suma importância para o entendimento completo da análise estatística e, para o estudo em questão, informações de como foram distribuídos os tratamentos e os pacientes, assim como sua estrutura, são necessárias para obtermos o modelo estatístico que mais se ajustará ao dados.

Na Tabela 1 está disposta uma parte da relação de pacientes, dentes e o tratamento recebido por cada dente. Os tratamentos foram distribuídos para cada dente de maneira aleatória.

Tabela 1 – Relação entre pacientes e tratamento para cada dente.

	DENTE 13	**DENTE 21**	**DENTE 23**
PACIENTE 1	Hemostop	Colírio	Placebo
PACIENTE 2	Colírio	Placebo	Hemostop
PACIENTE 3	Placebo	Hemostop	Colírio
PACIENTE 4	Hemostop	Colírio	Placebo
PACIENTE 5	Colírio	Placebo	Hemostop
PACIENTE 6	Placebo	Hemostop	Colírio
PACIENTE 7	Hemostop	Colírio	Placebo
PACIENTE 8	Colírio	Placebo	Hemostop
PACIENTE 9	Placebo	Hemostop	Colírio
⋮	⋮	⋮	⋮
PACIENTE 24	Hemostop	Colírio	Placebo

A Tabela 2 exibe parte da disposição dos valores do afastamento gengival vertical contidos no banco de dados, assim como sua estrutura hierárquica.

Tabela 2 – Disposição das avaliações do afastamento gengival vertival por pacientes e por dentes.

Pacientes - NÍVEL 3	**Dentes - Nível 2**	**Unidades de Análise - Nível 1**		
PACIENTES	**DENTES**	**MEDIDA 1**	**MEDIDA 2**	**MEDIDA 3**
Paciente 1	Dente 13	0,25778	0,187796	0,183607
Paciente 1	Dente 21	0,154878	0,136879	0,161837
Paciente 1	Dente 23	0,216886	0,203544	0,18065
Paciente 2	Dente 13	0,232762	0,279884	0,141471
Paciente 2	Dente 21	0,427033	0,383964	0,520655
Paciente 2	Dente 23	0,440261	0,27015	0,299254
Paciente 3	Dente 13	0,304009	0,203544	0,185695
Paciente 3	Dente 21	0,223543	0,248008	0,224964
Paciente 3	Dente 23	0,222686	0,204482	0,196372
⋮	⋮	⋮	⋮	⋮
Paciente 24	Dente 23	0,239643	0,248008	0,203231

O comportamento da variável resposta (afastamento) pode ser visto na Figura 4 separados por tratamento.

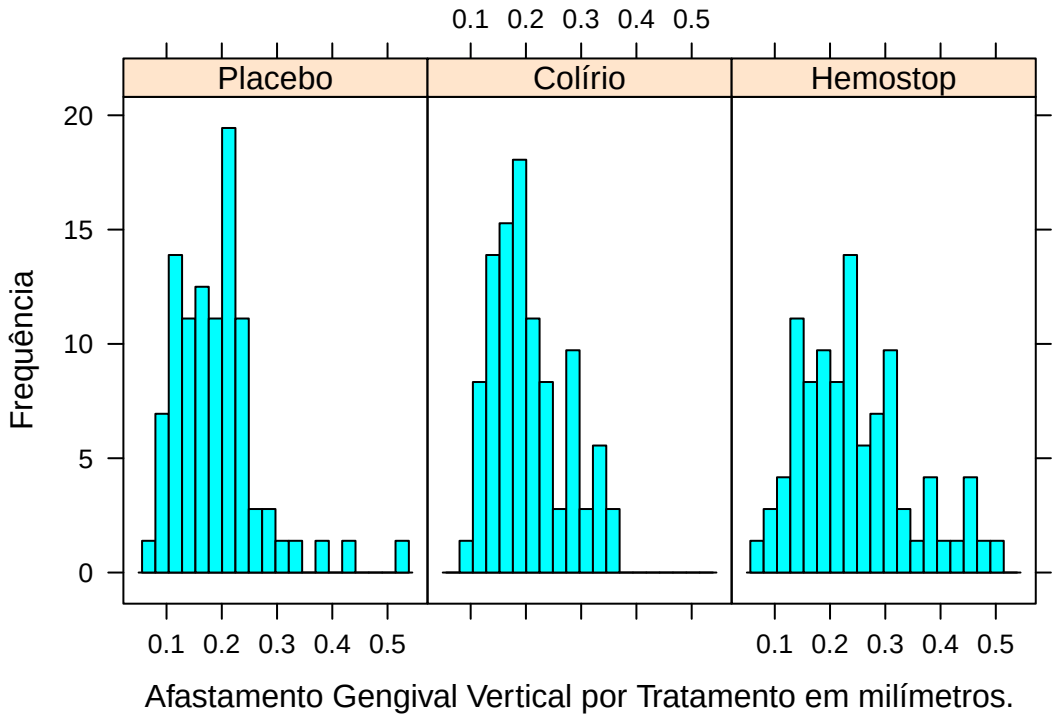


Figura 4 – Distribuição do afastamento gengival vertical por tratamento.

A Tabela 3 mostra as informações das medidas de posição e dispersão dos dados por tratamento, valor mínimo, máximo, mediana, média e desvio padrão.

Tabela 3 – Medidas de posição e dispersão por tratamento

TRATAMENTO	**Min.**	**Mediana**	**Média**	**Max.**	**d.p.**
Placebo	0.07361	0.18140	0.19200	0.52070	0.07767
Colírio	0.09614	0.19470	0.20550	0.36490	0.06669
Hemostop	0.07359	0.23590	0.24640	0.49850	0.10005

Juntamente com as informações da Tabela 3, a Figura 5 evidencia a distribuição do afastamento gengival separados por tratamento. Podemos notar que há pouca mudança da variância dos tratamentos, principalmente entre o placebo e o tratamento com o agente a ser testado, e somente uma pequena diferença entre os valores das medianas para cada tratamento.

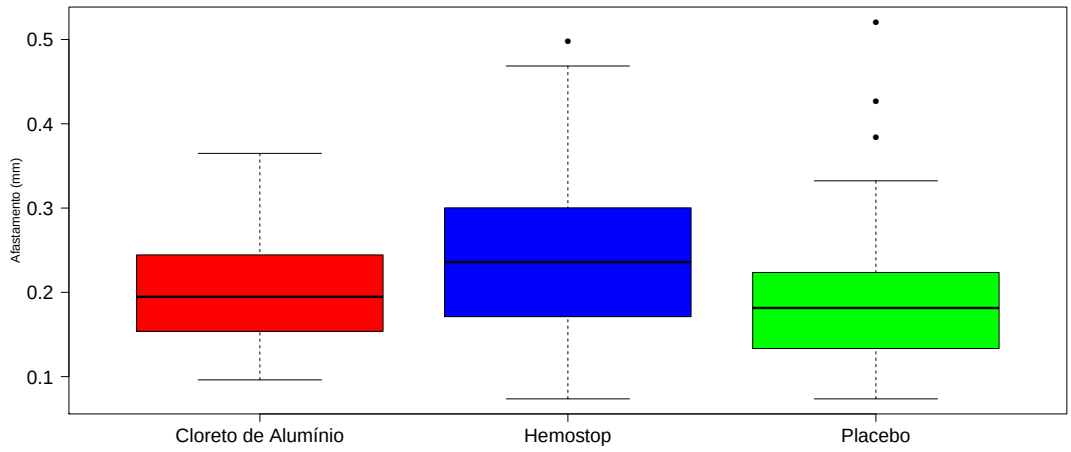


Figura 5 – Distribuição do afastamento gengival por tratamento.

Podemos perceber, apenas com a análise descritiva, que os dados do grupo controle (placebo) se assemelham bastante com os dados do grupo Colírio - Cloreto de Alumínio, que é o tratamento que queremos analisar, enquanto os dados do tratamento usual (Hemostop), apresenta valores superiores para o afastamento gengival.

4.2 O Modelo

O modelo multinível adotado para a observação do afastamento gengival no dente j para o paciente i é dado abaixo:

$$Afastamento_{ij} = \beta_0 + b_{0i} + \mathbf{b}_{0j|i} + \beta_1 \times TRAT_{ij} + \varepsilon_{ij} \tag{4.1}$$

- β_0 representa o efeito fixo associado à média;
- β_1 representa o efeito fixo associado ao tratamento por cada dente ($j = 1, 2, 3$);
- b_{0i} representa o efeito aleatório do paciente ($i = 1, \ldots, 24$);
- $\mathbf{b}_{0j|i}$ representa o efeito aleatório do dente dentro de cada paciente;
- ε_{ij} representa os erros.

O objetivo de qualquer modelo de regressão é obter o **melhor estimador linear não-enviesado** (*best linear unbiased estimator* -**BLUE**) para β e, no caso, o **melhor preditor linear não-enviesado** (*best linear unbiased predictor* - **BLUP**) para o vetor de efeitos aleatório ($\mathbf{b}$). Diferentes formas de obtenção do BLUP e BLUE, tanto sob o ponto de vista

clássico como o bayesiano e aplicações podem ser encontradas em Robinson (1991) e Searle et al. (2009).

Henderson (1950) motivou a estimação de β e predição de $\mathbf{b}$ através de um conjunto de equações de estimação do tipo mínimos quadrados. As "Equações de Henderson" são dadas a partir da formulação da distribuição conjunta de $\mathbf{b}$ e ε. Sejam g o número de elementos de $\mathbf{b}$ e n a dimensão de $\mathbf{y}$, sua distribuição conjunta é dada por

$$f(\mathbf{b},\varepsilon) = \frac{1}{(2\pi)^{(n+g)/2}} \left| \begin{array}{cc} \mathbf{G} & 0 \\ 0 & \mathbf{R} \end{array} \right|^{-\frac{1}{2}} \cdot \exp\left\{ -\frac{1}{2} \begin{bmatrix} \mathbf{b} \\ \mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{b} \end{bmatrix}' \begin{bmatrix} \mathbf{G}^{-1} & 0 \\ 0 & \mathbf{R}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{b} \end{bmatrix} \right\} \tag{4.2}$$

A ideia de maximizar $f(\mathbf{b},\varepsilon)$ em relação a β e $\mathbf{b}$ significa minimizar a parte exponencial da equação 4.2, ou seja, minimizar

$$Q = \begin{bmatrix} \mathbf{b} \\ \mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{b} \end{bmatrix}' \begin{bmatrix} \mathbf{G}^{-1} & 0 \\ 0 & \mathbf{R}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{b} \end{bmatrix} \tag{4.3}$$

$$= \mathbf{b}'\mathbf{G}^{-1}\mathbf{b} + (\mathbf{y} - \mathbf{X}\beta\mathbf{Z}\mathbf{b})'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{b}) \tag{4.4}$$

onde é aproveitado a independência de $\mathbf{b}$ e ε. Isso nos leva às Equações dos Modelos Mistos de Henderson

$$\frac{\partial Q}{\partial \beta} = 0 \Leftrightarrow \mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\hat{\beta} + \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\hat{\mathbf{b}} = \mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} \tag{4.5}$$

$$\frac{\partial Q}{\partial b} = 0 \Leftrightarrow \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}\hat{\beta} + (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})\hat{\mathbf{b}} = \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y} \tag{4.6}$$

ou na forma matricial

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}) \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y} \end{bmatrix}. \tag{4.7}$$

Para escrevermos a equação (4.7) para o nosso modelo, a matriz $\mathbf{Z}$ das especificações dos efeitos aleatórios deverá ser composta de ambas as matrizes, $\mathbf{Z}_1$ (quando levamos em consideração somente os pacientes como efeitos aleatórios) e $\mathbf{Z}_2$ (quando incluímos o efeito multinível), da seguinte maneira

$$\mathbf{Z} = [\mathbf{Z}_1 \quad \mathbf{Z}_2] = \left[\begin{pmatrix} Z_{1i}^{(1)} & Z_{1i}^{(2)} & \cdots & Z_{1i}^{(q)} \\ Z_{2i}^{(1)} & Z_{2i}^{(2)} & \cdots & Z_{2i}^{(q)} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n_i i}^{(1)} & Z_{n_i i}^{(2)} & \cdots & Z_{n_i i}^{(q)} \end{pmatrix} \quad \begin{pmatrix} Z_{1j|i}^{(1)} & Z_{1j|i}^{(2)} & \cdots & Z_{1j|i}^{(s)} \\ Z_{2j|i}^{(1)} & Z_{2j|i}^{(2)} & \cdots & Z_{2j|i}^{(s)} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n_i j|i}^{(1)} & Z_{n_i j|i}^{(2)} & \cdots & Z_{n_i j|i}^{(s)} \end{pmatrix} \right].$$

A matriz de covariância $\mathbf{G}$ dos efeitos aleatórios será composta pelas matrizes $\mathbf{G}_1$, representando os elementos de $\mathbf{b}_{0i}$, e $\mathbf{G}_2$, os elementos $\mathbf{b}_{0j|i}$, como uma matriz de blocos diagonais

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_1 & 0 \\ 0 & \mathbf{G}_2 \end{bmatrix}.$$

O objetivo é resolver o sistema (4.7) e, para isso, precisamos definir as matrizes $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$, $\mathbf{R}_2$ e $\mathbf{G}$.

Para definirmos algumas das matrizes, é necessário verificar como os dados estão dispostos no banco de dados. E a partir disso, a matriz $\mathbf{X}$ de incidência dos efeitos fixos e o vetor de observações $\mathbf{Y}$ são, respectivamente:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}_{216\times 3}, \mathbf{Y} = \begin{bmatrix} 0.257780 \\ 0.154878 \\ 0.216886 \\ 0.232762 \\ 0.427033 \\ 0.440261 \\ 0.304009 \\ 0.223543 \\ \vdots \\ 0.239643 \end{bmatrix}_{216\times 1}$$

Note que a ordem da matriz $\mathbf{X}$ é (216×3). São 216 linhas pois temos 24 pacientes, e de cada paciente utilizamos 3 dentes e de cada dente foram tiradas 3 medidas. Portanto, $24 \times 3 \times 3 = 216$. E temos 3 tratamentos, daí as 3 colunas.

Para as matrizes $\mathbf{R}_2$, e $\mathbf{G}$ neste trabalho, estimamos σ^2 usando $\sigma^2 = \frac{RSS_\varepsilon}{(n-IJ)}$, onde RSS_ε é a soma de quadrados dos resíduos do modelo (4.1) e $n = IJK$. Para estimarmos $\sigma^2_{b_{j|i}}$, usamos o modelo que resulta da média dos K valores no nível dos dentes, ou seja, $\bar{y}_{ij.} = \beta_0 + \beta_1 \times TRAT_{ij} + b_{0i} + b_{0j|i} + \frac{1}{K}\sum_{k=1}^{K}\epsilon_{ijk}$. Definindo $\varepsilon_{ij} = b_{0j|i} + \frac{1}{K}\sum_{k=1}^{K}\varepsilon_{ijk}$, temos que $var(\varepsilon_{ij}) = \sigma^2_{b_{0j|i}} + \frac{\sigma^2}{K}$, sendo os ε_{ij}'s variáveis aleatórias i.i.d. $N(0, \sigma^2_{b_{0j|i}} + \frac{\sigma^2}{K})$. Dessa forma, podemos escrever o modelo simplificado como

$$\bar{y}_{ij.} = \beta_0 + \beta_1 \times TRAT_{ij} + b_{0i} + \varepsilon_{ij}\,, \tag{4.8}$$

que é útil para estimar a variância residual $\hat{\sigma}^2_{b_{0j|i}} = \frac{RSS_{b_{0j|i}}}{(IJ-I-J+1)} - \frac{\hat{\sigma}^2}{K}$, em que $RSS_{b_{0j|i}}$ é a soma de quadrado dos resíduos de (4.8).

Fazendo a média da variável resposta para cada pacientes temos $\bar{y}_{i..} = \beta_0 + \frac{1}{J}\sum_{j=1}^{J}\beta_1 \times TRAT_{ij} + b_{0i} + \frac{1}{J}\sum_{j=1}^{J}\varepsilon_{ij}$. Definindo $\beta_i' = \beta_0 + \frac{1}{J}\sum_j \alpha_{ij}$ e $\varepsilon_i = b_{0i} + \frac{1}{J}\sum_j \varepsilon_{ij}$ temos

$$\bar{y}_{i..} = \beta_i' + \varepsilon_i \ , \tag{4.9}$$

em que $\varepsilon_i \sim N(0, \sigma_b^2 + \frac{\sigma^2_{b_{0j|i}}}{J} + \frac{\sigma^2}{JK})$. Assim, se RSS_b for a soma de quadrado dos resíduos do modelo (4.9), um estimador não enviesado de σ_b^2 é dado por $\hat{\sigma}_b^2 = \frac{RSS_b}{I-1} - \frac{\hat{\sigma}^2_{b_{0j|i}}}{J} - \frac{\hat{\sigma}^2}{JK}$.

Com isso, temos que

$$\mathbf{R}_2 = [Var(\boldsymbol{\varepsilon}_{ij})] = \sigma^2 \mathrm{I} = 0.03869367^2 \ \mathrm{I} = 0.0014972 \ \mathrm{I}_{216\times 216}$$

Ou seja,

$$\mathbf{R}_2 = \sigma^2 \mathrm{I} = 0.0014972 \times \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}_{216\times 216} .$$

Como nosso modelo apresenta um efeito multinível, precisamos expressá-lo. Como estamos interessados em encontrar os parâmetros para os efeitos aleatórios, a matriz $\mathbf{Z}$ será composta por 96 colunas (24 para os efeitos aleatório dos pacientes indicados por $P_1, P_2, \cdots, P_{24}$ e 72 para os efeitos aleatórios do dente, indicados por $P_1/D_{13}, P_1/D_{21}, \cdots, P_{24}/D_{23}$) e 216 linhas:

$$\mathbf{Z} = \left[\begin{array}{ccccc|cccccc} P_1 & P_2 & P_3 & \cdots & P_{24} & P_1/D_{13} & P_1/D_{21} & P_1/D_{23} & P_2/D_{13} & \cdots & P_{24}/D_{23} \\ 1 & 0 & 0 & \cdots & 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 & 1 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 & 1 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 & 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 1 & 0 & 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 1 & 0 & 0 & 0 & 0 & \cdots & 1 \end{array}\right]_{216\times 96} .$$

A matriz $\mathbf{G}$ levará em consideração a variância do efeito aleatório do intercepto e será uma matriz de ordem 96, pois teremos uma matriz de ordem 24 para os efeitos aleatórios dos pacientes e uma de ordem 72 para os dentes dentro de cada indivíduo.

$$\mathbf{G}_1 = [0.001540523] \times \mathrm{I}_{24\times 24}$$

$$\mathbf{G}_2 = [0.003947203] \times \mathrm{I}_{72\times 72},$$

o que implica

$$\mathbf{G} = \begin{bmatrix} \begin{bmatrix} 0.001540 & 0 & \cdots & 0 \\ 0 & 0.001540 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0.001540 \end{bmatrix}_{24\times 24} & 0 \\ 0 & \begin{bmatrix} 0.003947 & 0 & \cdots & 0 \\ 0 & 0.003947 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0.003947 \end{bmatrix}_{72\times 72} \end{bmatrix}_{96\times 96}$$

Com todas as matrizes implementadas no *software* R, como resultado do sistema (4.7) obtemos:

```
[1]   0.191998  0.013502  0.054398   -0.014034  0.060240  0.004622
[7]  -0.039918 -0.021236 -0.036653    0.053653  0.018053  0.003654
[13]  0.027100 -0.031677 -0.034517   -0.022178  0.002574 -0.013885
[19] -0.035041  0.001480  0.040138   -0.004210 -0.004791 -0.003178
[25]  0.004162  0.026968  0.018673   -0.035748 -0.020093  0.019882
[31] -0.042346  0.026561  0.170135   -0.002020 -0.016731  0.030594
[37]  0.044679 -0.106658 -0.040300   -0.012043 -0.038433 -0.003935
[43] -0.039806 -0.039377 -0.014731    0.049584  0.135353 -0.047465
[49] -0.031177  0.123404 -0.045970   -0.064282  0.144905 -0.071261
[55]  0.094518  0.011605 -0.036688   -0.038188 -0.039938 -0.003038
[61] -0.005458 -0.060839 -0.022144    0.063841 -0.091392 -0.029274
[67]  0.057947  0.043332 -0.094684    0.023792 -0.068370  0.009002
[73] -0.040841 -0.038816 -0.010126   -0.030150  0.001725  0.032217
[79] -0.011426  0.064968  0.049300    0.016467 -0.024251 -0.003003
[85] -0.029380 -0.019753  0.036858   -0.019494  0.006584  0.004767
[91] -0.025994 -0.009323  0.045981    0.028239  0.034396  0.006463
[97]  0.049286 -0.018861  0.017420
```

Podemos notar que os 3 primeiros valores (dentro da caixa) representam os efeitos fixos, os 24 próximos representam os efeitos aleatórios dos pacientes, enquanto os demais representam os efeitos aleatórios de cada dente dentro de cada paciente.

Utilizando o *software* estatístico `R` (R Core Team, 2015) com o uso do pacote `nlme` (PINHEIRO et al., 2016), e sua função `lme()`, a qual utiliza o método de máxima verossimilhança (ML) e máxima verossimilhança restrita (REML) para estimação dos parâmetros, comparamos o resultado obtido e obtivemos os mesmos valores.

4.3 Análise de Resíduos

Para que um modelo estatístico seja válido, sabemos que a aleatoriedade e a imprevisibilidade são componentes cruciais. Por tal fato, precisamos analisar os erros estatísticos do modelo.

Para fins práticos, exibições gráficas dos resíduos podem ser utilizados para detectar discrepâncias no modelo para a resposta média ou para a presença de observações *outliers* que necessitam de uma investigação mais profunda (FROST, 2012).

Como estamos trabalhando com modelos mistos, os resíduos possuem uma matriz de covariância **R**. E tal fato tem implicações importantes para a análise dos gráficos dos resíduos (FITZMAURICE et al., 2012).

Conforme visto na seção anterior, utilizando o *software* `R (Versão 3.2.1)`, com o uso do pacote `nlme` e de sua função `lme()`, a qual utiliza o método de máxima verossimilhança para a estimação dos parâmetros, foi obtido os mesmos resultados e, portanto, iremos utilizar essas informações para a realização da análise de resíduos.

Na Figura 6 observamos o gráfico de dispersão dos resíduos pelos valores ajustados. Nota-se que não há aparentemente alguma forma específica, e que os dados apresentam uma pequena dispersão (de -2 a 2), indicando já uma qualidade de ajuste.

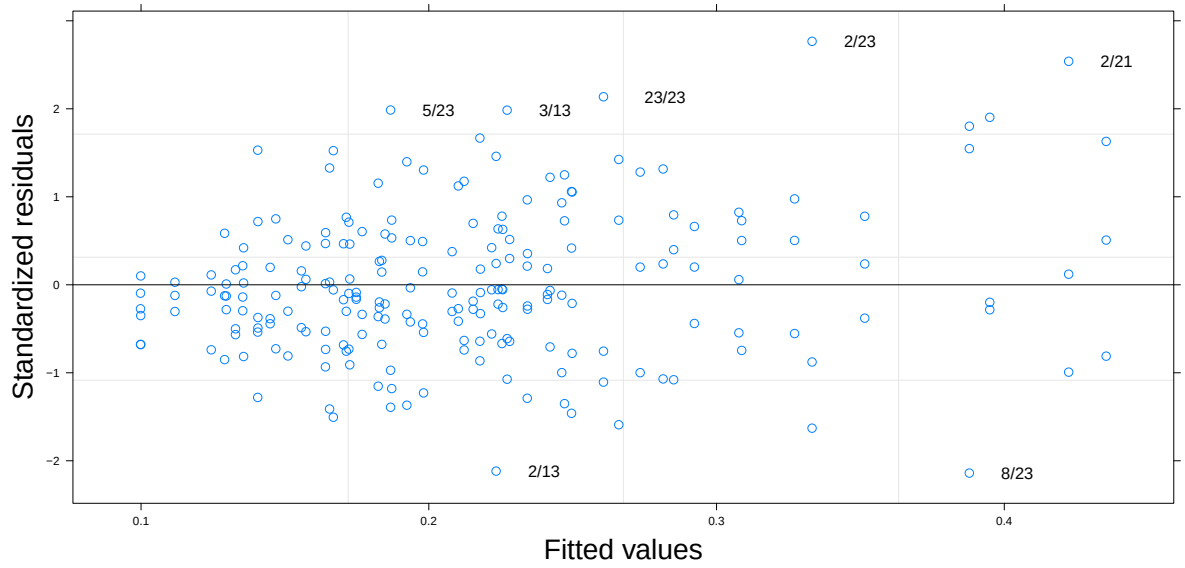


Figura 6 – Gráfico de dispersão dos resíduos por valores ajustados.

Para a Figura 7, podemos notar que os valores preditos estão próximos dos valores observados. Observamos essa informação pela proximidade dos pontos com a reta de regressão, melhorando a percepção de qualidade no ajuste.

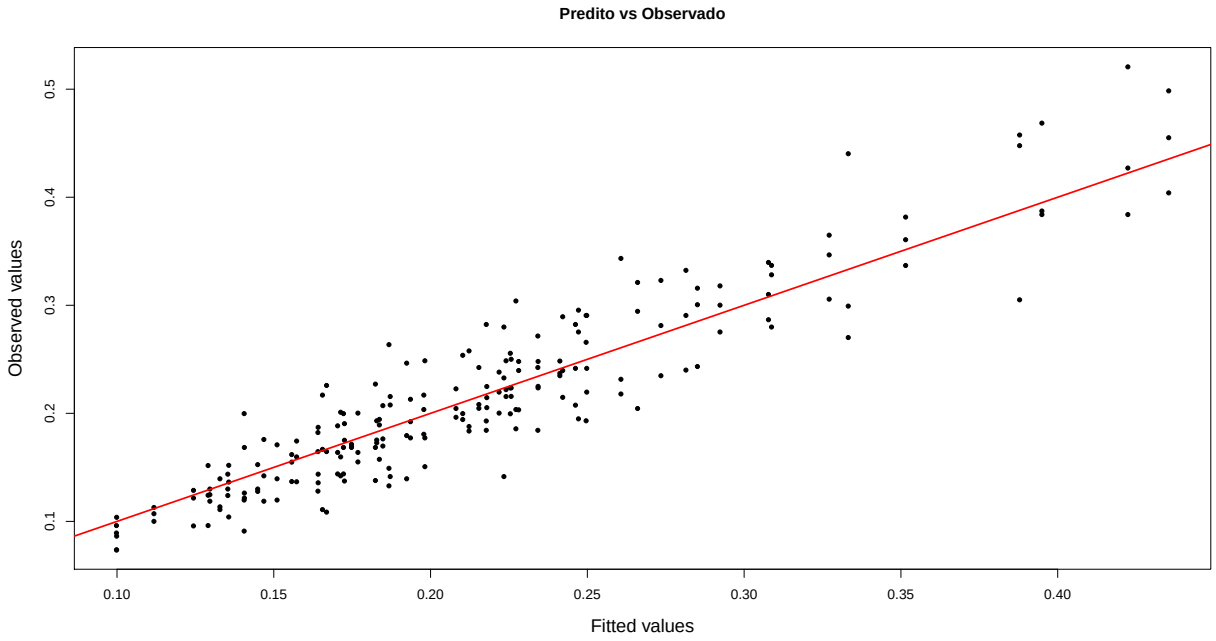


Figura 7 – Valores preditos por valores observados.

Conforme o esperado, podemos verificar na Figura 8 que os dados se encontram dentro do envelope formado pelo intervalo de confiança a 95% para a normalidade dos resíduos. O que indica que o modelo proposto está de acordo com os dados do experimento.

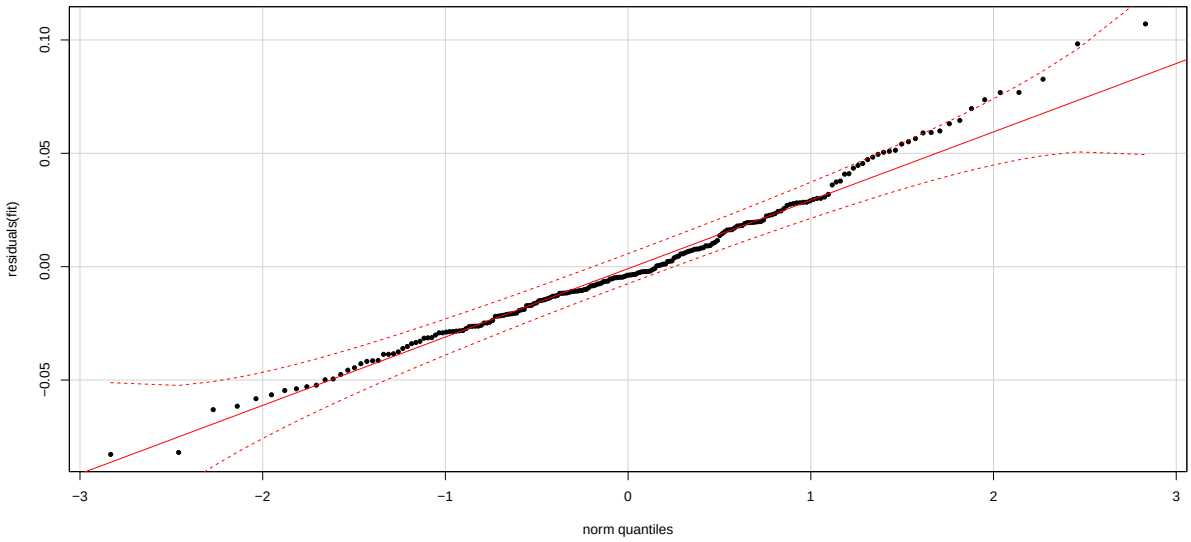


Figura 8 – Valores teóricos por valores amostrais.

A partir disso, temos o seguinte resultado para os estimadores do modelo.

Parâmetro	**Estimativa**	**E.P.**	**Lim. inferior**	**Lim. superior**	**p-valor**
Intercept	0.19199793	0.01579398	0.16077995	0.22321591	<0.0001
tratC	0.01350160	0.01924896	-0.02524455	0.05224774	0.4866
tratH	0.05439754	0.01924896	0.01565139	0.09314369	0.0070

Tabela 4 – Estimativas, Erros Padrão (E.P.), limites inferiores e superiores com 95% de confiança e p-valores para o modelo multinível.

Utilizando o placebo como base para a comparação, temos que o `tratC`, que representa o tratamento com o Colírio não apresentou um p-valor significativo. Ou seja, a hipótese nula de que eles não são diferentes não foi rejeitada. Já o `tratH`, o tratamento com o Hemostop, obteve um p-valor significativo na comparação com o placebo, indicando que há diferença entre os tratamentos (o que já era esperado, pois é a substância usual).

Portanto, de acordo com os dados, há evidências de que o Cloridrato de Nafazolina (Colírio *Legrand*), quando utilizado como agente de retração gengival **não** é capaz de aumentar o afastamento gengival em comparação com a substancia padrão.

4.4 Coeficiente de Correlação Intraclasse

Para um modelo multinível com três níveis com interceptos aleatórios, o coeficiente de correlação intracalsse (CCI) é uma medida que descreve a similaridade (ou homogeneidade) de observações dentro de um mesmo *cluster*. Para cada nível de agrupamento, um

valor de CCI pode ser definido como uma função dos componentes de variância. Como os componentes de variância são nulos ou positivos por definição, os resultados do CCI também são zero ou positivos (WEST et al., 2014). Esta estatística toma valores entre 0 e 1 em que:

- Se o seu valor é próximo de 0, não existe estrutura de agrupamento nos dados;
- Se o seu valor é próximo de 1, pode-se inferir que existe uma estrutura de agrupamento forte.

Conclui-se então que quanto maior a correlação entre os indivíduos, maior é a inadequação do modelo de regressão usual. O descrito traduz uma maior dependência entre indivíduos do mesmo grupo e, por conseguinte, uma maior necessidade de um método de regressão que respeite a estrutura de agregação dos dados.

Para o modelo adotado, consideramos σ_p^2 como a variância dos efeitos aleatórios associados aos pacientes, σ_d^2 a variância dos efeitos aleatórios associados aos dentes aninhados em cada paciente e σ^2 a variância dos resíduos. Assim, temos que o Coeficiente de Correlação Intraclasse para o nível de paciente é dado por:

$$\begin{aligned} CCI_p &= \frac{\sigma_p^2}{\sigma_p^2 + \sigma_d^2 + \sigma^2} \\ &= \mathbf{0.2205497} \end{aligned}$$

Com esse resultado, já considerado alto por alguns autores, temos uma indicação da estrutura de agrupamento dos dados. De maneira similar, o CCI para o nível dos dentes é definido como a proporção da variação total devido a variação aleatória entre pacientes e entre dentes. Esse CCI é alto se existe pouca variação nas respostas observadas dentro do mesmo dente comparado com o total da variação aleatória. Logo, o valor do CCI para o nível dos dentes é dado por:

$$\begin{aligned} CCI_d &= \frac{\sigma_p^2 + \sigma_d^2}{\sigma_p^2 + \sigma_d^2 + \sigma^2} \\ &= \mathbf{0.7856527} \end{aligned}$$

Com esse valor alto para o CCI no nível dos dentes, temos uma conclusão de que o modelo adotado, utilizando uma estrutura de hierarquia, satisfaz as necessidades de dependência dos indivíduos.

4.5 Comparação com um modelo misto de 1 nível

Para uma questão de comparação da qualidade do ajuste das estimativas, consideramos o seguinte modelo linear misto considerando somente o nível dos pacientes

$$Afastamento_{ij} = \beta_0 + b_{0i} + \beta_1 \times TRAT_{ij} + \varepsilon_{ij}$$

onde

- β_0 representa o efeito fixo associado à média;
- β_1 representa o efeito fixo associado ao tratamento em cada dente ($j = 1, 2, 3$);
- b_{0i} representa o efeito aleatório do i-ésimo paciente ($i = 1, \ldots, 24$);
- ε_{ij} representa os erros.

Utilizando o Critério de Informação de Akaike (AIC) e o Critério de Informação Bayesiano (BIC), além do teste da razão de verossimilhança e os gráficos de resíduos, temos que o modelo multinível proposto realmente é mais adequado para o banco de dados, conforme exibido no *output* a seguir

```
> anova(fit.multilevel, fit.mixedmodel)
               Model  df    AIC       BIC   logLik   Test  L.Ratio  p-value
fit.multilevel     1  6 -588.7879 -568.6202 300.3940
fit.mixedmodel     2  5 -488.1282 -471.3217 249.0641 1 vs 2 102.6598  <.0001
```

De acordo com a Tabela 5, temos uma comparação dos valores estimados tanto para o modelo linear misto como para o modelo linear misto multinível para o paciente 20, em relação a seus três dentes.

Paciente/Dente	**Observados**	**Estimados**	
		Multinível	**Misto**
$P_{20,13}$	0.22872	0.22406	0.18410
$P_{20,21}$	0.21935	0.22185	0.23850
$P_{20,23}$	0.16761	0.17132	0.19761

Tabela 5 – Valores Observados, valores estimados dos modelos misto e multinível.

4.6 Conclusão e trabalhos futuros

A partir das informações apresentadas após o ajuste do modelo, podemos afirmar que há evidências de que o novo tratamento proposto com Cloridrato de Nafazolina (Colírio Legrand), quando utilizado como agente de retração gengival, **não** é capaz de aumentar a retração gengival quando comparado com a substância padrão, corroborando com um estudo piloto conduzido anteriormente.

Obteve-se também, informações de uma melhor qualidade de ajuste para um modelo com mais de um nível para análise de dados com esse tipo de estrutra hierárquica. Conforme os valores obtidos para o Coeficiente de Correlação Intraclasse na Seção 4.4, observou-se a necessidade de um elemento que incorpore as dependências dos indivíduos, e conforme as informações da Seção 4.5, verificamos o melhor ajuste dos dados, utilizando medidas de qualidade de ajuste como AIC, BIC e o teste da razão de verossimilhança.

Como trabalhos futuros, além da continuidade de pesquisas juntamente com o programa de Pós graduação em Odontologia, estão alguns pontos:

- Discutir a estimação dos componentes de variação e a estimação da variância dos efeitos aleatórios;
- Estudar a inclusão de efeitos aleatórios para inclinações e não somente para interceptos em modelos multiníveis;
- Analisar o Poder do Teste;
- Verificar adequação de Curva ROC para modelos multiníveis.

Parte II

Modelo Multinível Multinomial na caracterização das ocorrências de câncer de mama por CID no estado do Paraná

Capítulo 5

Introdução

O câncer em geral, como uma neoplasia maligna, pode-se ser atribuído a vários fatores e entre um deles está a localização e o meio em que as pessoas vivem. Questões ambientais e culturais influenciam no desenvolvimento de tal doença conforme (MATOS et al., 2009; WÜNSCH; MONCAU, 2002; KLUTHCOVSKY et al., 2014). Como veremos mais adiante existem diferenças, por exemplo, entre o desenvolvimento do câncer de mama entre mulheres que vivem em áreas mais urbanizadas daquelas que vivem em áreas rurais.

Com base nessas informações, surgiu o interesse de se estudar o câncer de mama no estado do Paraná levando em consideração o espaço geográfico das Regionais de Saúde. Notou-se, tanto durante uma revisão de literatura, quanto durante encontros com pesquisadores do Hospital Universitário de Maringá e do Programa de Pós Graduação em Ciências da Saúde, a ausência e a necessidade de estudos com esse aspecto.

Ao levarmos o espaço geográfico em consideração, podem surgir semelhanças entre os indivíduos residentes em uma mesma região, e essa importante característica deve influenciar na escolha de uma modelagem estatística. Para o estado do Paraná, as Regionais de Saúde estão aninhadas em Macro Regionais conforme será descrito posteriormente, e essa estrutura pode indicar a presença de semelhanças dentro de Regionais de Saúde em uma mesma Macro Regional, mas também indicar diferenças entre Regionais de Saúde pertencentes a Macro Regionais diferentes.

Para resolvermos esse problema de analisar o câncer de mama no estado do Paraná levando em considerações diferenças entre as Regionais de Saúde inseridas dentro de Macro Regionais, no qual utilizaríamos um banco de dados obtidos através do DATASUS, que exibem as informações por internação e por município, surgiu a ideia de utilizarmos modelos multiníveis.

Implícito na noção de que “lugares importam” está o conceito de estrutura multinível,

nos quais existem “efeitos” ou diferenças que operam em diversas camadas ou escalas. Segundo (LONGLEY; BATTY, 1996), é axiomático que métodos quantitativos possuem importância em uma geografia regional. Essa metodologia particular, os modelos multiníveis, se ajustam às necessidades requeridas de uma geografia que se preocupa com a importância do contexto. Como o nome sugere, modelos multiníveis operam em mais de um nível ou escala, tal que um simples modelo pode lidar com situação em nível micro, como no caso, os pacientes portadores de câncer de mama, e em nível macro, com os lugares, no caso as Regionais de Saúde. Mais importante é que, diferenciando os níveis, procedimentos multiníveis permitem que as relações variem de acordo com o contexto.

Alguns artigos e informações sobre a utilização de modelos multiníveis em pesquisas cujo foco é em estabelecer diferenças geográficas estão presentes na literatura, como (JONES, 1993) que apresenta perspectivas sobre a importância de "lugar", (BOYLE; LIPMAN, 1998) apresentam uma análise multinível sobre comportamento de crianças sob uma perspectiva de localização no Canadá, e (DUNCAN et al., 1993) discutem uma análise multinível sobre variações regionais relacionadas a comportamento e saúde no Reino Unido. O mesmo autor, em seu trabalho (DUNCAN et al., 1998), comenta sobre a utilização de modelos multiníveis em pesquisas sobre saúde de uma maneira mais geral.

Outro aspecto importante sobre a pesquisa relacionada ao câncer de mama que não foi encontrado durante a revisão literária, é sobre as características físicas do desenvolvimento do câncer. Segundo a Classificação Internacional de Doenças (CID-10), o câncer de mama é subdividido em 9 tipos de câncer segundo sua topografia, e tal classificação se mostra importante no diagnóstico e no tratamento desse tipo de câncer.

Essa característica do câncer de mama, juntamente com seu comportamento espacial (as divisões das Regionais de Saúde no estado do Paraná) nos levou a consideração de uma modelagem estatística que fosse capaz de absorver essas informações, e pudesse dar respostas importantes tanto estatisticamente quanto na questão da saúde pública.

Para que o modelo incorporasse a questão da hierarquia presente na divisão entre as Regionais de Saúde que estão inseridas nas Macro Regionais, notou-se a necessidade de um modelo multinível. E para também analisarmos a situação do câncer de mama segundo essa classificação surgiu a idéia de utilizarmos como variável resposta essa classificação segundo o CID. Assim, surgiu a ideia de utilizarmos a classificação do câncer de mama de acordo com o CID como a variável resposta, ou seja, cada classificação topográfica representaria uma categoria da variável resposta. Para isso, notou-se a necessidade de utilizarmos um modelo multinomial.

Com esses fatores em consideração, e buscando um modelo que incorporasse tanto a estrutura hierárquica quanto as classificações do câncer de mama, optou-se por utilizar um modelo multinível multinomial, cujo principal objetivo é verificar a frequência de ocorrência de determinado tipo de câncer de mama segundo a classificação do CID em

mulheres no estado do Paraná, levando em consideração sua localidade de residência, assim como outros fatores.

Capítulo 6

Modelo Multinível Multinomial

6.1 Distribuição Multinomial

É comum na área da saúde serem utilizadas, como variáveis respostas, dados que não necessariamente são resultados de mensuração, e sim que possuem uma escala de medidas constituída de um conjunto de categorias. Por exemplo para medir respostas sobre a sobrevivência ou não de um paciente a uma cirurgia, ou ainda a gravidade de uma lesão ou doença como (*"leve"*, *"moderada"* ou *"grave"*) e estágio de uma doença (*"inicial"*, *"intermediário"* ou *"avançado"*) (AGRESTI, 1996).

Variáveis categóricas com escalas que possuem ordem são denominadas variáveis *ordinais* e as variáveis que não possuem escalas de ordem são denominadas variáveis *nominais* e possuem exemplos como tipos de câncer (*"mama", "útero", "próstata", "pulmão"*), tipo de sangue (*"O", "A", "B", "AB"*) entre outros. Para esse tipo de variável, a ordem da listagem das categorias é irrelevante, e a análise estatística não deve levar em consideração tal característica.

Quando os ensaios são independentes com a mesma probabilidade para cada categoria, entre um conjunto definido de categorias, a distribuição de contagem nos vários casos é a `multinomial`.

Segundo (MCCULLAGH; NELDER, 1989), a distribuição multinomial é, de muitas maneiras, a distribuição mais natural que possa ocorrer com esse tipo de variável resposta.

Suponha que indivíduos de uma população possuam somente um entre k atributos $A_1, A_2, \cdots, A_k$. Se a população for grande o suficiente, e uma simples amostra de tamanho n é tomada, quantos indivíduos serão observados que possuam o atributo A_j?

A resposta é dada pela distribuição multinomial, conforme

$$P(Y_1 = y_1, \cdots, Y_k = y_k | n, \pi) = \begin{pmatrix} n \\ y \end{pmatrix} \cdot \pi_1^{y_1} \cdots \pi_k^{y_k}, \tag{6.1}$$

tal que $\pi_1, \cdots, \pi_k$ são as frequências atribuídas à população e que

$$\begin{pmatrix} n \\ y \end{pmatrix} = \frac{n!}{y_1! \cdots y_k!}. \tag{6.2}$$

Note que se $k = 2$, temos a distribuição binomial.

A função geradora de momentos da distribuição multinomial, $M(n, \pi)$ é

$$M_Y(t) = E\left[\exp\left(\sum t_j Y_j\right)\right] = \left\{\sum \pi_j \exp(t_j)\right\}^n, \tag{6.3}$$

e a partir disso, temos que a distribuição multinomial possui média, variância e covariância dados por

- $E(Y_i) = n\pi_j$;
- $Var(Y_j) = n\pi_j(1 - \pi_j)$;
- $Cov(Y_i, Y_j) = -n\pi_i\pi_j$.

6.2 Modelos Lineares Generalizados - MLG

Uma das partes mais importantes de toda pesquisa em modelagem estatística envolve a procura de um modelo que seja o mais simples possível e que descreva bem os dados observados que surgem em diversas áreas do conhecimento e de diversas formas. Nelder e Wedderburn (1972) mostraram que uma série de técnicas estatísticas comumente estudas separadamente podem ser formuladas de uma maneira unificada, como uma classe de modelos de regressão. A essa teoria unificadora de modelagem estatística, uma extensão dos modelos clássicos de regressão, deram o nome de *modelos lineares generalizados (MLG)* (CORDEIRO; DEMÉTRIO, 2008).

Avanços na teoria estatística e computacional nos permitiram usar metodologias análogas às desenvolvidas para modelos lineares, como utilizar variáveis respostas não normalmente distribuídas, e que as relações entre variáveis respostas e explicativas possuírem uma forma não linear.

Um desses avanços foi o reconhecimento de que várias "boas" propriedades da distribuição normal são compartilhadas por uma ampla classe de distribuições chamada *família exponencial de distribuições* (DOBSON; BARNETT, 2008).

Muitas distribuições conhecidas pertencem à família exponencial. As distribuições normal, binomial, binomial negativa, gama, Poisson, normal inversa, multinomial, beta, logarítmica, entre outras.

6.2.1 Família Exponencial Multiparamétrica

A família exponencial multiparamétrica de dimensão k é caracterizada por uma função (de probabilidade ou densidade) da forma

$$f(y;\boldsymbol{\theta}) = h(y)\exp\left[\sum_{i=1}^{k}\eta(\boldsymbol{\theta})t_i(y) - b(\boldsymbol{\theta})\right], \tag{6.4}$$

em que $\boldsymbol{\theta}$ é um vetor de parâmetros, usualmente de dimensão k, e as funções $\eta_i(\boldsymbol{\theta}), b(\boldsymbol{\theta}), t_i(y)$ e $h(y)$ assumem valores em subconjuntos dos reais. Pelo Teorema da Fatoração, o vetor $\mathbf{T} = [T_1(Y), \cdots, T_k(Y)]^T$ é suficiente para o vetor de parâmetros $\boldsymbol{\theta}$. Quando $\eta_i(\boldsymbol{\theta}) = \theta_i$, $i = 1, \cdots, k$ obtém-se de 6.4 a família exponencial na forma canônica com parâmetros canônicos $\theta_1, \cdots, \theta_k$ e estatísticas canônicas $T_1(Y), \cdots, T_k(Y)$, escrito sob a forma

$$f(y;\boldsymbol{\theta}) = h(y)\exp\left[\sum_{i=1}^{k}\theta_i t_i(y) - b(\boldsymbol{\theta})\right]. \tag{6.5}$$

Como é o objetivo do estudo, estudaremos a distribuição multinomial como parte da família exponencial multiparamétrica. Portanto, seja a distribuição multinomial com função de probabilidade

$$f(y;\pi) = \frac{n!}{y_1!\cdots y_k!}\cdot \pi_1^{y_1}\cdots\pi_k^{y_k} \tag{6.6}$$

em que $\sum_{i=1}^{k} y_i = n$ e $\sum_{i=1}^{k}\pi_i = 1$.

Essa distribuição pertence à família exponencial multivariada canônica com parâmetro canônico $\boldsymbol{\theta} = (\log\pi_1, \cdots, \log\pi_k)^T$, estatística canônica $\mathbf{T} = (Y_1, \cdots, Y_k)^T$ e $\mathbf{t} = (t_1, \cdots, t_{k-1})^T$, ambos vetores de dimensão $(k-1)$, resultando na família exponencial multiparamétrica de dimensão $(k-1)$

$$f(y;\boldsymbol{\theta}) = \frac{n!}{y_1!, \cdots, y_k!}\exp\left[\sum_{i=1}^{k-1}\theta_i y_i - b(\boldsymbol{\theta})\right], \tag{6.7}$$

com $\theta_i = \log\left(\frac{\pi_i}{\pi_k}\right), i = 1, \cdots, k-1$ e $b(\boldsymbol{\theta}) = n\log\left(1 + \sum_{i=1}^{k-1} e^{\theta_i}\right)$ (CORDEIRO; DEMÉTRIO, 2008).

Pode-se demonstrar que os dois primeiros momentos da estatística suficiente $\mathbf{T} = [T_1(Y), \cdots, T_k(Y)]^T$ na família exponencial canônica são dados por

$$E(\mathbf{T}) = \frac{\partial b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \text{Cov} = \frac{\partial^2 b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}. \tag{6.8}$$

Para o modelo multinomial 6.7, utilizando as equações 6.8, temos que o valor esperado e a variância são dados por

$$\begin{aligned} E(Y_i|\eta) = n\frac{\partial}{\partial \theta_i} \log\left(1 + \sum_{i=1}^{k-1} e^{\theta_i}\right) &= \frac{ne^{\theta_i}}{1 + \sum_{i=1}^{k-1} e^{\theta_i}} \\ &= \frac{n \cdot \frac{\pi_i}{\pi_k}}{1 + \sum_{i=1}^{k-1} \frac{\pi_i}{\pi_k}} \\ &= n\pi_i. \end{aligned} \tag{6.9}$$

e para $i \neq j$

$$\begin{aligned} Cov(Y_i, Y_j|\eta) = n\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log\left(1 + \sum_{i=1}^{k-1} e^{\theta_i}\right) &= \frac{-ne^{\theta_i}e^{\theta_j}}{\left(1 + \sum_{i=1}^{k-1} e^{\theta_i}\right)^2} \\ &= -n\pi_i\pi_j \end{aligned} \tag{6.10}$$

e para $i = j$

$$\begin{aligned} Var(Y_i|\eta) = n \cdot \frac{\partial^2}{\partial \theta_i^2} \log\left(1 + \sum_{i=1}^{k-1} e^{\theta_i}\right) \\ = n\pi_i(1 - \pi_i). \end{aligned} \tag{6.11}$$

Todos os modelos lineares generalizados possuem três componentes:

- O ***componente aleatório*** que identifica a variável resposta e seleciona uma distribuição de probabilidade, que faz parte da família de distribuições descritas anteriormente;
- A ***função de ligação*** que especifica a função que relaciona os componentes sistemáticos e aleatórios. Ou seja, vincula a média ao preditor linear, isto é,

$$\eta_i = g(\mu_i), \tag{6.12}$$

sendo $g(.)$ uma função monótona e diferenciável.

Entre alguns exemplos de funções de ligação estão a função identidade, logit, log, probit, log-log. Assim como para a distribuição binomial, a função de ligação canônica

para a distribuição multinomial é a logit. Essa função de ligação é a mais popular nas ciências da saúde, parcialmente porque seus coeficientes podem ser interpretados em termos de *odds ratio*, ou seja, devido a existência de um maior apelo na área da saúde para obter informações de maneiras relevantes, sua utilização possui mais eficiência de interpretação dos resultados.

- O ***componente sistemático*** que especifica as variáveis explicativas. Elas entram na forma de uma soma linear de seus efeitos

$$\eta_i = \sum_{r=1}^{p} x_{ir}\beta_j = \mathbf{x}_i^T\boldsymbol{\beta}$$
$$\text{ou} \quad \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}, \tag{6.13}$$

sendo $\mathbf{X} = (\mathbf{x_1}, \cdots, \mathbf{x_n})^T$ a matriz do modelo, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ o vetor de parâmetros e $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_n)^T$ o preditor linear;

A extensão que leva de uma regressão binomial para uma regressão multinomial é dada por um modelo que se baseia em varias equações do modelo, dependendo da quantidade de classes existentes na variável resposta. Para uma variável resposta nominal com k categorias, o modelo multinomial estima $k-1$ equações `logit`, pois utiliza uma das categorias como referência/*baseline* para as demais.

Como descrito em (JR; LEMESHOW, 2004), utilizando um exemplo cuja variável resposta assume apenas três níveis, digamos 0, 1 e 2, o modelo logístico terá duas funções `logit`: a razão entre Y=1 e Y=0 e a razão Y=2 e Y=0. Nesse caso, o nível Y=0 foi utilizado como *baseline*.

$$\eta_1(x) = \log\left[\frac{P(Y=1)}{P(Y=0)}\right] = \beta_{10} + \beta_{11}x_1 + \cdots + \beta_{1p}x_p \tag{6.14}$$

$$\eta_2(x) = \log\left[\frac{P(Y=2)}{P(Y=0)}\right] = \beta_{20} + \beta_{21}x_1 + \cdots + \beta_{2p}x_p \tag{6.15}$$

A partir das funções $\eta_i(x)$, é possível calcular as probabilidades condicionais de ocorrência de cada categoria da variável resposta Y dado um vetor de observações X.

A primeira equação exibe a probabilidade condicional para a categoria utilizada como referência e é resultado da condição das somas das probabilidades serem iguais a 1, e as demais exibem as demais probabilidades

$$P(Y=0|x) = \frac{1}{1 + e^{\eta_1(x)} + e^{\eta_2(x)}}, \tag{6.16}$$

$$P(Y=1|x) = \frac{e^{\eta_1(x)}}{1+e^{\eta_1(x)}+e^{\eta_2(x)}}, \tag{6.17}$$

$$P(Y=2|x) = \frac{e^{\eta_2(x)}}{1+e^{\eta_1(x)}+e^{\eta_2(x)}}. \tag{6.18}$$

A generalização do modelo logístico binomial para variáveis respostas com k níveis é direta, assim, na regressão logística multinomial, a probabilidade de uma dada observação X pertencer a uma das classes y_i é dada por

$$P(Y=y_i|x) = \frac{\exp\{\eta_i(x)\}}{1+\sum_{j=1}^{k-1}\exp\{\eta_j(x)\}}, \tag{6.19}$$

com $i = 1, 2, \cdots, k-1$. Assim, a função `logit`, assumindo o nível y_k como base, é dada por

$$\eta(x) = \log\left[\frac{P(Y=y_i|x)}{P(Y=y_k|x)}\right] = \beta_{i0} + \beta_{i1}x_1 + \cdots + \beta_{ip}x_p. \tag{6.20}$$

Qualquer uma das categorias podem servir como referência com a decisão baseada na pesquisa e, a partir disso, é possível comparar resultados de duas categorias não consideradas como *baseline* usando a seguinte equação:

$$\log\left(\frac{P(Y=y_i)}{P(Y=y_m)}\right) = \log\left(\frac{P(Y=y_i)}{P(Y=y_j)}\right) - \log\left(\frac{P(Y=y_m)}{P(Y=y_j)}\right) \tag{6.21}$$

onde *baseline* é a categoria y_j e o objetivo é comparar as categorias y_i e y_m

6.3 Modelos Lineares Generalizados Mistos - MLGM

Segundo McCulloch e Neuhaus (2001), a ideia básica por trás dos modelos lineares generalizados mistos é conceitualmente direta: incorporar efeitos aleatórios na proporção linear de um modelo linear generalizado. Essa simples mudança nos permite acomodar uma correlação no contexto de uma ampla classe de modelos para dados não normalmente distribuídos. Ou seja, é uma maneira conveniente de construir distribuições multivariadas para dados não normais que possam acomodar alguma flexibilidade na estrutura de associação, assim como um rico conjunto de variáveis preditoras. (PINHEIRO; BATES, 2006) afirmam que os modelos multiníveis lineares generalizados (MMLG) são uma classe estendida dos MLGM denotando cada nível como nível hierárquico de efeitos aleatórios. Isto é, um MLGM é equivalente a um MMLG com 1 nível.

Essa linha de pensamento sugere que os efeitos aleatórios devem ser incorporados da mesma maneira e na mesma porção do modelo que os efeitos fixos. Nosso modelo

linear básico possuía média $E[y] = \mathbf{X}\boldsymbol{\beta}$ conforme 6.13. Nós incorporamos efeitos aleatórios ampliando o modelo como $E[y|\mathbf{b}] = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zb}$.

Isso sugere uma extensão direta dos modelos lineares generalizados descritos na seção anterior, acrescentando o efeito aleatório na forma $\mathbf{Zb}$ ao preditor linear $\mathbf{X}\boldsymbol{\beta}$. Isso alcançará os dois principais objetivos: incorporar correlação e permitir uma inferência mais ampla. Porém, a natureza não linear do modelo cria complicações (MCCULLOCH; SEARLE, 2001).

Os MLGM são condicionados aos efeitos aleatórios $\mathbf{b}_i$, assumindo que os elementos Y_{ij} de $\mathbf{Y}_i$ (vetor da variável resposta) sejam independentes, seguindo um modelo linear generalizado, mas com preditor linear estendido com parâmetros de regressão específicos para o sujeito $\mathbf{b}_i$ (VERBEKE; MOLENBERGHS, 2005).

Com isso, a média μ_{ij} é modelada através de um preditor linear contendo parâmetros de regressão fixos β assim como parâmetros específicos para o sujeito $\mathbf{b}_i$, isto é,

$$E(Y|\mathbf{b}) = g^{-1}(\eta_{ij}^{\mathbf{b}}) \text{ com } \eta_{ij}^{\mathbf{b}} = x'_{ij}\boldsymbol{\beta} + z'_{ij}\mathbf{b}_i \tag{6.22}$$

para uma função de ligação conhecida $g(\mu_{ij}^{\mathbf{b}}) = \eta_{ij}^{\mathbf{b}}$, tal que $\mu_{ij}^{b} = g^{-1}(\eta_{ij}^{\mathbf{b}})$ e por $\mathbf{x}_{ij}$ e $\mathbf{z}_{ij}$ dois vetores contendo covariáveis conhecidas.

Denotando o vetor de observações por Y, e as matrizes de design x'_{ij} e z'_{ij} por **X** e **Z**, a média condicional satisfaz

$$\mu^{\mathbf{b}} = g^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Zb}). \tag{6.23}$$

Logo, a esperança condicional se torna

$$\mu_{ij}^{b} = g^{-1}(\eta_{ij}^{\mathbf{b}}) = \psi'(\theta_{ij}) \equiv \frac{\partial\psi(\theta_{ij})}{\partial\theta_{ij}}, \tag{6.24}$$

e a variância condicional

$$Var(Y_{ij}|\eta_{ij}^{\mathbf{b}}) = \psi''(\theta_{ij}) = \phi V(\mu_{ij}^{\mathbf{b}}), \tag{6.25}$$

onde $V(\mu_{ij}^{\mathbf{b}})$ é a função de variância, e ϕ é o parâmetro de dispersão, que para a distribuição multinomial consideramos igual a 1.

A função de variância é utilizada para modelar variabilidade não sistemática, e modela a relação entre a variância de y e μ, de acordo com

$$\begin{aligned} V(\mu_{ij}) &= V(g^{-1}(\mathbf{X_{ij}}\boldsymbol{\beta})) \\ &= V(g^{-1}(\eta_{ij}^{\mathbf{b}})) \end{aligned} \tag{6.26}$$

e, de acordo com 6.11, temos que a variância condicional para a distribuição multinomial é dada por

$$Var(Y_{ij}|\eta_{ij}^{\mathbf{b}}) = n\pi_i(1-\pi_i). \tag{6.27}$$

Consequentemente, temos que

$$\begin{aligned} Var(Y_{ij}|\eta_{ij}^{\mathbf{b}}) &= n \cdot \frac{e^{\eta_{ij}^{\mathbf{b}}}}{1+\sum e^{\eta_{ij}^{\mathbf{b}}}} \cdot \left(1 - \frac{e^{\eta_{ij}^{\mathbf{b}}}}{1+\sum e^{\eta_{ij}^{\mathbf{b}}}}\right) \\ &= n \cdot \frac{e^{\eta_{ij}^{\mathbf{b}}}}{1+\sum e^{\eta_{ij}^{\mathbf{b}}}} - n \cdot \left(\frac{e^{\eta_{ij}^{\mathbf{b}}}}{1+\sum e^{\eta_{ij}^{\mathbf{b}}}}\right)^2. \end{aligned} \tag{6.28}$$

Segundo Breslow e Clayton (1993), essa formulação engloba situações onde os efeitos aleatórios estão aninhados entre os sujeitos, e quando não estão.

A média de Y pode ser obtida pelo artifício de esperanças iteradas, e como $g(\mu) = \eta$, temos que $\mu = g^{-1}(\eta)$, ou seja, $\mu^{\mathbf{b}} = g^{-1}(X_i'\beta + Z_i'\mathbf{b})$. Logo,

$$\begin{aligned} E[y_i] &= E[E[y_i|\mathbf{b}]] \\ &= E[\mu_i^{\mathbf{b}}] \\ &= E[g^{-1}(X_i'\beta + Z_i'\mathbf{b})]. \end{aligned}$$

Isso não pode ser simplificado, em geral, devido à função não linear g^{-1}.

Para ilustrar em uma função particular $g(.)$, suponhamos que tenhamos uma ligação logarítmica tal que $g(\mu) = \log(\mu)$ e $g^{-1}(x) = \exp\{x\}$. Assim temos,

$$\begin{aligned} E[y_i] &= E[\exp\{X_i'\beta + Z_i'\mathbf{b}\}] \\ &= \exp\{X_i'\beta\} \cdot E[\exp\{Z_i'\mathbf{b}\}] \\ &= \exp\{X_i'\beta\} \cdot M_{\mathbf{b}}(Z_i), \end{aligned}$$

onde $M_{\mathbf{b}}(Z_i)$ é a função geradora de momentos de $\mathbf{b}$ avaliado em Z_i.

Suponha a seguir que cada coluna de Z tenha uma simples entrada de 1's com o restante sendo 0. Assim, $M_{\mathbf{b}}(Z_i) = \exp\{\frac{\sigma_b^2}{2}\}$ e

$$E[y_i] = \exp\{X_i'\beta\} \cdot \exp\{\frac{\sigma_b^2}{2}\} \tag{6.29}$$

ou

$$\log E[y_i] = X_i'\beta + \frac{\sigma_b^2}{2}. \tag{6.30}$$

Esse modelo é completado assumindo que, condicionalmente nos efeitos específicos do sujeito $\mathbf{b}_i$, as respostas Y_{ij} são independentes e que $\mathbf{b}_i$ tenha uma distribuição normal multivariada com média $\mathbf{0}$ e matriz de covariância $\mathbf{G} = \mathbf{G}(\boldsymbol{\theta})$, dependendo de um vetor $\boldsymbol{\theta}$ desconhecido de componentes de variância (BRESLOW; CLAYTON, 1993).

Como estimamos os efeitos fixos diretamente, os componentes de efeitos aleatórios são modelados como desvios do efeito fixo, portanto possuem média zero. A matriz de covariância $\mathbf{G}$, se o modelo possuir um intercepto aleatório e uma inclinação aleatória, seria dado como

$$\mathbf{G} = \begin{bmatrix} \Sigma^2_{int} & \Sigma^2_{int,incl} \\ \Sigma^2_{int,incl} & \Sigma^2_{incl} \end{bmatrix} \tag{6.31}$$

onde Σ^2_{int} representa a variância do intercepto, Σ^2_{incl}, a variância da inclinação e $\Sigma^2_{int,incl}$ a variância da interação.

Para o caso do modelo possuir somente um efeito aleatório para o intercepto, a matrix $\mathbf{G}$ é uma matriz 1×1, com a variância do intercepto aleatório.

Assim, o elemento final do modelo é a matriz de covariância dos resíduos ε, ou a matriz de covariância condicional, denotada por $\mathbf{R}$. A estrutura de covariância residual mais comum é dado por

$$\mathbf{R} = \mathbf{I}\Sigma^2_{\varepsilon}, \tag{6.32}$$

onde $\mathbf{I}$ é a matriz identidade, e Σ^2_{ε} é a variância residual.

A relação entre o preditor linear $\boldsymbol{\eta}$ e o vetor de observações $\boldsymbol{y}$ é modelado por

$$y|\mathbf{b} \sim (g^{-1}(\boldsymbol{\eta}), \mathbf{R}), \tag{6.33}$$

ou seja, a distribuição condicional de y dado $\mathbf{b}$ possui média $g^{-1}(\boldsymbol{\eta})$ e variância $\mathbf{R}$. A distribuição condicional de $y|\mathbf{b}$ é usualmente referida como distribuição dos erros.

Por último, de acordo com Fikret (2011), a matriz de covariância das observações é dada por

$$V(Y) = \mathbf{ZGZ^T} + \mathbf{A}^{1/2}\mathbf{RA}^{1/2}, \tag{6.34}$$

onde a matriz **A** é uma matriz diagonal que contém as funções de variância do modelo conforme 6.26.

A verossimilhança de um modelo linear generalizado misto envolve uma integral, a qual geralmente não pode ser resolvida explicitamente. Segundo Verbeke e Molenberghs (2005), a verossimilhança para o indivíduo i é da forma

$$f(y_i|\boldsymbol{\beta}, \mathbf{G}, \phi) = \int \prod_{i=1}^{n_i} f_{ij}(y_{ij}|\mathbf{b}, \boldsymbol{\beta}, \phi) f\mathbf{b}_i|\mathbf{G})d\mathbf{b}_i, \tag{6.35}$$

da qual segue a verossimilhança para $\boldsymbol{\beta}, \mathbf{G}$ e ϕ,

$$\begin{aligned} L(\boldsymbol{\beta}, \mathbf{G}, \phi) &= \prod_{i=1}^{N} f_i(y_i|\boldsymbol{\beta}, \mathbf{G}, \phi) \\ &= \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\mathbf{b}, \boldsymbol{\beta}, \phi) f\mathbf{b}_i|\mathbf{G})d\mathbf{b}_i, \end{aligned} \tag{6.36}$$

onde $\mathbf{G}$ representa a matriz de covariância dos efeitos aleatórios $\mathbf{b}$.

6.4 Modelo Multinível Multinomial

Segundo Hartzel e Agresti (2001) uma grande quantidade de artigos tem desenvolvido o MLGM para distribuições *binomial* e *Poisson*, porém, pouca ênfase em desenvolvimento para respostas multinomiais, com a maioria das pesquisas focadas em modelos ordinais com função de ligação `logit` e `probit` para probabilidades acumuladas. Harville e Mee (1984) propuseram um modelo de efeitos aleatórios de probabilidade acumulada com função de ligação `probit` que utilizava aproximações de séries de Taylor para integrais intratáveis provenientes da função de Verossimilhança. Jansen (1990), Ezzet e Whitehead (1991) propuseram modelos com interceptos aleatórios `probit` e `logit`, respectivamente, e utilizaram técnicas de quadratura para aproximação da função de verossimilhança.

Ainda segundo Hartzel e Agresti (2001), modelagem de respostas nominais com efeitos aleatórios não têm recebido muita atenção, e sua maior concentração está na área de econometria e psicometria.

6.4.1 Modelo Marginal

De acordo com Kuss e McLerran (2007), para especificar um modelo multinomial logístico para dados correlacionados como um modelo marginal, precisamos reorganizar o vetor resposta.

Escrevemos, então, Y_{ij} como um vetor Y_{ij}^* de tamanho $((C-1)\times 1)$ de um indicador de variáveis binárias Y_{ijc}^* tal que $Y_{ij} = 2, \cdots, C$ resulta em $Y_{ij}^* = 1$ na coluna c e 0 nas demais colunas. Ou seja, nós rescrevemos cada resposta categórica como um vetor de resposta *dummy* tal que

$$Y^*_{ijc} = \begin{cases} 1, \text{ se } Y^*_{ij} = c, c = 2, \cdots, C. \\ 0, \text{ caso contrário.} \end{cases}$$

No caso de $Y_{ij} = 1$ (categoria de referência), $\mathbf{Y}^*_{ij} = 0$ em todas as $C - 1$ colunas.

Essa reorganização do vetor resposta pode ser interpretado como transformando o modelo multinomial em um modelo binário multivariado, no qual Hartzel e Agresti (2001) utiliza o termo MGLMM (Multivariate Generalized Linear Mixed Model).

Assim, seja $\mathbf{Y}^*_i = (Y^*_{i1}, Y^*_{i1}, \cdots, Y^*_{in_i})'$ o vetor resposta de tamanho $(n_i(C-1) \times 1)$ para o i-ésimo *cluster* com esperança π^*_i e matriz de covariância V^*_c.

Essa matriz de covariância V^*_c é uma matriz diagonal de dois blocos, onde o bloco $(C-1) \times (C-1)$ para (c, c') no "bloco de dentro" da diagonal principal de V^*_c é uma matriz de covariância multinomial para a j-ésima observação no i-ésimo *cluster*, e os demais elementos no "bloco exterior" especifica a covariância entre duas observações diferentes (j, j') no i-ésimo *cluster*.

De maneira mais formal, temos que a matriz de covariância multinomial é

$$V^*_i = \text{cov}(Y^*_{ijc}, Y^*_{ij'c'}) = \begin{cases} \pi^*_{ijc}(1 - \pi_{ijc}), \text{ se } j = j', c = c'; \\ -\pi^*_{ijc} \cdot \pi^*_{ijc'}, \text{ se } j = j', c \neq c'; \\ \frac{\text{corr}(Y^*_{ijc}, Y^*_{ij'c'})}{\left[\pi^*_{ijc}\left(1-\pi^*_{ij'c'}\right) \cdot \pi^*_{ij'c'}\left(1-\pi^*_{ij'c'}\right)\right]^{1/2}}, \text{ se } j \neq j', \end{cases}$$

onde as primeiras duas linhas correspondem ao bloco "interior" de V^*_c, a terceira linha o bloco "exterior", e $\pi^*_{ijc} = E(Y^*_{ijc} = 1)$.

Deve-se notar que a terceira linha não constitui uma definição circular. Ao invés disso, corr$(Y^*_{ijc}, Y^*_{ij'c'})$ deve ser indicada como um padrão de correlação que funcione na análise.

Assim, a equação do modelo é

$$\log\left(\frac{\pi^*_{ic}}{1 - \pi^*_{ic}}\right) = \theta^*_c + X'_{ij}\beta^*_c, \tag{6.37}$$

com $c = 2, \cdots, C$, e onde π^*_{ic} denota a esperança de todos os elementos de $\mathbf{Y}^*_i$ pertencentes à categoria c da resposta. Note que não há referência a um efeito aleatório na equação do modelo.

Várias escolhas são possíveis para uma forma funcional para a matriz de covariância V^*_c, desde a mais simples suposição de independência entre os *clusters*, (corr$(Y^*_{ijc}, Y^*_{ij'c'}) \equiv 0$ se $j \neq j'$), a uma forma mais complexa onde todos os $\left((C-1)\frac{(C-1)}{2}n_i\frac{(n_i-1)}{2}\right)$ parâmetros variam.

Escolher V_c^* o mais próximo possível da verdadeira matriz de correlação geralmente resulta em ganho de eficiência (KUSS; MCLERRAN, 2007).

6.4.2 Especificação do modelo

Sejam i as unidades de nível 2 (clusters) e j as unidades de nível 1 (unidades observadas). Assumimos que há $i = 1, 2, \cdots N$ unidades de nível 2 e $j = 1, 2, \cdots, n_i$ unidades de nível 1 aninhadas em cada unidade de nível 2.

Sejam y_{ij} o valor da variável nominal associada com a unidade i de nível 2 e j de nível 1, e considerando os valores correspondentes às categorias não ordenadas da variável resposta, assumimos que as C categorias respostas são dadas por $c = 1, 2, \cdots, C$.

Ao adicionarmos efeitos aleatórios ao modelo de regressão logístico usual, temos que a probabilidade que $y_{ij} = c$ (a resposta ocorra na categoria c) para uma dada unidade de nível 2 i, condicional ao efeito aleatório $\mathbf{b}$ é dado por

$$p_{ijc} = P(y_{ij} = c|\mathbf{b}) = \frac{\exp(\eta_{ijc})}{1 + \sum_{h=1}^{C-1} \exp(\eta_{ijh})}, \tag{6.38}$$

para $c = 1, 2, 3, \cdots, C-1$, onde $\eta_{ijc} = x'_{ij} \cdot \boldsymbol{\beta}_c + z'_{ij} \cdot \mathbf{b}_{ic}$ é o preditor linear dependendo da categoria c. Com x_{ij} o vetor $s \times 1$ de covariáveis, z_{ij} o vetor para os r efeitos aleatórios, ambos sendo para a j-ésima unidade de nível 1 aninhada na unidade i de nível 2, $\boldsymbol{\beta}_c$ um vetor $s \times 1$ de parâmetros fixos desconhecidos e $\mathbf{b}_{ic}$ um vetor $r \times 1$ de efeitos aleatórios desconhecidos para a unidade i de nível 2. A distribuição dos efeitos aleatórios é assumido ser normal multivariado com vetor média 0 e matriz de covariância $\mathbf{G}$ (HEDEKER, 2005).

A partir da Equação 6.38, podemos chegar na equação do modelo logístico multinomial multinível com interceptos aleatórios.

$$p_{ijc} = \frac{\exp(\eta_{ijc})}{1 + \sum_{h=1}^{C-1} \exp(\eta_{ijh})}$$

$$\Rightarrow p_{ijc} = \left[\frac{1}{1 + \sum_{h=1}^{C-1} \exp(\eta_{ijh})} \right] \cdot \exp(\eta_{ijc}) \text{ mas, por 6.16,}$$

$$\Rightarrow p_{ijc} = p_{ij1} \cdot \exp(\eta_{ijc})$$

$$\Rightarrow \frac{p_{ijc}}{p_{ij1}} = \exp(\eta_{ijc}) \text{ aplicando a função logarítmica,}$$

$$\Rightarrow \log\left(\frac{p_{ijc}}{p_{ij1}}\right) = \eta_{ijc}$$

$$\Rightarrow \log\left(\frac{p_{ijc}}{p_{ij1}}\right) = x'_{ij} \cdot \boldsymbol{\beta}_c + z'_{ij} \cdot \mathbf{b}_{ic}. \tag{6.39}$$

onde $p_{ijc} = P(Y_{ij} = c)$ são as probabilidades da resposta, $p_{ij1} = P(Y_{ij} = 1)$ é a categoria de referência e as influências das covariáveis são avaliados através dos componentes $\beta_c = (\beta_{1c,\cdots,\beta_{pc}})'$. $\boldsymbol{\beta}_c$ são considerados efeitos fixos. Para os efeitos aleatórios $\mathbf{b}_{ic}$, assumimos uma distribuição normal multivariada com média zero e matriz de covariância $\mathbf{G}$.

A partir disso, a contribuição de verossimilhança do i-ésimo grupo é dado por:

$$l_i(\boldsymbol{\theta}_c, \boldsymbol{\beta}_c, \mathbf{G}) = \int_{-\infty}^{\infty} \left(\prod_{j=1}^{n_i} \left[\frac{\exp(x'_{ij}\boldsymbol{\beta}_c + z'_{ij}\mathbf{b}_{ic})}{\sum_{q=1}^{C} \exp(x'_{ij}\boldsymbol{\beta}_q + z'_{ij}\mathbf{b}_{iq})} \right]^{I(Y_{ij}=c)} \right) f_{\mathbf{b}}(\mathbf{b}_i, \mathbf{G}) d\mathbf{b}_i, \tag{6.40}$$

onde $f_b(\mathbf{b_i}, \mathbf{G})d\mathbf{b_i}$ é a densidade normal multivariada e I() a função indicadora (KUSS; MCLERRAN, 2007).

A função de verossimilhança geral é o produto das contribuições l_i para todos os grupos.

Com efeitos aleatórios seguindo uma distribuição normal, a distribuição marginal da resposta, obtida integrando os efeitos aleatórios, não possui uma forma fechada. Integrações numéricas usando Quadratura de Gauss-Hermite (ANDERSON; AITKIN, 1985), ou técnicas de Monte Carlo ((MCCULLOCH, 1997; BOOTH; HOBERT, 1999)) ou métodos de aproximação tais qual a Aproximação de Laplace e expansões de séries de Taylor ((BRESLOW; CLAYTON, 1993; WOLFINGER; O'CONNELL, 1993)), são utilizadas para aproximar a distribuição marginal e, assim, a função de verossimilhança e no final as estimativas de máxima verossimilhança e seus erros padrão.

6.5 Algumas Considerações Sobre Métodos de Estimação

Métodos iniciais de estimação propostos para MLGMs, quase-verossimilhança penalizada (PQL) e quase-verossimilhança marginal (MQL) (BRESLOW; CLAYTON, 1993), são baseados em ideias de linearização similares aos propostos por (LINDSTROM; BATES, 1990) no contexto de modelos mistos não lineares (MMNL). Tais métodos possuem a vantagem de possuírem simples implementação em *softwares* existentes para modelos lineares mistos, incluindo o caso de alguns MLGMs multiníveis.

Entretanto, ambos PQL e MQL têm sido reportados como produzindo estimativas enviesadas em alguns casos. Outro problema com a utilização de PQL e MQL na prática é que a função objetivo subjacente, a qual é otimizada para produzir os parâmetros, não é realmente uma aproximação da função de verossimilhança. Isso, em particular, não permite o uso de testes de razão de verossimilhança para comparar modelos lineares generalizados mistos aninhados baseados nas funções objetivos PQL ou MQL. Mais precisos, e computacionalmente intensivos, aproximações para as verossimilhanças de MMNL e MLGM foram propostas na literatura, incluindo aproximações de Laplace e métodos de quadratura ((PINHEIRO; CHAO, 2006) e (FITZMAURICE et al., 2008) entre outros).

Segundo Bolker et al. (2009), embora muitas ferramentas de estimação somente estão disponíveis em poucos pacotes estatísticos, ou são de grande dificuldade de utilização, a situação está gradualmente melhorando, com desenvolvedores de *softwares* melhorando suas publicações. Qual técnica de estimação é mais útil em dada situação depende da complexidade do modelo, assim como do tempo e poder computacional, bem como a disponibilidade de *softwares* e a aplicabilidade de diferentes métodos de inferência.

Em simulações realizadas por Capanu et al. (2013) no *software* `SAS`, os resultados indicaram que para uma grande quantidade de observações por efeito aleatório o método de quadratura adaptativo trabalha extremamente bem quando a implementação e tempo computacional são viáveis. Em tais cenários, ambas implementações (`PROC GLIMMIX` e `PROC NLMIXED`) produziram resultados semelhantes e a escolha entre os dois procedimentos pode ser feita em de acordo com a familiaridade do usuário com as sintaxes correspondentes. O método de quase-verossimilhança penalizada não demonstrou superioridade, pelo contrário, houve situações no qual foi bastante enviesado, corroborando com os autores citados anteriormente.

A Tabela 6 a seguir mostra alguns programas existentes para a análise de MLGM, assim como suas capacidades estatísticas.

Tabela 6 – Capacidade de diferentes *softwares* para análise de MLGM: Métodos de estimação, alcance dos modelos estatísticos que podem ser ajustados e métodos de inferência disponíveis (BOLKER et al., 2009).

		QVP	Laplace	QGH	EAC	Wald, χ^2 ou testes F de Wald	Graus de Liberdade	MCMC	CCT	Sobre.
SAS	**PROC GLIMMIX**	✓	✓^{a}	✓^{a}	✓	✓	BW, S, KR		✓	QL
	PROC NLMIXED			✓		✓	BW, S, KR		✓	Dist
R	**glmmPQL**	✓				✓	BW		✓	QL
	glmmML		✓	✓						
	glmer		✓	(✓)	✓			✓		QL
	glmmADMB		✓							Dist
	GLMM	✓			✓	✓			✓	QL
GenStat/ ASREML			✓	✓	✓			✓		Dist
AD Model Builder		✓	✓		✓					✓
HLM				✓						
GLLAAMM (Stata)								✓		Dist
WinBUGS					✓			✓		

QVP - Quase-Verossimilhança Penalizada, **QGH** - Quadratura de Gauss-Hermite, **EAC** - Efeitos Aleatórios Cruzados, **CCT** - Correlação Contínua Temporal, **Sobre** - Sobredispersão, **BW** - "entre-dentre", **Dist** - Distribuição específica (Binomial Negativa), **KR** - Kenward-Roger, **QL** - Quase-verossimilhança, **S** - Satterthwaite. ^{a} - Versão 9.2

Utilizado como método para aproximação da função de verossimilhança para os MLGM, a quadratura aproxima uma dada integral por uma soma de pesos sobre abscissas predefinidas para os efeitos aleatórios. Uma boa aproximação pode geralmente ser obtida com uma quantidade adequada de pontos de quadraturas (nós). A Quadratura Gaussiana Adaptativa, como descrito por Pinheiro e Bates (1995), centra e escala os pontos de quadratura usando Estimativas Empírica de Bayes (EEB), dos efeitos aleatórios e a matriz Hessiana da suboptimização da EEB (CAPANU et al., 2013).

Essa centralização e escala melhora a aproximação de verossimilhança colocando as abscissas de acordo com a função de densidade dos efeitos aleatórios. Além disso, o número de pontos de quadratura pode ser adaptado avaliando a função de log-verossimilhança nos valores iniciais do parâmetro em uma quantidade maior de nós até um ponto de tolerância for alcançado. A aproximação para a log-verossimilhança pode ser melhorado aumentando a precisão da integração numérica, e então é esperado que métodos de quadratura Gaussiana adaptativa tenham uma performance melhor do que alternativas baseadas em linearização.

Métodos de quadratura Gaussiana adaptativa estão atualmente implementados no software `SAS (Versão 9.4)` através das `PROC GLIMMIX` com a opção `"METHOD=QUAD"` e no `PROC NLMIXED`. Em ambos, como maneira padrão, a quantidade de pontos de quadratura é selecionado de acordo com a necessidade. Entretanto, há restrições quanto aos modelos que tais procedimentos conseguem trabalhar. A classe de modelos que, atualmente, podem ser estimados por quadratura Gaussiana adaptativa no `PROC GLIMMIX` é consideravelmente menor do que aquele feito com aproximação de Laplace (`PROC GLIMMIX METHOD=LAPLACE`).

Por exemplo, modelos de efeitos aleatórios cruzados ou modelos com sujeitos não aninhados não podem ser ajustados utilizando a abordagem de quadratura (devemos ser capazes de especificar o sujeito (opção `SUBJECT`) na parte aleatória (`RANDOM`) para conseguir utilizar o `PROC GLIMMIX METHOD=QUAD`). Uma limitação da `PROC NLMIXED` é que ela somente aborda modelos com apenas uma componente de variância, e portanto, modelos com multiplos componentes de variância ou modelos mistos multiníveis não lineares não são acomodados.

Ademais, aumentando o número de efeitos aleatórios, a abordagem de quadratura se torna computacionalmente inviável devido a alta dimensionalidade da integral, e isso limita a utilização dos `PROC GLIMMIX` e `PROC NLMIXED` para ajustar métodos de quadratura Gaussiana adaptativa.

6.5.1 Técnicas para verificação do ajuste

6.5.1.1 Teste da Razão de Verossimilhança Restrita

Uma maneira existente para compararmos a qualidade do ajuste de dois modelos, um dos quais é um caso particular do outro, é através do teste da razão de verossimilhança. O teste é baseado na razão de verossimilhança, a qual expressa o quão melhor os dados estão ajustados a um modelo do que no outro.

Quando a diferença entre dois modelos ajustados é somente a implementação de um efeito aleatório, tais modelos são "encaixados", possibilitando a realização do teste de razão de verossimilhança e, como estamos estudando os Modelos Lineares Generalizados Mistos, a presença dos efeitos aleatórios leva a necessidade de utilizar um teste de razão de verossimilhança baseado na estimativa de máxima verossimilhança restrita (REML), pois a estimação de verossimilhança estima componentes de variação enviesados conforme Pinheiro e Bates (2006), e uma maneira de corrigir esse problema é utilizar o teste que se baseia na estimação de máxima verossimilhança restrita, a qual tem por ideia básica a remoção dos parâmetros β da verossimilhança, fazendo com que esta seja definida somente em termos da matriz de covariância.

O teste se dá por

$$TRVR = -2\left[l(\theta_0) - l(\theta_1)\right] \tag{6.41}$$

onde $l(\theta_0)$ representa o modelo que possuímos, e $l(\theta_1)$, o modelos que queremos comparar.

Assim, desejamos testar:

$$\begin{cases} H_0 : l(\theta_0) = l(\theta_1) \\ H_1 : l(\theta_0) \neq l(\theta_1) \end{cases} \tag{6.42}$$

Para o caso em estudo, $l(\theta_0)$ representa o modelo multinível multinomial já com o acréscimo do efeito aleatório representado a hierarquia dos dados, e $l(\theta_1)$ representa o modelo multinomial, composto somente por efeitos fixos. Desejamos testar se ambos os modelos são igualmente representativos de acordo com a hipótese nula, ou se são diferentes, o que nos levaria a utilizar o modelo mais completo.

6.5.1.2 Coeficiente de Correlação Intraclasse

O grau de semelhança entre "micro unidades" pertencentes à mesma "macro unidade", ou então as unidades de nível 1 pertencences à mesma unidade de nível 2, podem ser expressas pelo Coeficiente de Correlação Intraclasse (CCI). Esse coeficiente pode ser definido como

$$\text{CCI} = \frac{\text{Variância da população de macro unidades}}{\text{Variância total}}. \tag{6.43}$$

A variância da população de macro unidades para este trabalho se torna a variância encontrada entre as Macro Regionais de Saúde.

A extensão desse coeficiente para um modelo multinível multinomial segundo Snijders e Bosker (1999), Grilli e Rampichini (2007) é dada para cada categoria da variável resposta, a menos da utilizada como referência. Ou seja, para o caso em estudo, serão verificadas 8 valores de coeficientes de correlação intraclasse, que identificarão o grau de dependência das unidades dentro das Regionais de Saúde em uma mesma Macro Regional para cada tipo do câncer de mama segundo sua categoria CID. Também de acordo com os autores supracitados, o coeficiente de correlação intraclasse para um modelo multinomial com variância do intercepto dado por τ^2_{0c} para cada categoria c, é dado por

$$\text{CCI}_c = \frac{\tau^2_{0c}}{\tau^2_{0c} + \frac{\pi^2}{3}}. \tag{6.44}$$

6.5.1.3 Análise de Resíduos

O comportamento dos resíduos através de exibições gráficas é uma boa ferramenta para verificar a análise dos erros estatísticos do modelo, e para determinar discrepâncias no modelo ou a presença de *outliers* que necessitam de uma investigação mais profunda.

Entretanto a análise de resíduos para dados multiníveis, especialmente para dados não Gaussianos, é menos desenvolvida do que em outros casos. O *software* utilizado neste estudo, `SAS` Versão 9.4, com a `PROC GLIMMIX`, não disponibiliza os valores de resíduos, assim como gráficos de diagnósticos para a distribuição multinomial (SCHABENBERGER, 2005), dificultando essa análise.

6.5.2 Inferência e Predição

Além da estimação dos parâmetros, pode-se também conduzir inferências sobre os parâmetros ou obter predições dos efeitos aleatórios. Pode-se conduzir testes dos parâmetros de efeitos fixos utilizando técnicas de inferências usuais de máxima verossimilhança como testes de razão de verossimilhança, porém testar os componentes de variância para os efeitos aleatórios não é direto.

Predições dos efeitos aleatórios $\mathbf{b}_i$, $i = 1, \cdots, n$ ou combinações lineares de efeitos fixos e aleatórios são baseadas na esperança condicional de $\mathbf{b}_i$, com os dados e a estimativa dos parâmetros finais.

Segundo Hartzel e Agresti (2001), para um modelo multinomial de efeitos aleatórios, essa esperança tem a forma

$$E\left[\mathbf{b_i}|y_i, \hat{\beta}, \hat{\mathbf{G}}\right] = \frac{\int \cdots \int \mathbf{b_i} \left[\prod_{j=1}^{T_i} f(y_{ij}|\hat{\beta}; \mathbf{b_i})\right] g(\mathbf{b_i}; \hat{\mathbf{G}}) d\mathbf{b_i}}{\int \cdots \int \left[\prod_{j=1}^{T_i} f(y_{ij}|\hat{\beta}; \mathbf{b_i})\right] g(\mathbf{b_i}; \hat{\mathbf{G}}) d\mathbf{b_i}} \tag{6.45}$$

e necessita de uma aproximação por integrais, como o uso da quadratura adaptativa de Gauss-Hermite, ou integração de Monte Carlo. Embora as esperanças envolvam somente os dados para o *cluster* i, as estimativas de $\mathbf{b_i}$ emprestam informações de todos os *clusters* desde que $\hat{\beta}$ e $\hat{\mathbf{G}}$ são obtidas usando todos os dados.

Capítulo 7

Características das internações por câncer de mama no estado do Paraná de 2008 a 2016

A neoplasia maligna de mama é o tipo de câncer mais comum entre as mulheres no mundo e o segundo no Brasil depois apenas do câncer de pele não melanoma, respondendo por cerca de 25% dos casos novos a cada ano. O câncer de mama também acomete homens, porém é raro, representando apenas 1% do total de casos da doença. Relativamente raro antes dos 35 anos, acima desta idade sua incidência cresce progressivamente, especialmente após os 50 anos.

Segundo o Sistema de Informação sobre Mortalidade (SIM), em 2013 foram registrados 14387 óbitos devido ao câncer de mama, sendo 181 homens e 14206 mulheres. A previsão para 2016 segundo o INCA (2016) é de 57960 novos casos.

Todo ano, 23% dos novos casos de câncer em mulheres estão relacionados com câncer de mama e segundo Organization (2016), até 2030 podemos esperar 27 milhões de casos de câncer, 17 milhões de mortes relacionadas com câncer e 75 milhões vivendo com câncer anualmente.

Nas últimas três décadas, o câncer de mama tem se constituído na primeira causa de morte por câncer na população feminina, registrando-se aumento das taxas de mortalidade, ajustadas por idade pela população mundial, de 38% (variação percentual relativa), entre os anos de 1979 a 2011 (RUIZ; JUNIOR, 2015).

No Brasil tem-se observado um crescimento na proporção de mortes por câncer de mama em relação ao total de mortes de mulheres. No período de 1991 até 1994, 1.8% das mortes de mulheres no Brasil eram devido ao câncer de mama. Já no período de

2007 a 2010 esse número passou para 2.6% com uma variação proporcional de 39.4% (KLUTHCOVSKY et al., 2014).

De acordo com a Figura 9, nota-se a diferença entre a taxa de mortalidade do câncer de mama com outros tipos de câncer em mulheres no Brasil.

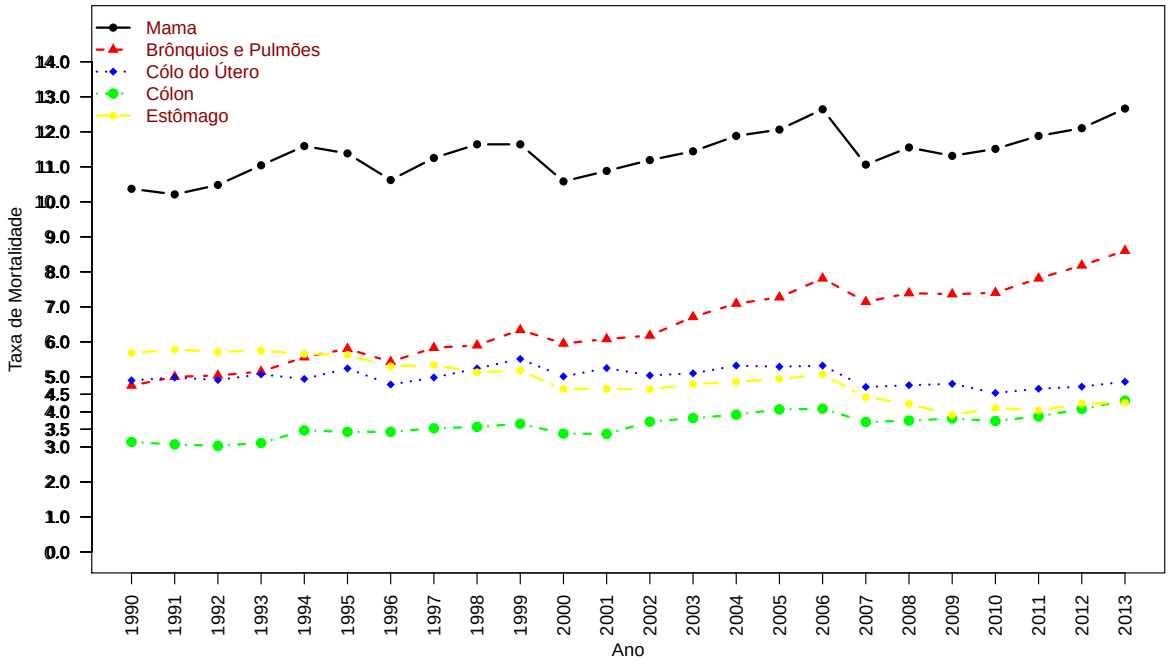


Figura 9 – Taxas de mortalidade para as 5 localizações primárias mais frequentes de câncer em mulheres, de 1990 a 2013, no Brasil por cada 100.000 habitantes.

Mulheres de países mais desenvolvidos têm mais “chance” de desenvolver câncer de mama, porém conseguem diagnósticos mais precoces e tratamentos mais precisos. O que leva à relação INCIDÊNCIA/MORTALIDADE ser menor em tais localidades. Diagnósticos em estágios avançados da doença reduzem as chances de cura, e é um dos fatores responsáveis pela alta taxa de mortalidade. Acesso limitado da população ao tratamento, seja devido à distribuição desigual de renda ou escassez de atendimentos do serviço público levam ao aumento de óbitos no Brasil (RODRIGUES et al., 2015).

Segundo Ruiz e Junior (2015), 30 a 40% dos diagnósticos de câncer de mama no Brasil são detectados com tamanhos maiores que 3 cm. Isso é relevante pois câncer de mama levam, em média, 8 a 10 anos para alcançarem 1 cm, o que, em princípio daria tempo suficiente para um diagnóstico precoce. Com essa dimensão, se nada for feito, pode levar a morte em até 3 anos, mas o paciente raramente morre por problemas locais, mas geralmente como resultados de metástases, geralmente nos ossos, pulmões e eventualmente no sistema nervoso central.

A Tabela 7 exemplifica a necessidade e importância de um diagnóstico precoce, mostrando a sobrevida esperada de 5 anos, dependendo do tamanho do tumor, em uma pesquisa realizada entre 1975 e 1999 nos Estados Unidos, segundo (ELKIN et al., 2005).

Tamanho do Tumor	**Linfonodo-negativo***	**Linfonodo-positivo***
Menor que 1cm	100%	93%
1-1.9 cm	100%	91%
2-2.9 cm	93%	85%
3-3.9 cm	86%	71%
4-4.9 cm	82%	68%
Maior que 5cm	81%	63%

Tabela 7 – Sobrevida relativa de 5 anos com câncer de mama. *Status do Linfonodo: Mostra se o câncer se espalhou ou não para os linfonodos.

Considerando os cânceres que não se espalharam para os linfonodos, podemos notar uma diferença de 14% na sobrevida de 5 anos entre pacientes diagnosticados com tumores menores que 1cm em comparação com pacientes diagnosticados com tumores entre 3 e 4 cm.

No estado do Paraná e seus 399 municípios, os dados não diferem muito do resto do Brasil. A Figura 10 evidencia a taxa de mortalidade do câncer de mama para o estado do Paraná, comparado com as neoplasias mais comuns, em mulheres, de 1990 a 2013.

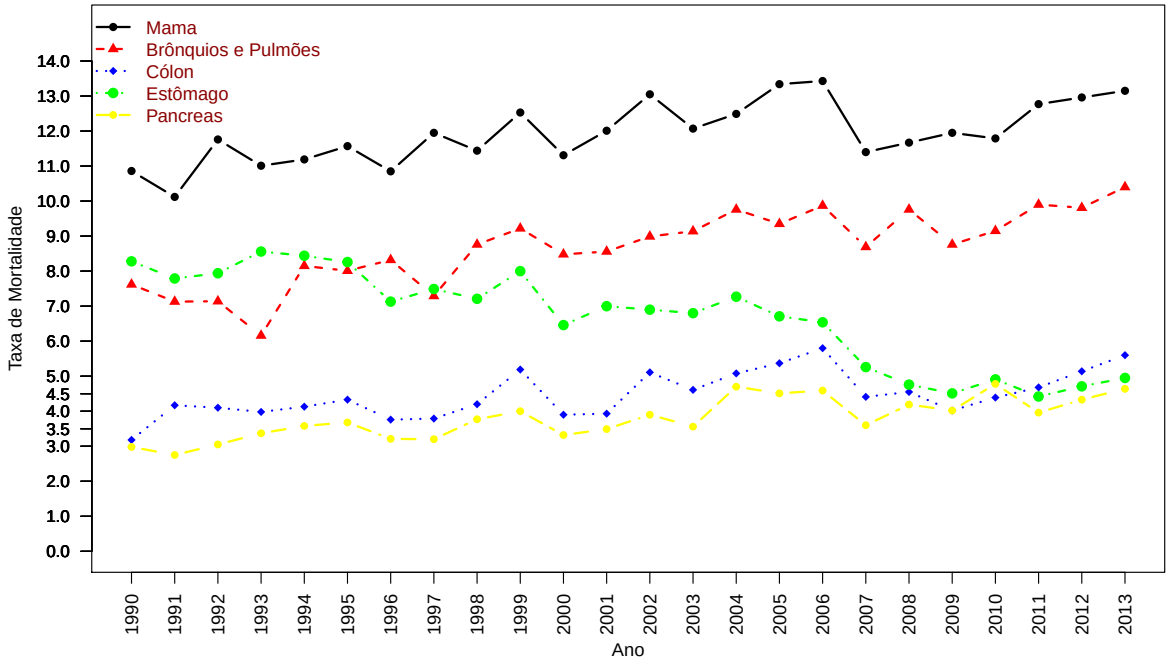


Figura 10 – Taxas de mortalidade para as 5 localizações primárias mais frequentes de câncer em mulheres, de 1990 a 2013, no estado do Paraná por cada 100.000 habitantes.

Podemos verificar a diferença entre a taxa de mortalidade devido ao cânder mama comparado com os demais, representando uma taxa de aproximandamente 3 pontos a mais para o segundo tipo de câncer com maior taxa, o câncer de Brônquio e Pulmões, e uma diferença de aproximadamente 8 pontos quando comparada com o câncer de Pâncreas.

7.1 Descrição dos Dados

7.1.1 População do Estudo

A população do estudo foi constituída por todas as internações por câncer de mama ocorridas em hospitais que atendem pelo Sistema Único de Saúde (SUS), no estado do Paraná, entre 2008 a 2016. Os dados secundários sobre as internações foram obtidos do Sistema de Informações Hospitalares do SUS (SIH-SUS) e de domínio público. O SIH-SUS é alimentado pelos dados da Autorização de Internação Hospitalar (AIH), que se referem a morbidade e mortalidade, além de informações financeiras.

O banco de dados utilizado possui 32541 entradas e cada entrada caracteriza uma internação realizada pelo SUS entre os anos de 2008 a 2016. Cada linha de informação foi criada após todo o tratamento do paciente, ou seja, possui todas as informações de gasto final e o resultado em morte ou não. Suas variáveis são *Município de Residência, Município*

de Atendimento, Sexo, Idade, Raça, Diagnóstico Principal, Morte, Valor Total US, Ano, Faixa Etária, Regional de Atendimento, Regional de Residência e Macro Regional de atendimento.

7.1.2 CID-10

A Classificação Internacional de Doenças e Problemas Relacionados com a Saúde, frequentemente designada pela sigla CID (em inglês: International Statistical Classification of Diseases and Related Health Problems - ICD) fornece códigos relativos à classificação de doenças e de uma grande variedade de sinais, sintomas, aspectos anormais, queixas, circunstâncias sociais e causas externas para ferimentos ou doenças.

No CID-10, as neoplasias malignas estão destacadas entre C00 e C97, e estão separado pela topografia (localização anatômica do tumor, ou seja, a parte do corpo humano em que o tumor se encontra instalado) do câncer.

7.1.2.1 C50

Foram utilizados somente classificações do CID-O/3 (Classificação Internacional de Doenças para Oncologia, 3ª edição) que representam a seção topográfica dos tumores. O código CID-O/3 topográfico da mama é C50 (exclui a pele da mama). Os códigos do câncer de mama com caracterização topográfica são:

- C50.0 - Mamilo;
- C50.1 - Porção central da mama;
- C50.2 - Quadrante superior interno da mama;
- C50.3 - Quadrante inferior interno da mama;
- C50.4 - Quadrante superior externo da mama;
- C50.5 - Quadrante inferior externo da mama;
- C50.6 - Prolongamentos axilar da mama;
- C50.8 - Lesão sobreposta da mama;
- C50.9 - Mama, sem outras especificações.

Note que a classificação C50.9 representa uma generalização do câncer de mama. Quanto o diagnóstico não é preciso, a doença é caracterizada com essa classificação. O decréscimo de quantidade de casos identificados com esse CID é uma importante implicação de uma melhoria em termos de diagnóstico.

O total de internações realizadas no período de 2008 a 2016 estão separados pelo CID conforme a Figura 11.

Nota-se que os 11157 casos para o C50.9 (Mama, sem outras especificações) e os 8895 casos para o C50.8 (Lesão sobreposta da mama) representam 61.62% de todos os 32541 casos, conforme a Tabela 8, assim como representam 81.18% de todos os 2126 óbitos.

As classificações do CID para o câncer de mama que possuem a menor quantidade de ocorrências são C50.2 e C50.3, que representam os quadrantes superior e inferior internos da mama, respectivamente. Para as demais categorias, existem uma menor variação na quantidade de casos.

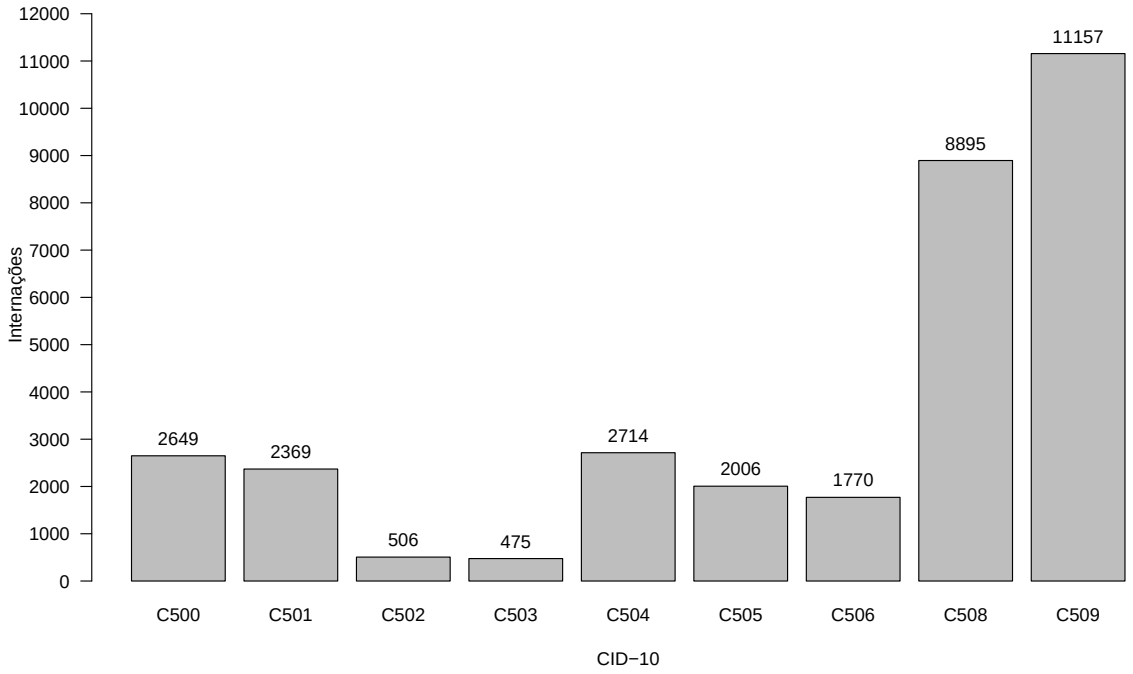


Figura 11 – Total de internações por câncer de mama separadas pelo CID de 2008 a 2016 no Paraná.

A partir da Tabela 8, nota-se que alguns tipos de câncer de mama possuem porcentagens de óbito maiores que as demais. Por exemplo, para o CID C50.9, 10.98% das internações resultam em óbito, enquanto que para o CID C50.6 somente 0.96%. A média das porcentagens de óbito é de 4.09% das internações.

Tabela 8 – Quantidade de internações e óbitos separados por CID para o câncer de mama no Paraná entre 2008 e 2016

Internações Totais			**Óbitos**		
Diag. Principal		**Porcentagem**		**Quantidade**	**Porcentagem**
C50.0	2649	8.14%	**SIM**	172	6.49%
			NÃO	2477	93.51%
C50.1	2369	7.28%	**SIM**	48	2.03%
			NÃO	2321	97.97%
C50.2	506	1.55%	**SIM**	14	2.77%
			NÃO	492	97.23%
C50.3	475	1.46%	**SIM**	12	2.53%
			NÃO	463	98.47%
C50.4	2714	8.34%	**SIM**	109	4.01%
			NÃO	2605	95.99%
C50.5	2006	6.16%	**SIM**	28	1.40%
			NÃO	1978	98.60%
C50.6	1770	5.44%	**SIM**	17	0.96%
			NÃO	1753	99.04%
C50.8	8895	27.33%	**SIM**	501	5.63%
			NÃO	8394	94.37%
C50.9	11157	34.29%	**SIM**	1225	10.98%
			NÃO	9932	89.02%

A Figura 12 representa a mudança da quantidade de ocorrências, separados por CID, de 2008 a 2016. Nota-se que para os casos C50.1, C50.6 houve uma diminuição da quantidade de casos (43.68% e 89.65% respectivamente), e para os casos C.50.4 e C50.8, houve um crescimento na quantidade de internações até 2015 de (313.14% e 214.07% respectivamente), ocorrendo uma diminuição em ambos para o ano de 2016. As demais alteram entre a diminuição e o aumento do número de occorências ao longo dos anos.

Já a Figura 13 caracteriza como as internações estão separados por CID, ano por ano. Nota-se que somente a partir do ano de 2014, o C50.9 não é o caso com o maior número de ocorrências, sendo ultrapassado pelo C50.8. Pode-se perceber que a diferença da quantidade entre os casos C50.9 e C50.8 para com os demais tende a crescer ao longo dos anos. Em 2008, os dois juntos representaram 50.86% dos atendimentos, já em 2016, ambos representaram 71.97% do total de internações do ano.

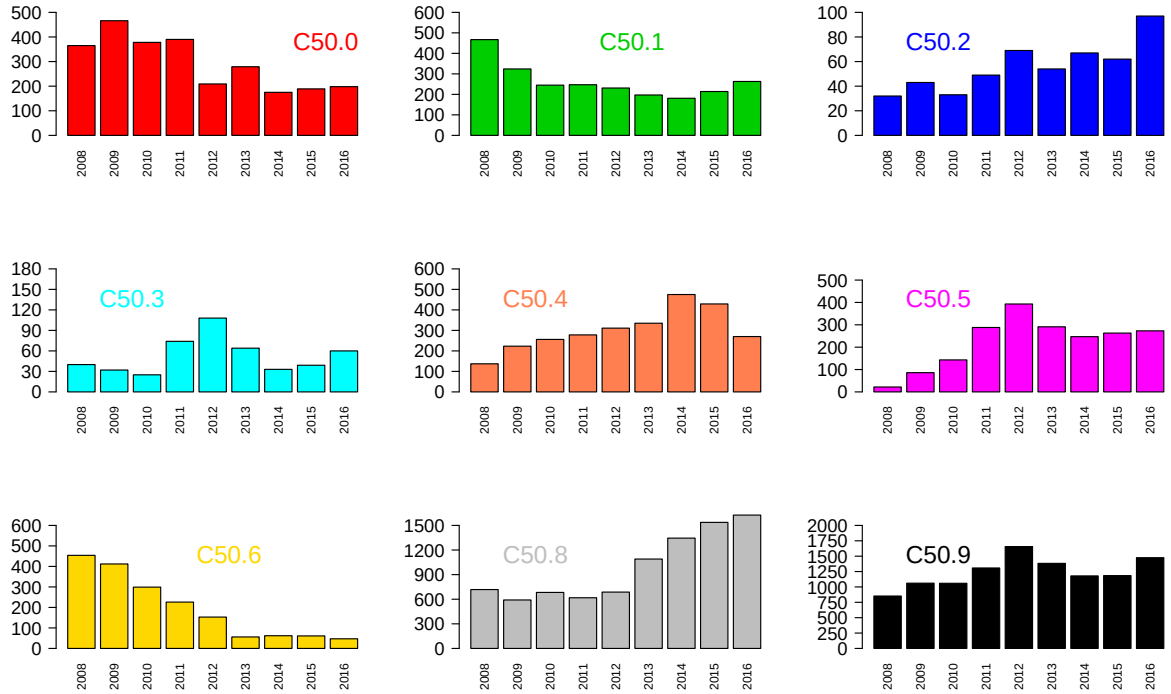


Figura 12 – Quantidade de internações por câncer de mama separados pelo ano e por CID no Paraná entre 2008 e 2016.

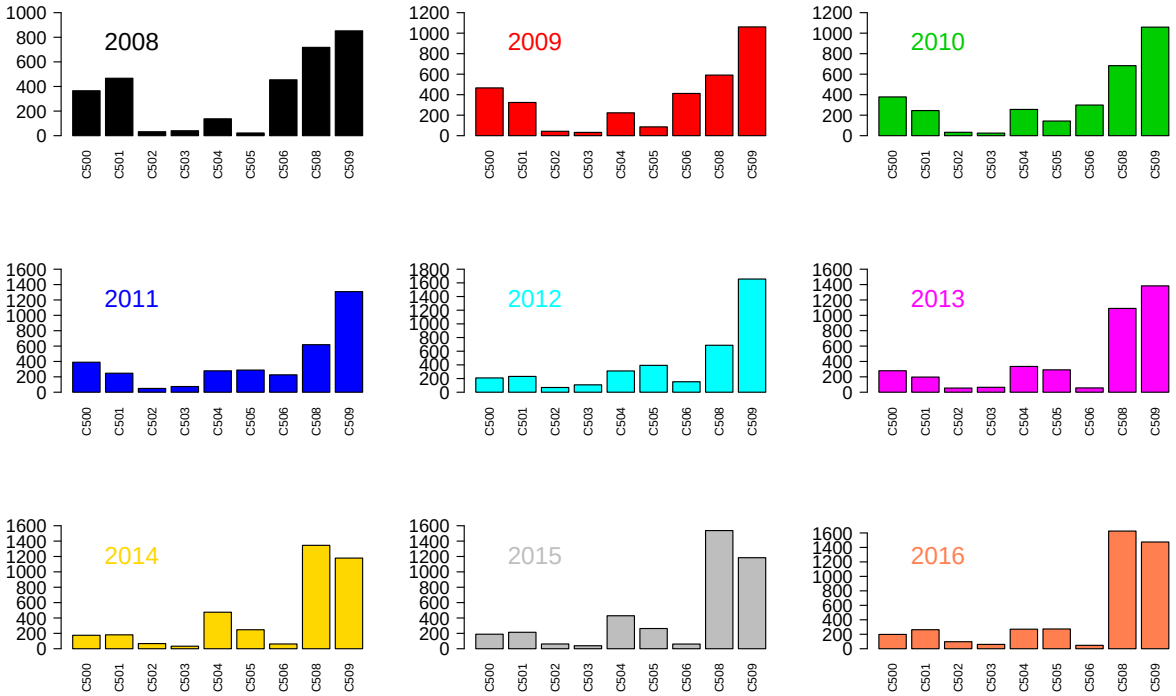


Figura 13 – Total de internações por câncer de mama separados pelo CID entre 2008 e 2016 no Paraná.

Assim como na Figura 13, a Figura 14 mostra o perfil das internações separados por CID, ao longo dos anos. Exibindo crescimento de alguns e diminuição de outros.

Não considerando os casos C50.9 e C50.8, os demais casos, mesmo variando suas quantidades, não passaram de 500 ocorrências cada.

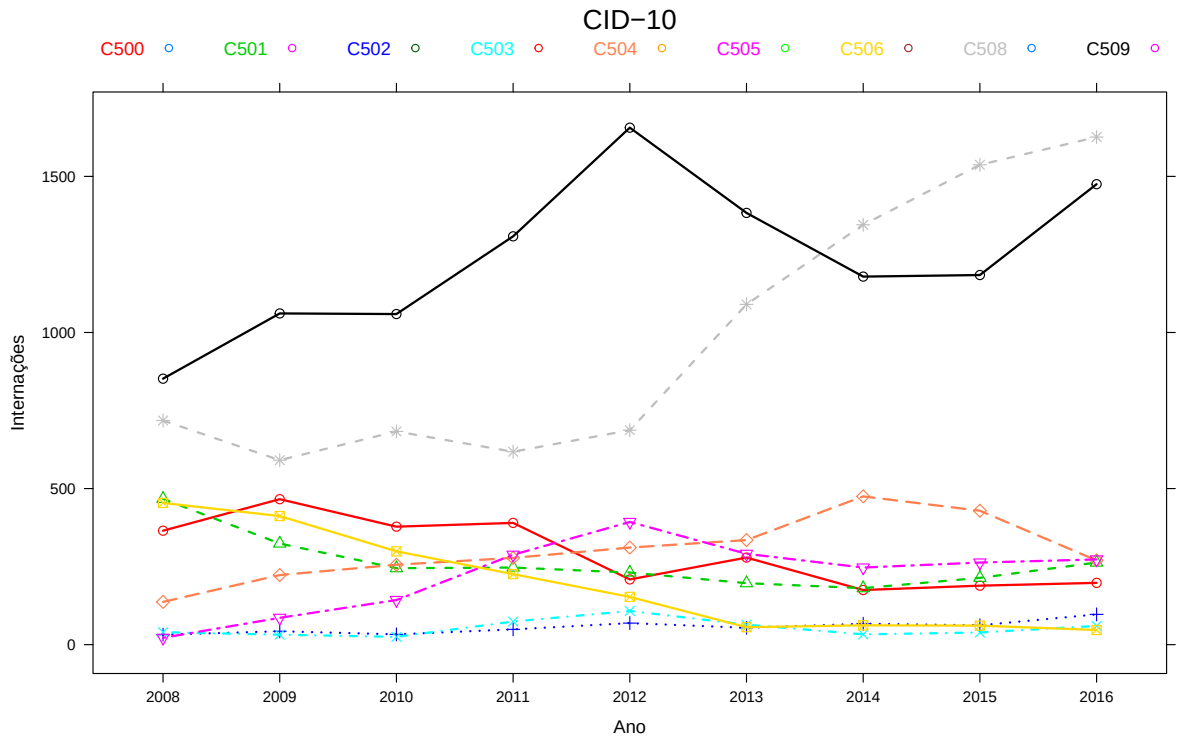


Figura 14 – Gráfico de perfil da classificação do câncer de mama segundo o CID entre 2008 e 2016 no Paraná.

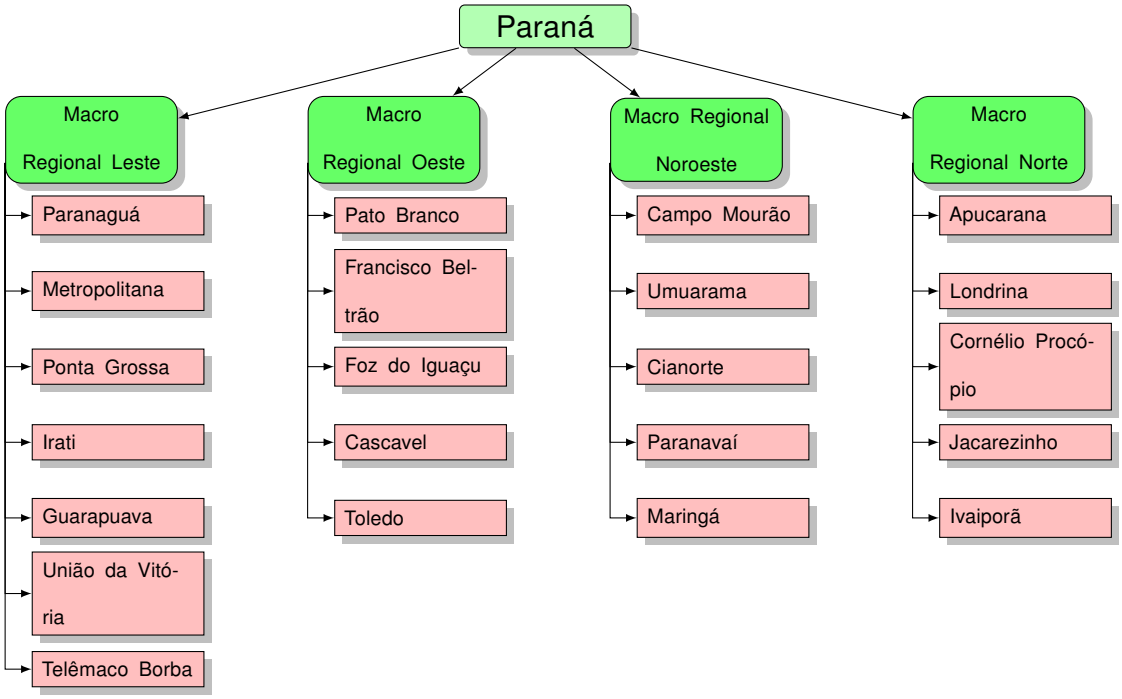
Nota-se, a partir de 2012, a diminuição do caso C50.9, que representa o câncer de mama indicado como “sem outras especificações”. Isso pode levar a crer que a qualidade da descrição do câncer, e sua caracterização topográfica tenha melhorado nos diagnósticos e internações no estado do Paraná.

7.1.3 Regionais de Saúde

O Estado do Paraná com seus 399 municípios e população de 10.577.755 (censo IBGE/2010) está subdividido em 22 Regionais de Saúde e segundo informações da Secretaria de estado da saúde do paraná (SESA/PR, 2016):

> *"As Regionais de Saúde constituem a instância administrativa intermediária da Secretaria de Estado da Saúde do Paraná (SESA/PR). Principalmente através delas o Estado exerce o seu papel. Este papel é menos o de executar ações e serviços de saúde e mais de apoio, cooperação técnica e investimentos nos municípios e nos consórcios. Os municípios, isoladamente ou aglutinados em módulos intermunicipais, devem assumir todas as ações e serviços que possam por eles ser absorvidos. À Regional de Saúde cabe desenvolver a inteligência necessária para apoiar o município em todas as áreas e para influenciar na gestão das questões regionais, fomentando a busca contínua e crescente da eficiência com qualidade. As 4 Macrorregionais de Saúde não constituem novas instâncias administrativas, não têm sede e nem funcionários. Seu objetivo é articular as Regionais de Saúde em conjuntos para que possam, também entre si, somar esforços na solução de problemas comuns (como por exemplo o encaminhamento de doentes para centros de referência) e trocar experiências. Cada Macrorregional conta com um Assessor de Macrorregião que tem a incumbência de assessorar as suas regionais e o conjunto delas nas articulações necessárias"*

A seguir segue a lista das 22 Regionais de Saúde e as 4 Macro Regionais as quais pertencem, lembrando que a Regional Metropolitada caracteriza a Regional que comporta a cidade de Curitiba.



As características das regionais são abrangentes em termos de IDH:

- IDH médio (média ponderada, utilizando a população como peso) de 0.674 na regional de Telêmaco Borba;
- IDH médio (média ponderada, utilizando a população como peso) de 0.777 na regional Metropolitana.

Em termos de renda per capta:

- renda per capta média de R$512.50 na regional de Ivaiporã;
- renda per capta média de R$1149.52 na regional Metropolitana.

E em termos de população total:

- população total de 137169 na regional de Ivaiporã;
- população total de 3285851 na regional Metropolitana (que inclui Curitiba).

A Organização e Divisão Judiciárias do Estado do Paraná, definidas pelo Tribunal de Justiça e atualizadas pela Lei Estadual nº 16.352/2009-PR - de 22 de dezembro de 2009, possui características e objetivos diferentes da divisão da área do estado em Regionais de Saúde - e, ainda além, obedece “critérios de democratização da gestão e do acesso à Justiça” (art. 3º da Lei Estadual nº 14.277/2003-PR - 30/12/2003).

Apesar disso, e provavelmente devido às características geográficas, existem poucas diferenças entra as divisões. Porém, a única relação direta entra as divisões, devido ao tamanho dos municípios como base para escolha como sede, é que todas as cidades sedes de Regional de Saúde são, também, sedes de Seção Judiciária.

Segue, abaixo, lista com as 14 comarcas que possuem municípios, num total de 25, com Regionais de Saúde diferentes da regional da própria comarca a que pertencem (TJPR, 2009):

- Andirá - Regional: 18ª RS - Cornélio Procópio
  Barra do Jacaré - Regional: 19ª RS - Jacarezinho
- Assaí - Regional: 17ª RS - Londrina
  Nova América da Colina - Regional: 18ª RS - Cornélio Procópio
  São Sebastião da Amoreira - Regional: 18ª RS - Cornélio Procópio
- Cidade Gaúcha - Regional: 13ª RS - Cianorte
  Nova Olímpia - Regional: 12ª RS - Umuarama
  Tapira - Regional: 12ª RS - Umuarama

- Cruzeiro do Oeste - Regional: 12ª RS - Umuarama
  Tapejara - Regional: 13ª RS - Cianorte
  Tuneiras do Oeste - Regional: 13ª RS - Cianorte

- Curiúva - Regional: 21ª RS - Telêmaco Borba
  Figueira - Regional: 19ª RS - Jacarezinho
  Sapopema - Regional: 18ª RS - Cornélio Procópio

- Faxinal - Regional: 16ª RS - Apucarana
  Cruzmaltina - Regional: 22ª RS - Ivaiporã

- Grandes Rios - Regional: 16ª RS - Apucarana
  Rio Branco do Ivaí - Regional: 22ª RS - Ivaiporã
  Rosário do Ivaí - Regional: 22ª RS - Ivaiporã

- Imbituva - Regional: 4ª RS - Irati
  Ivaí - Regional: 3ª RS - Ponta Grossa

- Mallet - Regional: 4ª RS - Irati
  Paulo Frontin - Regional: 6ª RS - União da Vitória

- Matelândia - Regional: 9ª RS - Foz do Iguaçu
  Céu Azul - Regional: 10ª RS - Cascavel
  Diamante d'Oeste - Regional: 20ª RS - Toledo
  Vera Cruz do Oeste - Regional: 10ª RS - Cascavel

- Paranacity - Regional: 15ª RS - Maringá
  Cruzeiro do Sul - Regional: 14ª RS - Paranavaí
  Inajá - Regional: 14ª RS - Paranavaí
  Jardim Olinda - Regional: 14ª RS - Paranavaí
  Paranapoema - Regional: 14ª RS - Paranavaí

- Pitanga - Regional: 5ª RS - Guarapuava
  Mato Rico - Regional: 22ª RS - Ivaiporã
  Santa Maria do Oeste - Regional: 22ª RS - Ivaiporã

- Ribeirão do Pinhal - Regional: 18ª RS - Cornélio Procópio
  Jundiaí do Sul - Regional: 19ª RS - Jacarezinho

- Uraí - Regional: 18ª RS - Cornélio Procópio
  Jataizinho - Regional: 17ª RS - Londrina

Todos as informações de internações realizadas levam em consideração a regional de RESIDÊNCIA dos pacientes.

A maioria das internaçoes analisadas são registradas à 2ª Regional de Saúde (Metropolitana), a qual contempla a capital do estado, Curitiba. Como podemos ver na Figura 15, existe uma grande diferença da quantidade de internações totais dessa regional com as demais.

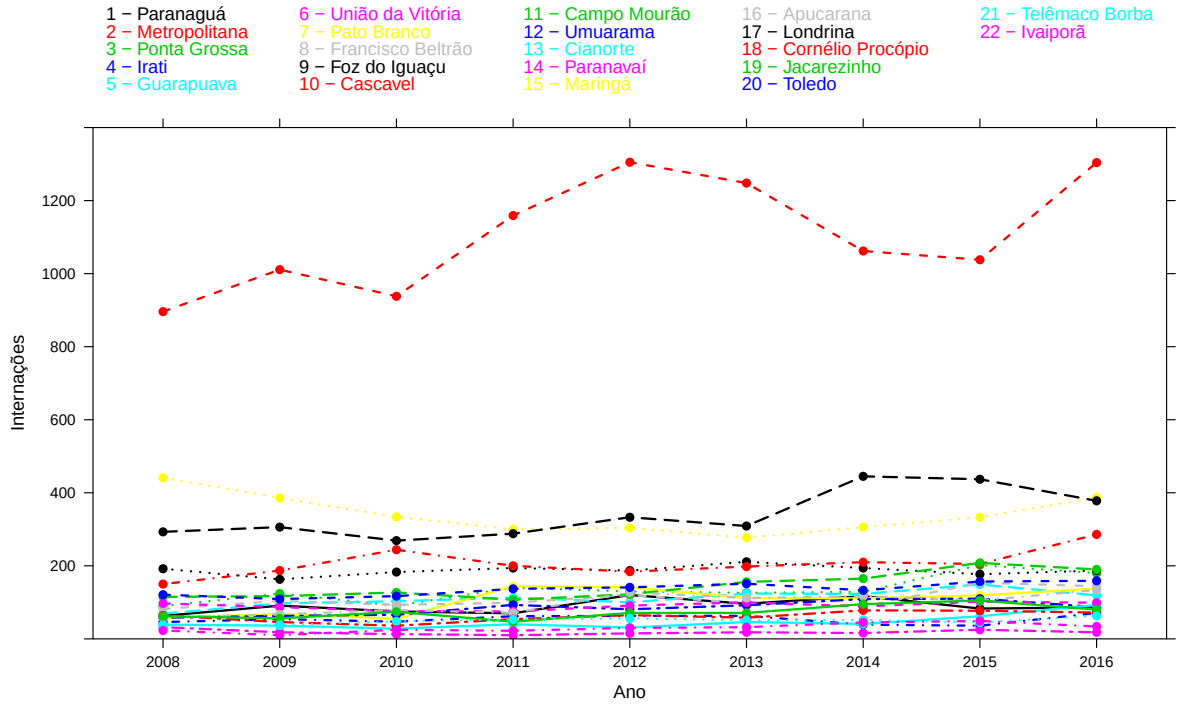


Figura 15 – Quantidade de internações por câncer de mama segundo as Regionais de Saúde de 2008 a 2016 no Paraná.

Porém, quando tratamos as internações como taxas, levando em consideração a população residente em cada regional, é notável a diferença.

As Figuras 16 a 20 descrevem a taxa de internação das regionais de saúde para cada 100 mil habitantes no período de 2008 a 2016. A Figura 16 mostra a taxa de todas as 22 regionais de saúde. A Figura 17 mostra a taxa para a Macro Região Oeste, a Figura 18 mostra a taxa para a Macro Região Noroeste, a Figura 19 mostra a taxa para a Macro Região Norte e a Figura 20 mostra a taxa para a Macro Região Leste.

Para a Macro Regional Oeste, damos ênfase para a regional de Pato Branco, que sofreu um grande aumento em sua taxa de internação entre 2010 e 2011. As demais regionais seguem aproximadamente o mesmo padrão, e aumentando levemente suas taxas de internação. Nota-se também um leve aumento na quantidade de internações nas Regionais de Pato Branco e Cascavel de 2015 para 2016.

A Figura 18 que representa a taxa de internações para a Macro Regional Noroeste mostra a regional de Maringá sofrendo uma queda em sua taxa de internações, enquanto que as demais sofrem um leve aumento ou se mantém constante, com excessão da Regional de Saúde de Cianorte que apresenta um grande aumento a partir de 2014 ultrapassando, em 2016, a faixa das 60 internações por câncer de mama para cada 100 mil habitantantes.

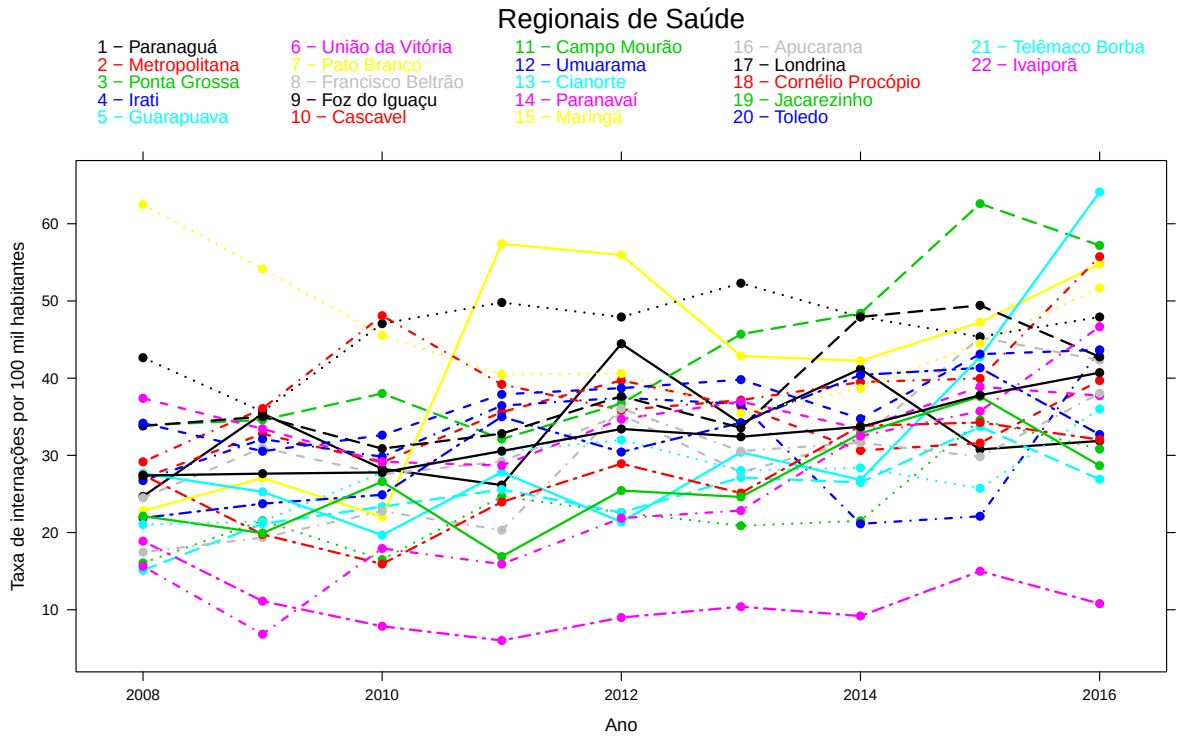


Figura 16 – Perfil da taxa de internação por câncer de mama entre 2008 e 2016 das Regionais de Saúde do Paraná.

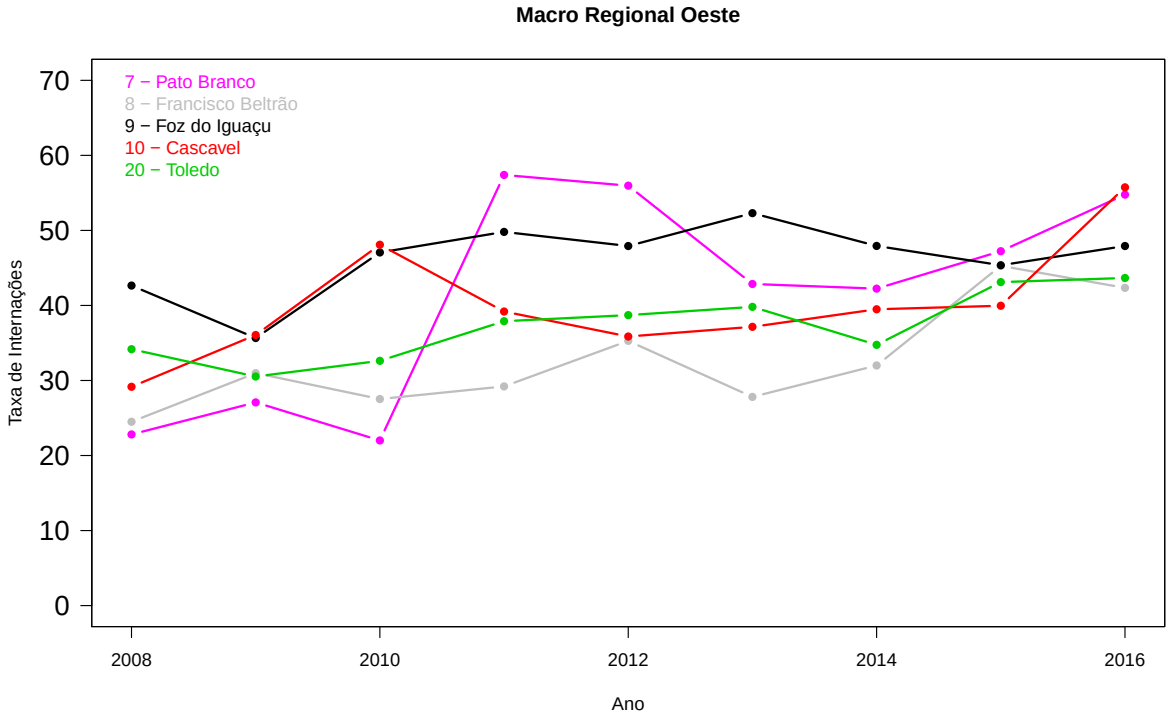


Figura 17 – Perfil da taxa de internação por câncer de mama entre 2008 a 2016 na Macro Regional Oeste.

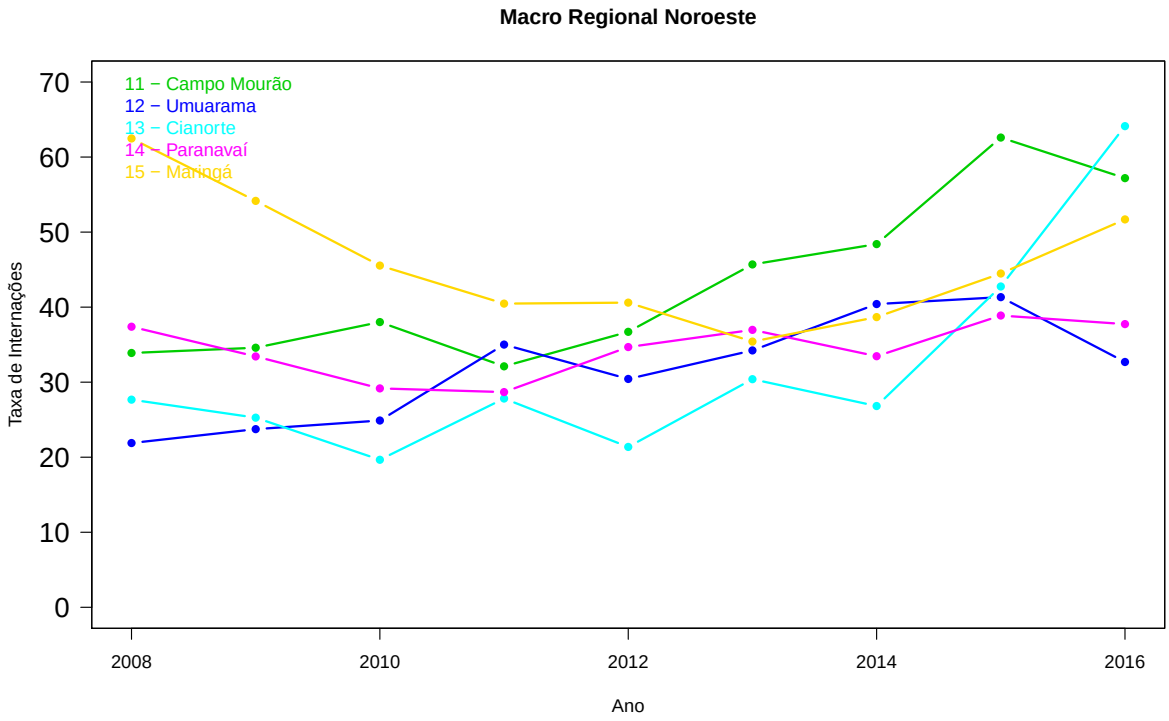


Figura 18 – Perfil da taxa de internação por câncer de mama entre 2008 e 2016 na Macro Regional Noroeste.

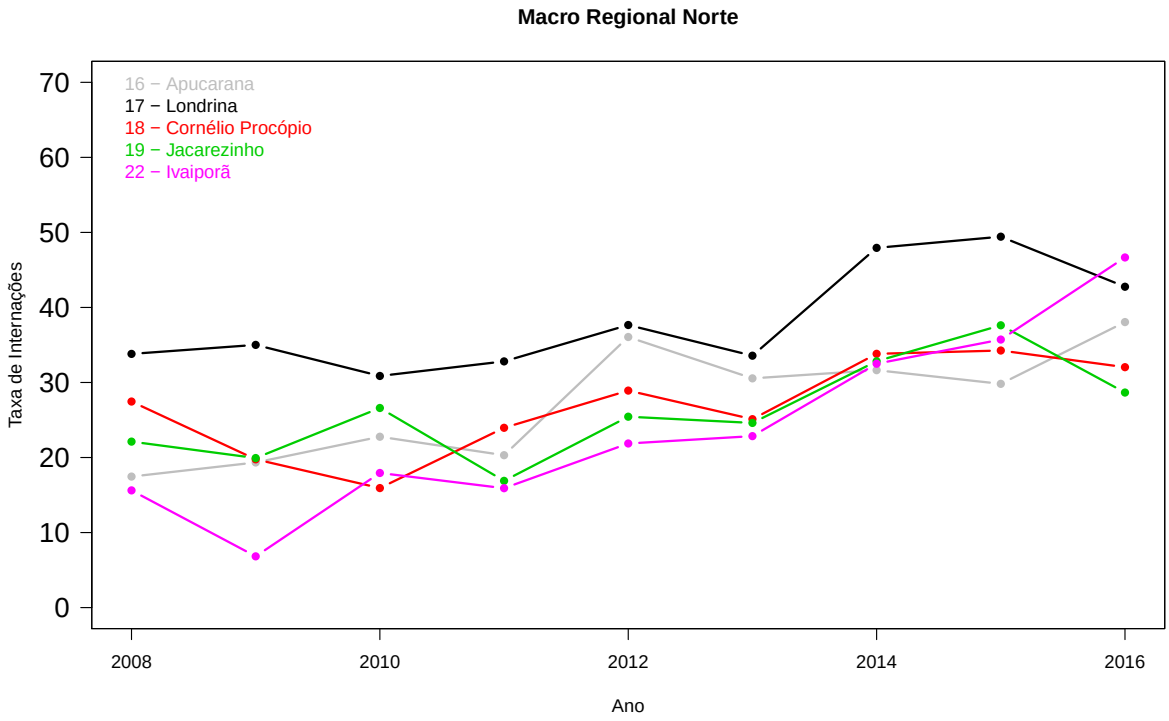


Figura 19 – Perfil da taxa de internação por câncer de mama entre 2008 e 2016 na Macro Regional Norte.

As taxas de internações para a Macro Regional Norte exibidas na Figura 19, mostram um padrão para todas as regionais, seguindo um leve crescimento de suas taxas.

Para a Macro Regional Leste, nota-se que a regional de União da vitória se mantém com a menor taxa entre todas as regionais através de todos os anos, enquanto as demais regionais dessa Macro Região sofrem aumento de suas taxas ao longo dos anos.

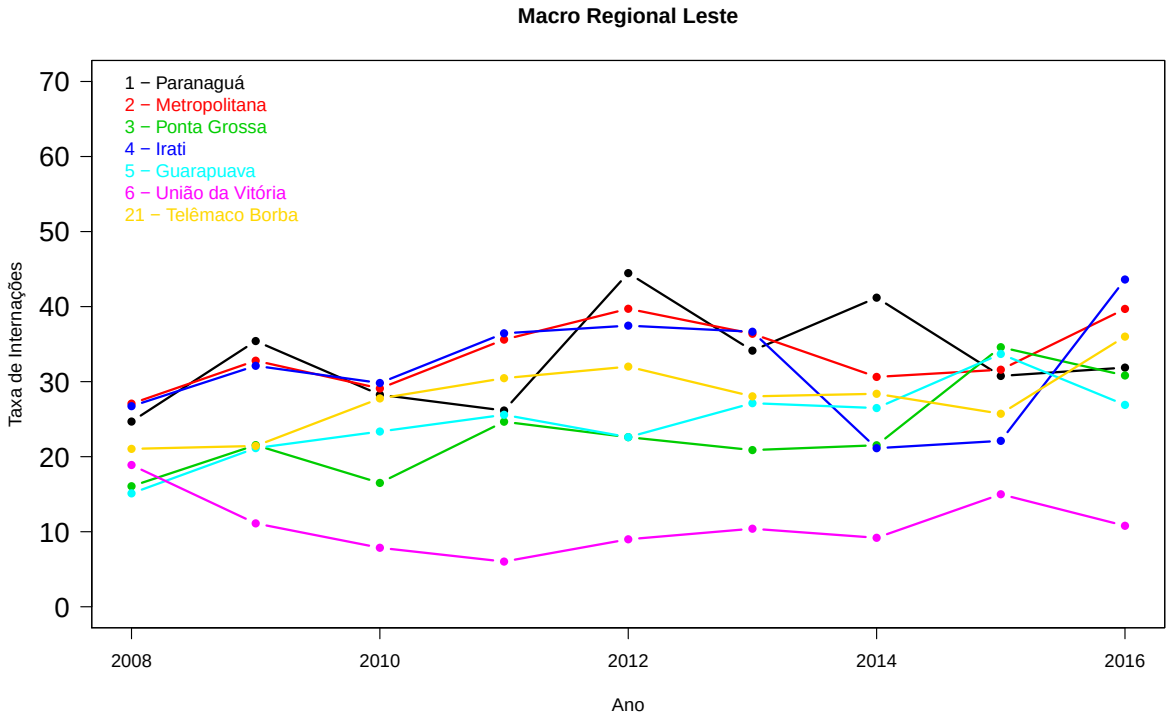


Figura 20 – Perfil da taxa de internação por câncer de mama entre 2008 e 2016 na Macro Regional Leste.

A Tabela 9 apresenta as taxas dos internações, divididas por Regional e por ano. Considerando todos os anos e todas as regionais, somente em 11 ocasiões a taxa de internação foi superior a 51 para cada 100 mil habitantes, porém somente 2016 foi responsável por 5 destes casos. Em 2008 e 2009, na Regional de Maringá, com 62.48 e 54.16 respectivamente; em 2011 e 2012, na Regional de Pato Branco com 57.39 e 55.96 respectivamente, em 2013 na Regional de Foz do Iguaçu com 52.30 internações, em 2015 na Regional de Campo Mourão com 62.60 casos, e em 2016 nas Regionais de Pato Branco com 54.77, Cascavel com 55.74, Campo Mourão com 57.19, Cianorte com 64.12 e Maringá com 51.69.

Tabela 9 – Taxa de internações por câncer de mama nas Regionais de Saúde entre 2008 e 2016 no Paraná.

Regional de Saúde	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**MÉDIA por Reg.**
1ª Paranaguá	24,68	35,41	28,26	26,15	44,46	34,13	41,19	30,76	37,87	**32,99**
2ª Curitiba	27,07	32,79	29,09	35,60	39,71	36,38	30,63	31,59	39,69	**33,62**
3ª Ponta Grossa	16,06	21,54	16,50	24,66	22,60	20,88	21,52	34,59	30,82	**23,24**
4ª Irati	26,73	32,11	29,82	36,44	37,46	36,65	21,13	22,11	43,61	**31,78**
5ª Guarapuava	15,11	21,14	23,35	25,58	22,60	27,13	26,48	33,69	26,90	**24,66**
6ª União da Vitória	18,89	11,11	7,86	6,02	8,99	10,40	9,19	14,99	10,79	**10,92**
7ª Pato Branco	22,81	27,08	22,01	**57,39**	**55,96**	42,86	42,24	47,23	54,77	**41,37**
8ª Francisco Beltrão	24,5	30,98	27,53	29,21	35,28	27,81	32,00	45,28	42,34	**32,77**
9ª Foz do Iguaçu	42,65	35,66	47,06	49,80	47,92	**52,30**	47,93	45,36	47,92	**46,29**
10ª Cascavel	29,16	36,06	48,09	39,19	35,86	37,15	39,49	39,96	55,74	**40,08**
11ª Campo Mourão	33,89	34,59	38,01	32,11	36,71	45,69	48,40	**62,60**	57,19	**43,24**
12ª Umuarama	21,89	23,74	24,89	35,01	30,43	34,23	40,41	41,33	32,69	**31,62**
13ª Cianorte	27,67	25,27	19,66	27,82	21,37	30,40	26,82	42,75	**64,12**	**31,76**
14ª Paranavaí	37,39	33,43	29,17	28,68	34,68	36,97	33,46	38,88	37,74	**34,49**
15ª Maringá	**62,48**	**54,16**	45,54	40,47	40,60	35,41	38,67	44,48	51,69	**45,94**
16ª Apucarana	17,46	19,34	22,77	20,31	36,06	30,55	31,64	29,82	38,05	**27,33**
17ª Londrina	33,81	35,01	30,87	32,81	37,66	33,57	47,94	29,43	42,76	**38,21**
18ª Cornélio Procópio	27,46	19,76	15,93	23,96	28,92	25,11	33,82	34,26	32,04	**26,81**
19ª Jacarezinho	22,12	19,93	26,60	16,89	25,44	24,61	32,83	37,62	28,66	**26,08**
20ª Toledo	34,17	30,53	32,62	37,90	38,71	39,81	34,75	43,11	43,66	**37,25**
21ª Telêmaco Borba	21,04	21,43	27,76	30,46	32,00	28,04	28,38	25,72	36,00	**27,87**
22ª Ivaiporã	15,62	6,83	17,94	15,91	21,87	22,85	32,50	35,72	46,66	**23,99**
MÉDIA por ANO	**27,39**	**27,63**	**27,79**	**30,56**	**33,42**	**32,41**	**33,70**	**37,78**	**40,71**	**32,38**

Considerando todas as regionais em todos os anos, a regional de Foz do Iguaçu obteve a maior média, com 46.29 internações para cada 100 mil habitantes e a segunda maior média ocorreu na regional de Maringá, com 45.94 internações. Entre todos os valores, as duas maiores taxas foram encontradas para a Regional de Cianorte em 2016 com 64.12 e Campo Mourão em 2015 com 62.60 internações para cada 100 mil habitantes.

7.1.4 Distribuição Espacial para o Estado do Paraná das taxas de internação

A Figura 21 mostra a distribuição espacial das internações por câncer de mama no estado do Paraná separados por Regionais de Saúde. Cada mapa representa um ano, e suas cores representam as taxas de internação para cada 100 mil habitantes. Seguindo da cor mais clara para a mais escura, demonstrando aumento na taxa das internações. As cores estão caracterizadas seguindo os intervalos: Menor que 20, entre 21 e 30, entre 31 e 40, entre 41 e 50 e entre 51 a 65 internações por 100 mil habitantes.

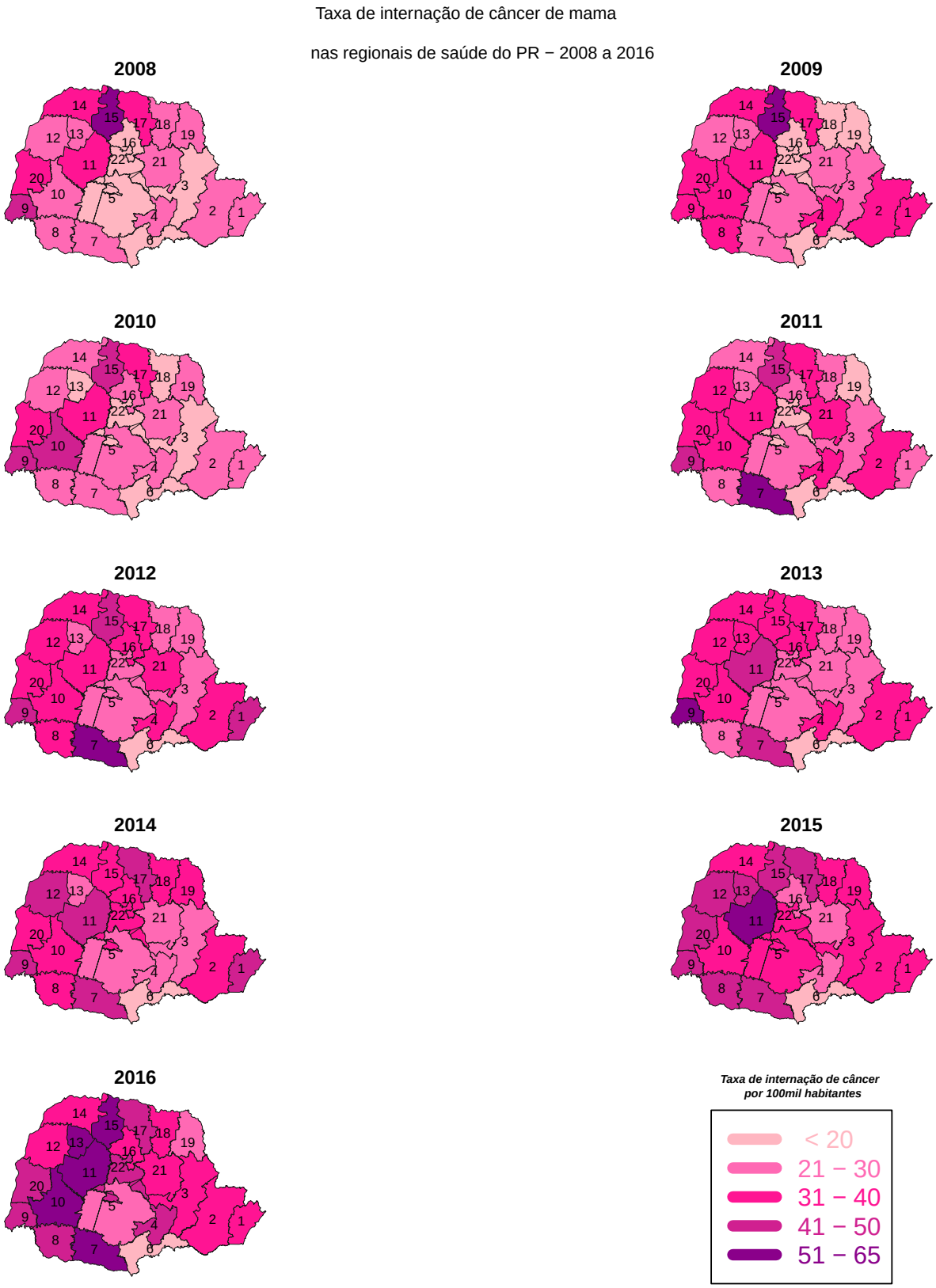


Figura 21 – Mapa do estado do Paraná com as taxas das internações por câncer de mama das Regionais de Saúde entre 2008 e 2016.

A partir da Figura 21 e da Tabela 9, nota-se que nos primeiros anos da comparação, as regionais pertencentes à Macro Regional Leste e Norte possuem uma taxa de internação predominantemente pequena, enquanto que as regionais da Macro Região Noroeste possuem uma taxa maior, obtendo a terceira maior taxa entre todas (e entre todos os anos), com 62.48 e com a Regional 11 (Campo Mourão (2015)) com a segunda maior taxa entre todas as regionais (e entre todos os anos), com 62.6 internações para cada 100 mil habitantes. Enquanto ao longo dos anos a regional de Maringá (15) diminui sua taxa, as regionais próximas aumentam seu valor. A partir de 2014, somente a regional de Umuarama (no ano de 2014), das regionais pertencentes à Macro Regional Noroeste, não se encontra na faixa superior a 30. Mesmo esta ultrapassando essa marca a partir de 2015.

As regionais pertencentes à Macro Regional Oeste também sofrem um aumento em sua taxa, resultando em todas as regionais presentes a essa Macro estarem com taxas superiores a 31 internações para cada 100 mil habitantes.

No ano de 2014, não considerando a regional de Umuarama, e algumas regionais presentes na Macro Regional Leste, todas as regionais estão na faixa de taxa superior a 31. Da Macro Regional Leste, as regionais 3, 4, 5, 6 e 21 (Ponta Grossa, Irati, Guarapuava, União da Vitória e Telêmaco Borba) estão na faixa inferior a 30.

Embora algumas regionais tenham diminuído sua taxa de internações, com a regional de Maringá a única a diminuir suficientemente para trocar de faixa, a média total das regionais cresceu o suficiente para essa troca. Passou de 27.39 em 2008 para 33.70 em 2014 e 40.71 em 2016. Ou seja, na média, a quantidade de internações por câncer de mama foi superior ao crescimento populacional por ano.

A Figura 22 permite comparar a quantidade de internações separados por CID para cada Macro Regional. É notável a diferença entre cada uma das figuras. Na Macro Regional Leste, há predominância de casos com o CID C50.9 com 40.49% de todos os casos. Para a Macro Regional Oeste, os CIDs C50.9 e C50.8 juntos somam 75.83% de todos as internações. Já para as macro regionais Norte e Noroeste, os atendimentos estão distribuídos sem tamanha diferença, mas ainda com os CIDs C50.2, C50.3 e C40.5 com uma quantidade de internações menor do que as demais.

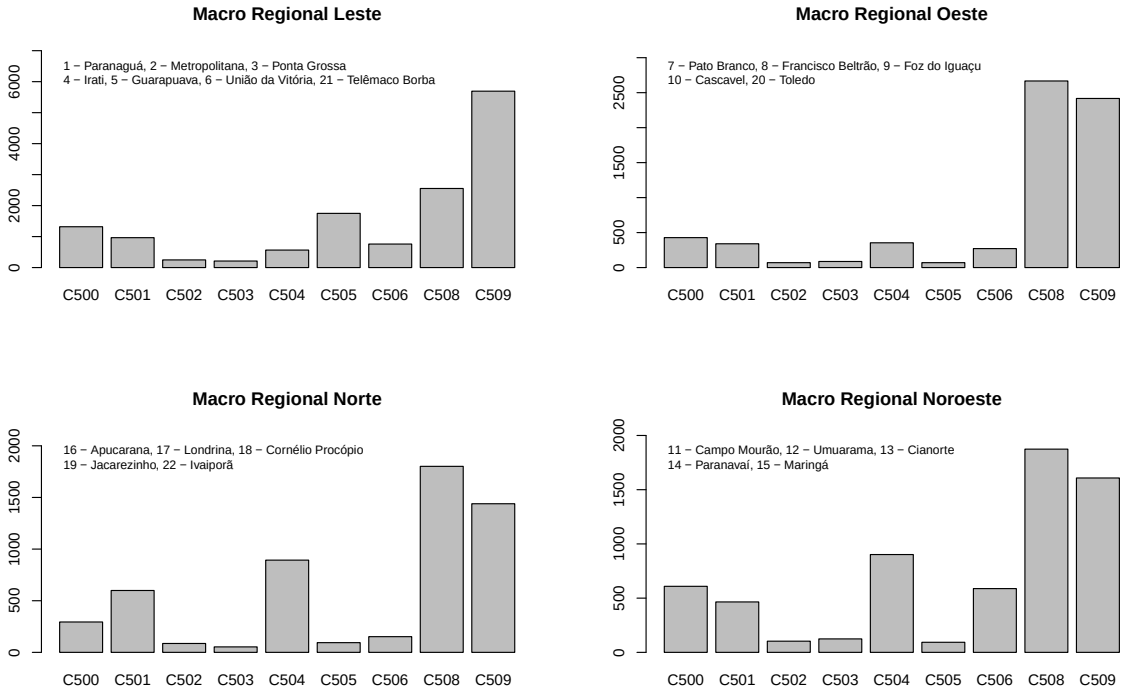


Figura 22 – Número de internações de câncer de mama separadas pela classificação do CID pelas Macro Regionais no estado do Paraná entre 2008 e 2016.

7.1.5 Faixa Etária

A faixa etária para caracterização das internações foi separada em cinco classes: 0 a 14 anos; 15 a 24 anos; 25 a 44 anos; 45 a 64 anos; e 65 anos ou mais. Estas categorias foram baseadas nas diretrizes internacionais das Nações Unidas (ORGANIZATION, 1982).

A Figura 23 mostra a separação das internações totais por faixa etária. A Figura está de acordo com estudos do INCA (2016), no qual a predominância dos casos é na faixa etária entre 45 a 64 anos.

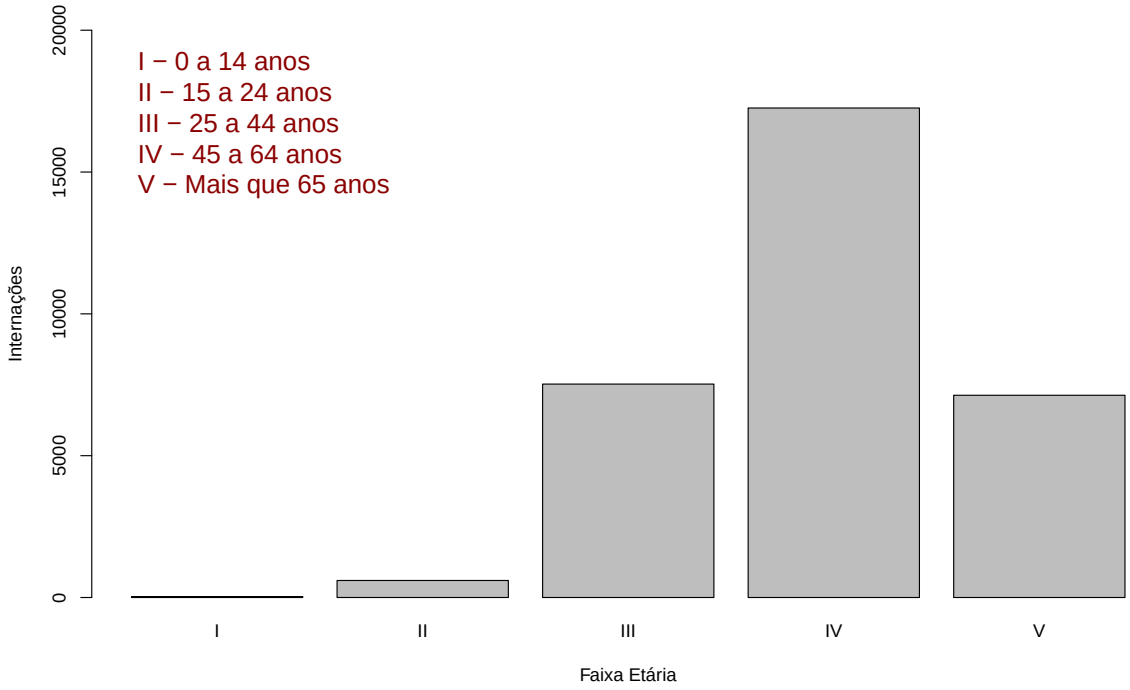


Figura 23 – Total de internações por câncer de mama por faixa etária no Paraná entre 2008 e 2016.

As seguintes figuras exibem como estão distribuídos as internações dependendo do CID dentro de cada faixa etária.

A Figura 24 representa o número de internações separados por CID dentro da primeira faixa etária, que varia entre 0 e 14 anos. Nota-se que a quantidade de casos é pequena comparado com as demais faixas etárias. Para as topografias C50.1, C50.2, C50.3 e C50.4 houveram somente um caso entre 2008 e 2016. Outro fato que merece atenção é a diminuição dos casos ao longo dos anos. Em 2012 tiveram 2 casos, em 2013 apenas 1 caso e nenhum caso registrado em 2014. Porém os casos voltaram a surgir em 2015 e 2016 com ênfase para a topografia C50.6 que, embora tenha timinuido as quantidades entre 2012 e 2015, voltou a crescer em 2016.

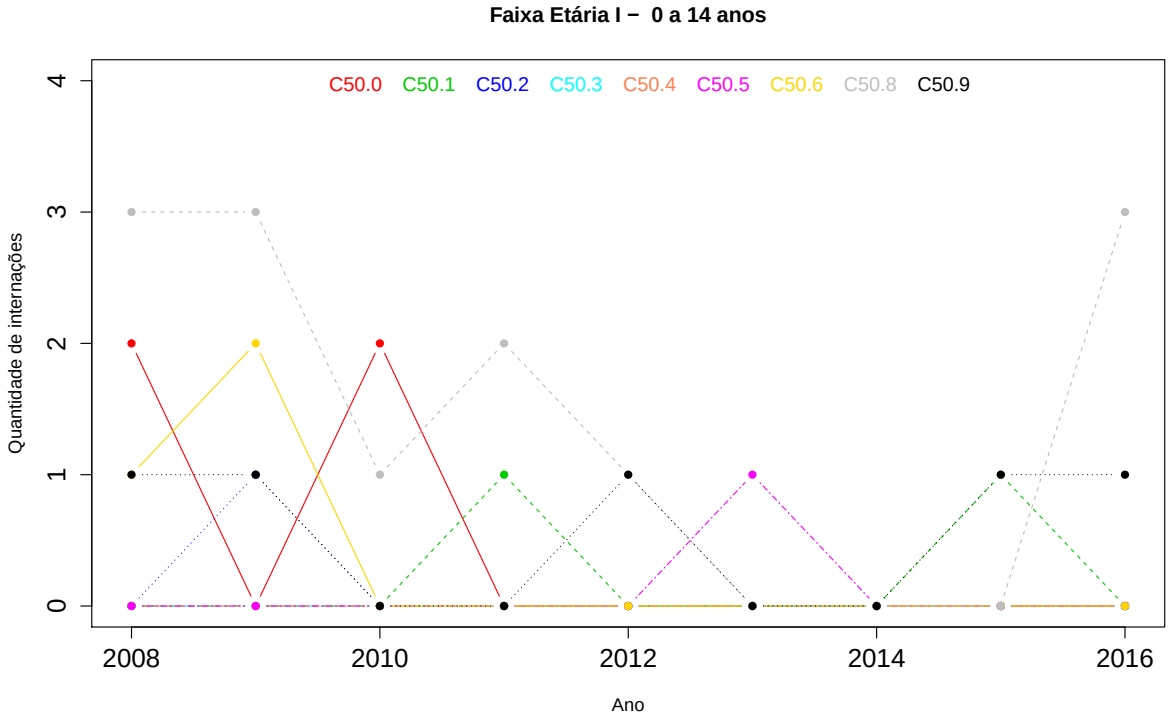


Figura 24 – Internações por câncer de mama na 1ª faixa etária separadas pelo CID entre 2008 e 2016 no Paraná.

A Figura 25 apresenta a distribuição das internações separados por CID para a segunda faixa etária, de 15 a 24 anos. Para tal faixa etária podemos notar, com exceção do CID C50.8, a diminuição significativa dos casos. Considerando todos as classificações, houve grande diminuição dos casos, de 125 em 2008 para 42 em 2014, com 24 casos somente no CID C50.8. Os valores se mantiveram em 2015, porém a quantidade de internações voltou a crescer no ano de 2016.

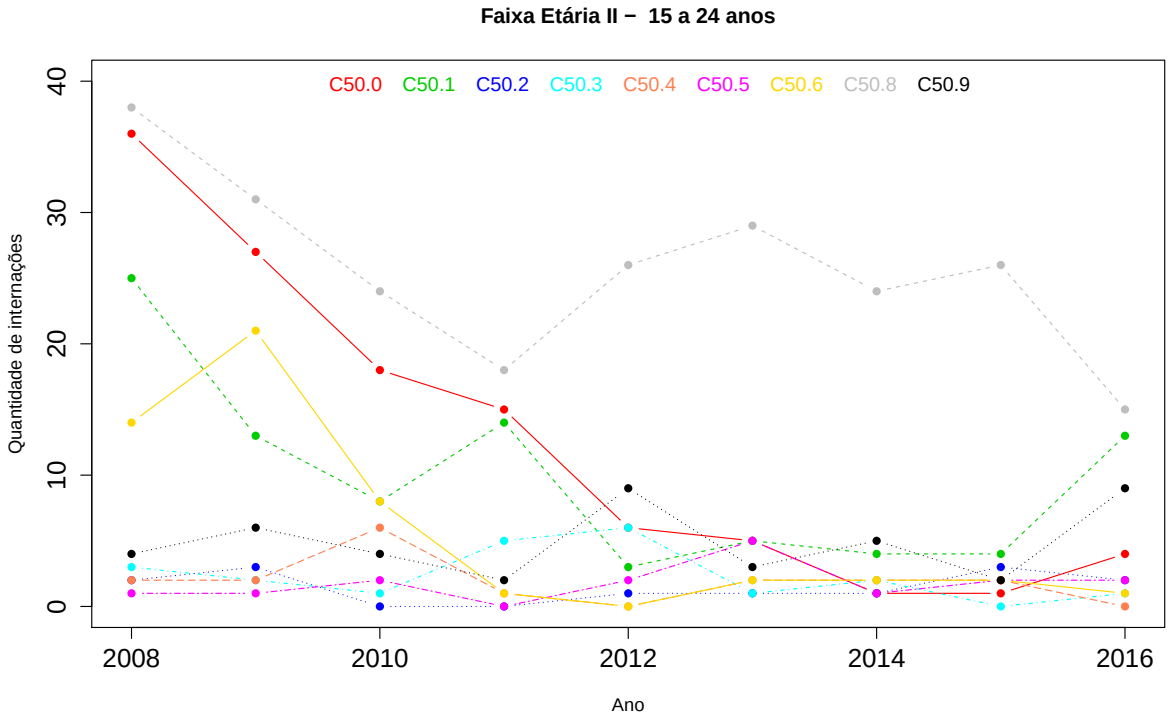


Figura 25 – Internações por câncer de mama na 2ª faixa etária separadas pelo CID entre 2008 e 2016 no Paraná.

As Figura 26 e 27 mostram a distribuição das internações separados por CID para a terceira faixa etária, de 25 a 44 anos e para a quarta faixa etária, de 45 a 64 anos, respectivamente. A partir de tais Figuras, que se comportam praticamente da mesma maneira, nota-se a diferença entre as classificações C50.8 e C50.9 com as demais. Enquanto elas sofrem grandes variações e se encontram, em 2014, com um aumento na quantidade de casos, as demais classificações se mantém estáveis durante os anos em estudo.

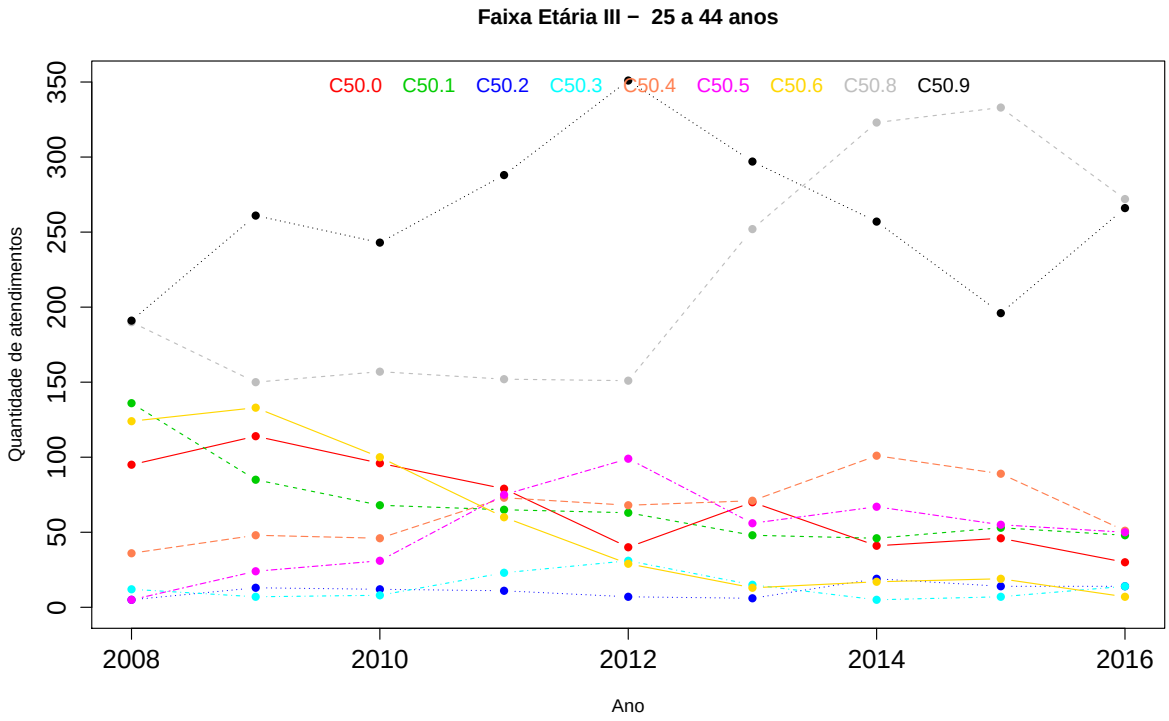


Figura 26 – Internações por câncer de mama na 3ª faixa etária separadas pelo CID entre 2008 e 2016 no Paraná.

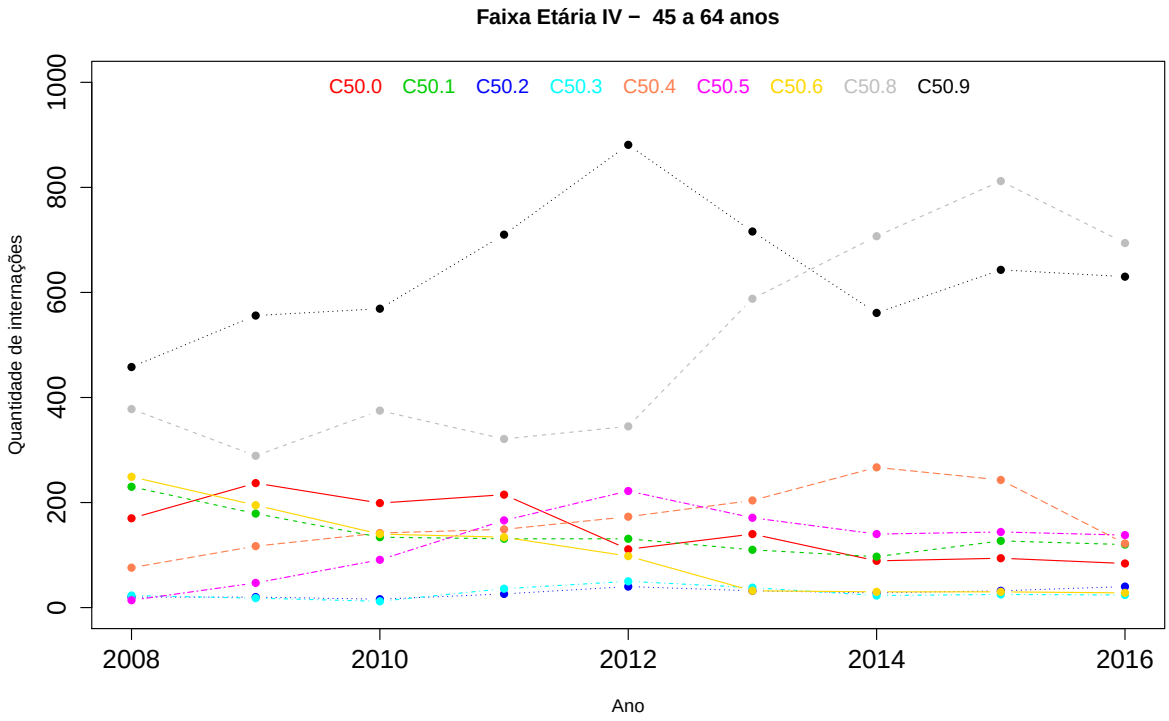


Figura 27 – Internações por câncer de mama na 4ª faixa etária separadas pelo CID entre 2008 e 2016 no Paraná.

Seguindo a mesma ideia das Figura 26 e 27, a Figura 28, que representa a distribuição das internações separadas por CID para a última faixa etária, maiores que 65 anos, apresenta as classificações C50.8 e C50.9 muito acima das demais. Enquanto a classificação C50.9 mostra uma certa tendência a diminuir ou estabilizar seu valor, a classificação C50.8 segue eu um crescimento grande ao longo dos anos de 2008 a 2015.

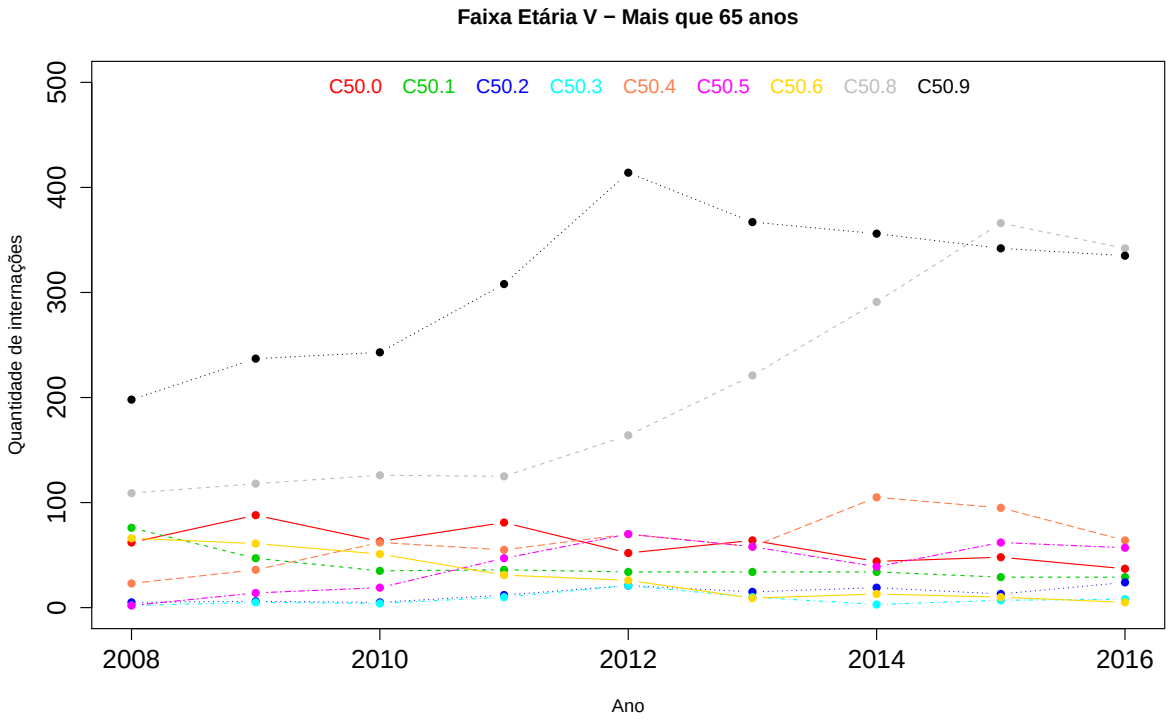


Figura 28 – Internações por câncer de mama na 4ª faixa etária separadas pelo CID entre 2008 e 2016 no Paraná.

A Figura 29 mostra as internações, por faixa etária, separados pelo CID. Nota-se que a distribuição das internações por faixa etária é aproximadamente constante para todos os CID. Porém, nota-se diferenças entre a quantidade de ocorrências dos CIDs na segunda e quinta faixas etárias. As proporções de internações mudam de acordo com o CID.

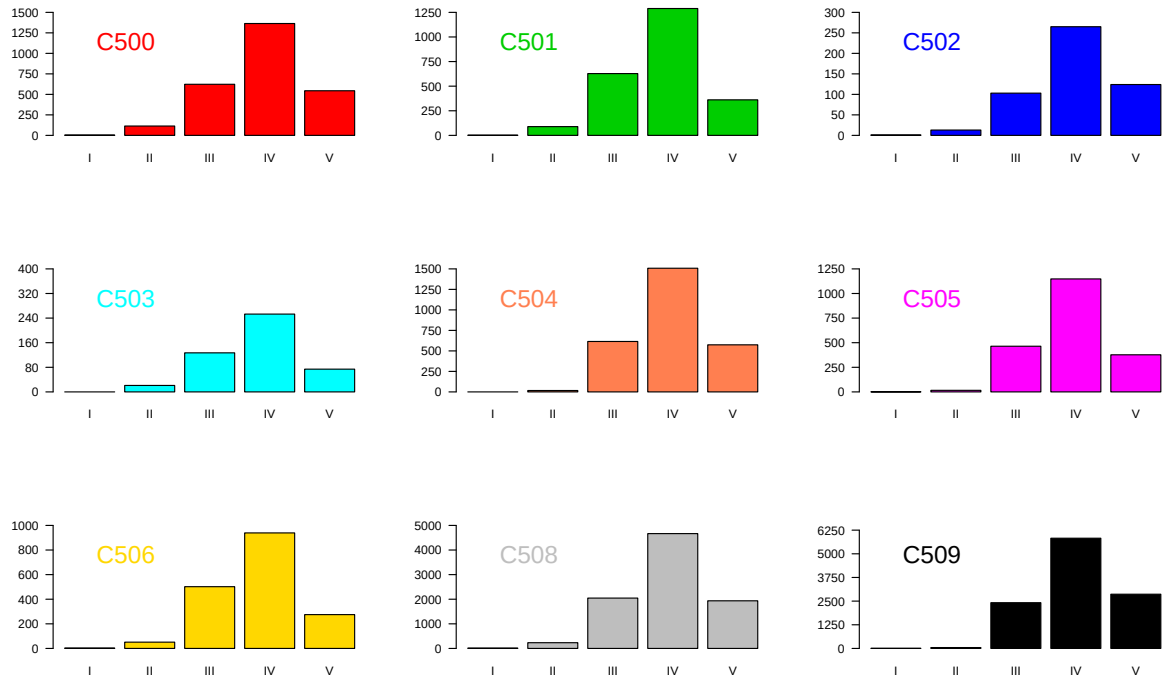


Figura 29 – Total de internações por câncer de mama por faixa etária separados por CID entre 2008 e 2016 no Paraná.

Assim como na Figura 29, a Figura 30, mostra que não há grandes diferenças entre as Macro Regionais quanto à faixa etária com maior número de internações.

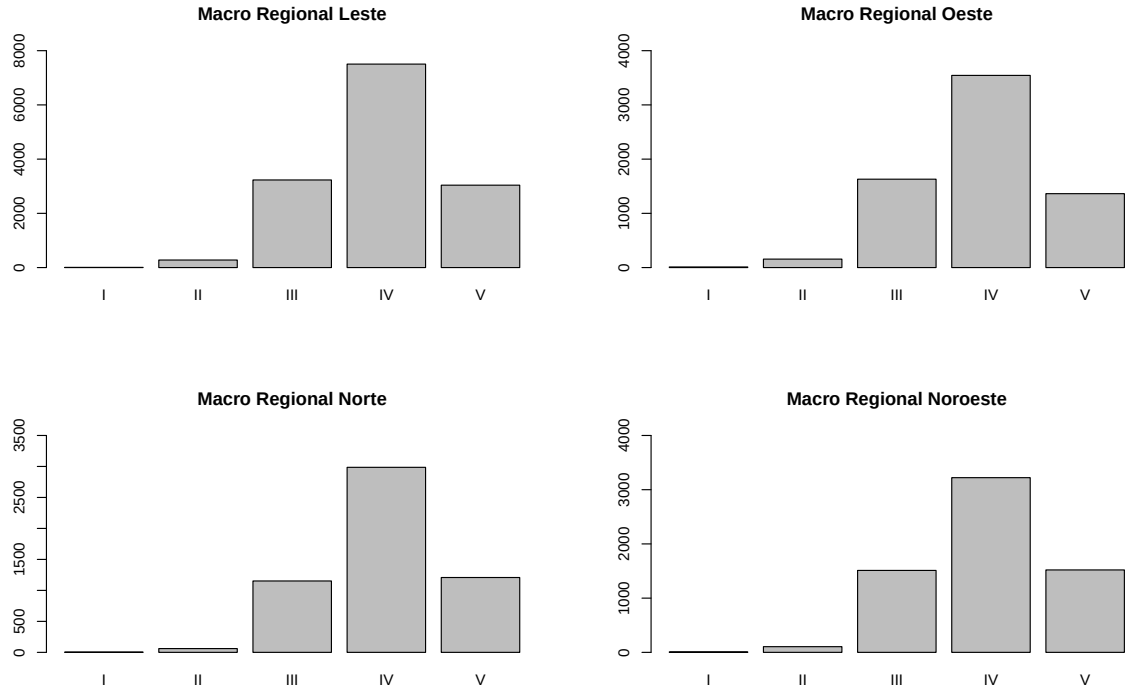


Figura 30 – Total de internações por câncer de mama por faixa etária separado por Macro Regional do estado do Paraná entre 2008 e 2016.

7.1.6 Raça

A variável raça presente no banco de dados segue os padrões do IBGE e está dividida em cinco categorias: branca, negra, parda, amarela e indígena. Além disso há a possibilidade da internação não ter informação sobre a raça sendo então, indicado como "Sem Informação".

A caracterização total das internações por raça é apresentada no Figura 31 onde é notável predominância da raça branca com 25844 (79.42%). As internações nos quais não foi informada a raça do paciente totalizam 2964 (9.11%) e, segundo a Figura 32 percebe-se uma grande diminuição dos registros sem a informação de raça de 784 casos em 2008 para 72 casos em 2016 (90.82% de diminuição).

A população indígena foi a menos frequente nas internações com apenas 11 internações (0.03%) durante os nove anos do estudo. Tais internações ocorreram somente nos anos de 2008 (9 casos), 2009 e 2010 com 1 caso.

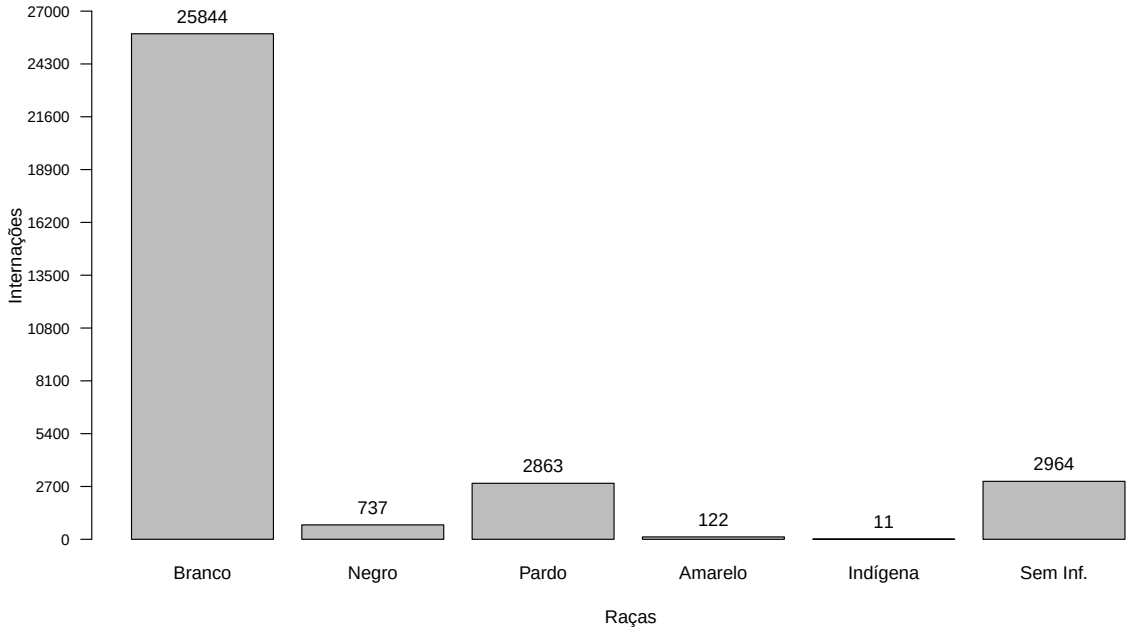


Figura 31 – Total de internações por câncer de mama separados por raça entre 2008 e 2016 no Paraná.

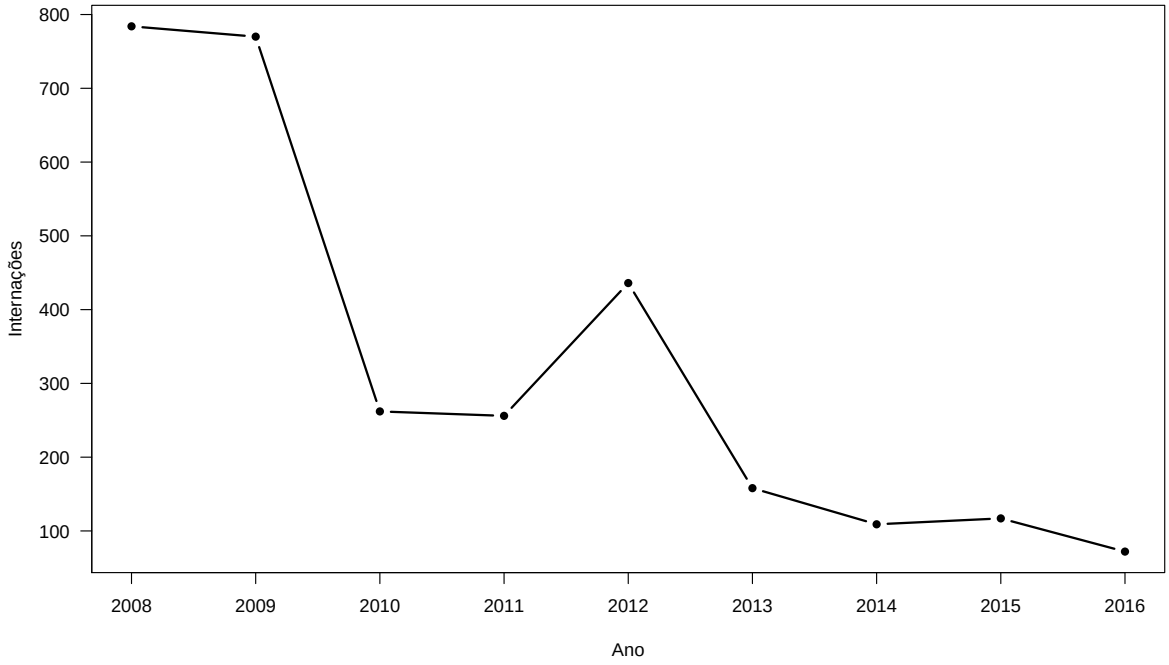


Figura 32 – Internações por câncer de mama sem raça informada entre 2008 a 2016 no Paraná.

Quando separado por Macro Regional, as divisões por raça se mantém aproximadamente as mesmas para as 4 Macro Regionais, com o predomínio sempre da raça branca conforma a Figura 33. Nota-se que não aconteceram internações por câncer de mama em indíginas nas Macro Regionais Oeste e Noroeste do estado do Paraná entre 2008 a 2016.

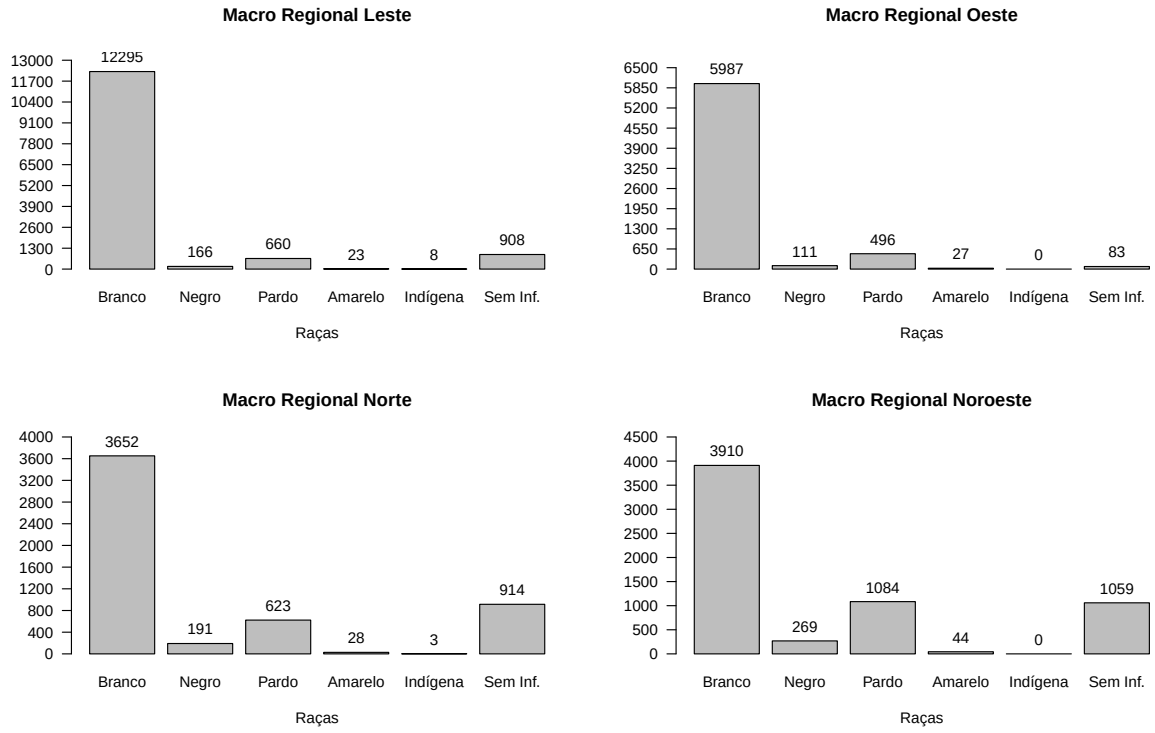


Figura 33 – Total de internações por câncer de mama separados por raça e por Macro Regional no Paraná entre 2008 e 2016.

Analisando os dados pela classificação do câncer de mama segundo o CID, a Figura 34 mostra que para todas as topografias do câncer de mama, a distribuição entre as raças se mantém parecidas.

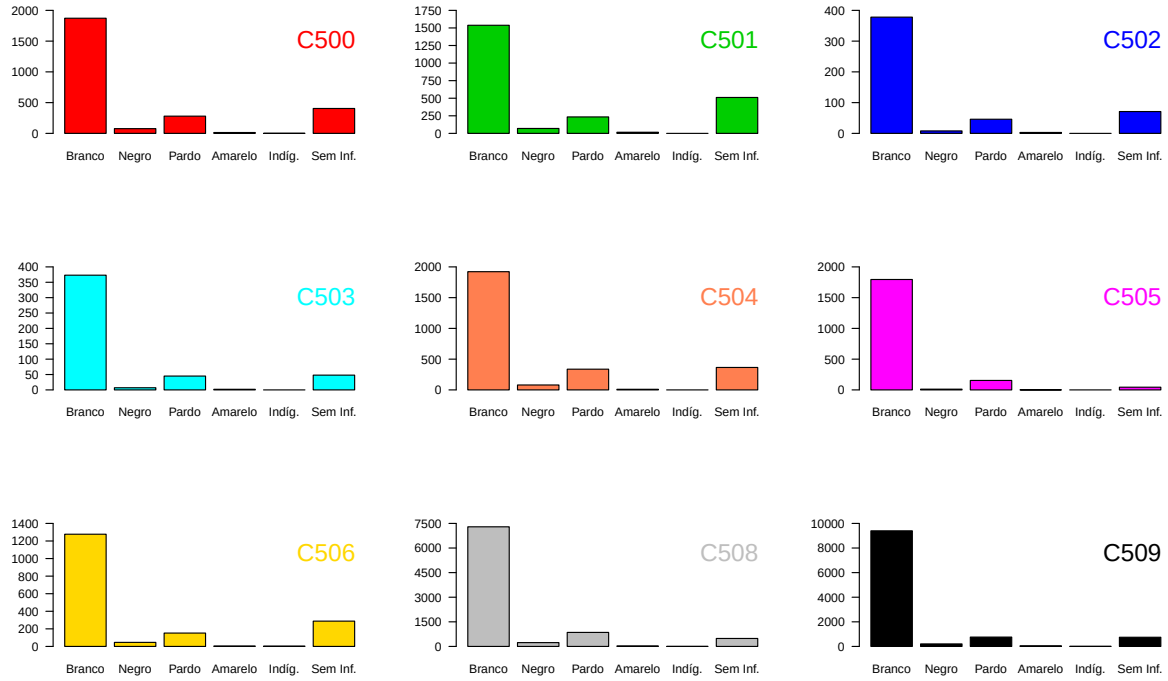


Figura 34 – Total de internações por câncer de mama separados por raça e por CID entre 2008 e 2016 no Paraná.

7.1.7 Valor gasto por internação

Os valores obtidos para os gastos das internações por câncer de mama no estado do Paraná entre 2008 a 2016, através do DATASUS, estão em dólares americanos (US$).

A Tabela 10 e a Figura 35 mostram o resumo dos valores obtidos por internação individual e por ano. Entre 2008 e 2012, nota-se pouca variação entre as médias obtidas (diferença máxima de 107.10 dólares). Porém, a partir de 2013 houve um brusco crescimento do valor (209.63% entre 2012 e 2013).

Assim como o valor por internação, o valor da soma total gasta sofre um aumento de 206.97% entre 2012 a 2013, conforme a Figura 36. A Tabela e a Figura também mostram que a partir de 2015, os valores voltam a diminuir até chegarem a uma média de aproximadamente 650.00 dólares por internação.

Tabela 10 – Valores gastos em dólares por internação de câncer de mama entre 2008 a 2016 no Paraná.

ANO	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2008	17.33	190.10	293.40	374.20	519.20	3457.00
2009	18.65	206.70	309.60	426.70	556.10	3072.00
2010	24.43	229.30	361.60	481.30	638.80	5905.00
2011	22.30	227.00	367.70	474.40	642.30	4907.00
2012	10.16	206.90	350.00	419.40	543.10	5524.00
2013	8.93	215.00	761.10	879.20	1365.00	8766.00
2014	7.82	190.20	951.90	906.20	1361.00	8179.00
2015	5.13	145.00	656.90	641.90	971.90	4628.00
2016	5.25	146.40	677.40	690.40	1022.00	4978.00

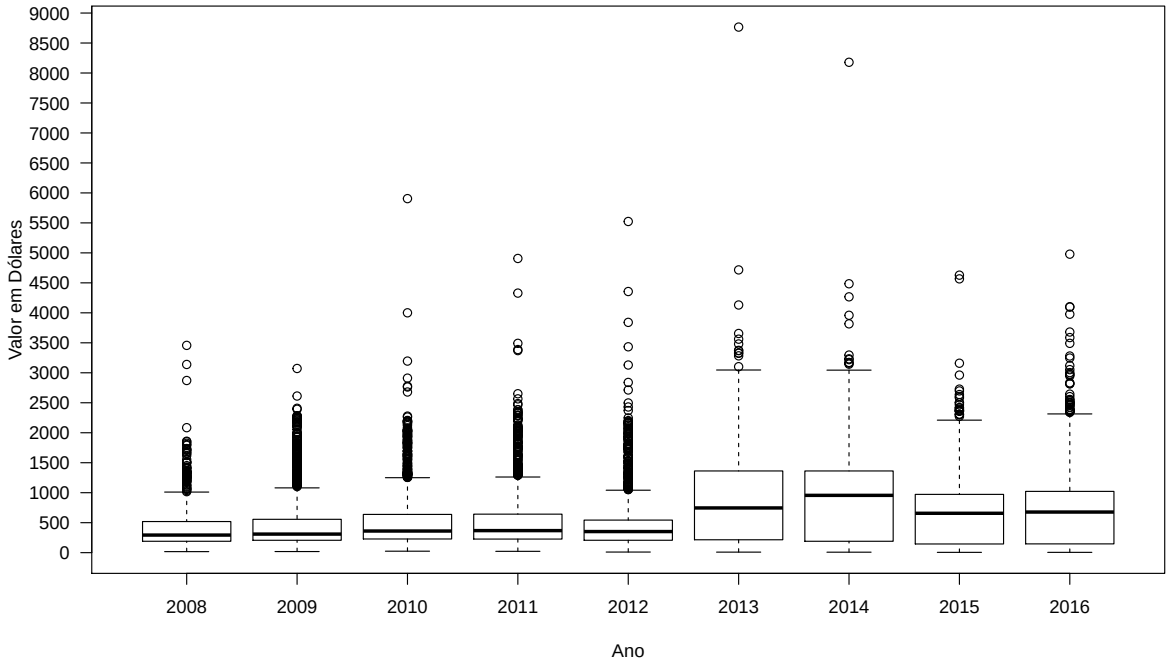


Figura 35 – Valores gastos por ano em dólares entre 2008 e 2016 no Paraná.

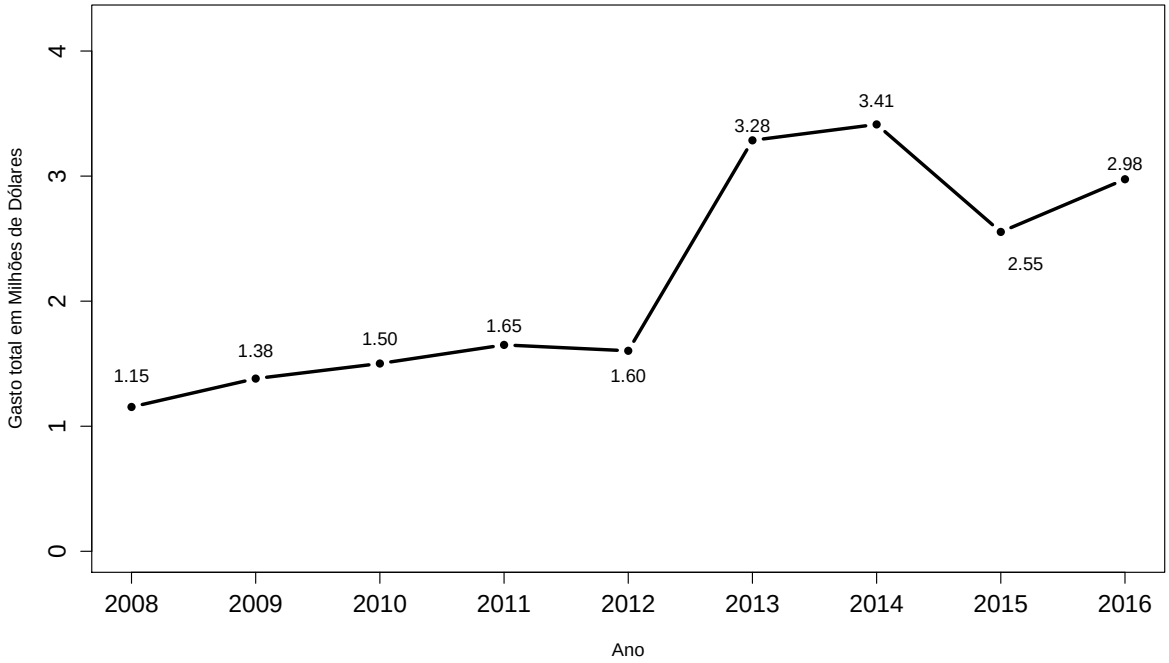


Figura 36 – Soma total de gastos em dólares para internações de câncer de mama entre 2008 e 2016 no Paraná.

Considerando valores gastos para as classificações do CID na Figura 37, notamos que, mesmo contendo os dois maiores *outliers*, os valores de C50.0 e C50.9 se mantém com uma média pequena, e com uma variação menor, juntamente com o C50.6. Já nos demais, aparecem com médias e variações maiores.

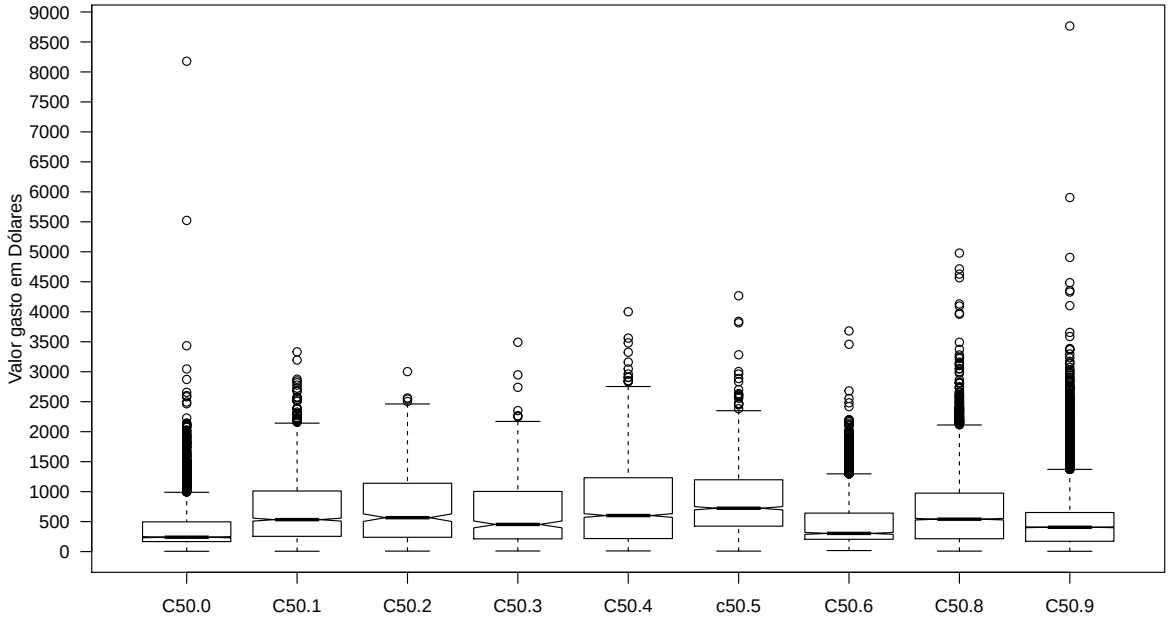


Figura 37 – Valor gasto em dólares por internação de câncer de mama separados por CID entre 2008 e 2016 no Paraná.

Diferentemente da Figura 23, onde a quantidade de internações é maior na quarta faixa etária, quando o valor é tratado, isso muda. Quanto maior a idade, a variação do gasto aumenta e surgem *outliers* maiores. Os 3 maiores valores gastos em uma internação estão na última faixa etária.

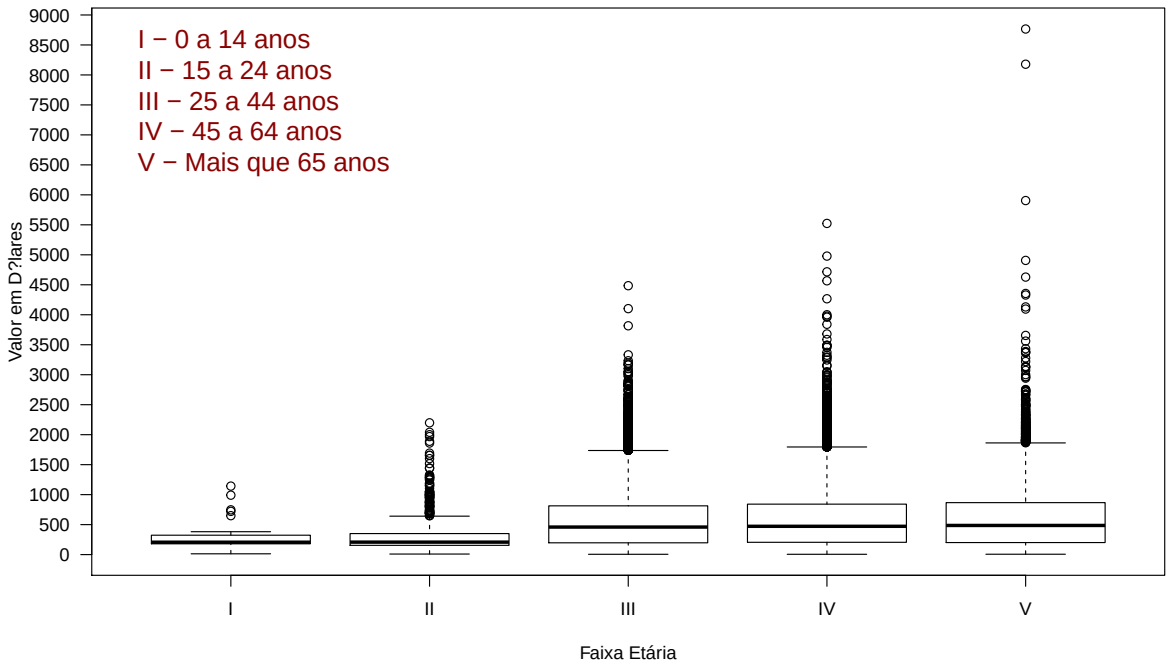


Figura 38 – Valores gastos em dólares por internação de câncer de mama separados por faixa etária entre 2008 e 2016 no Paraná.

A Tabela 11 mostra os valores gastos separados por Macro Regional, e a Figura 39 apresenta os valores gastos por regional.

Tabela 11 – Medidas descritivas de posição do valor gasto em dólares por internação de câncer de mama separados por Macro Regional entre 2008 e 2016 no Paraná.

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
Macro Leste	5.25	203.30	483.30	588.00	788.00	5524.00
Macro Oeste	7.82	207.60	456.90	613.30	841.30	8179.00
Macro Norte	5.13	184.90	434.80	556.80	826.80	8766.00
Macro Noroeste	5.84	195.60	398.50	647.10	984.20	4567.00

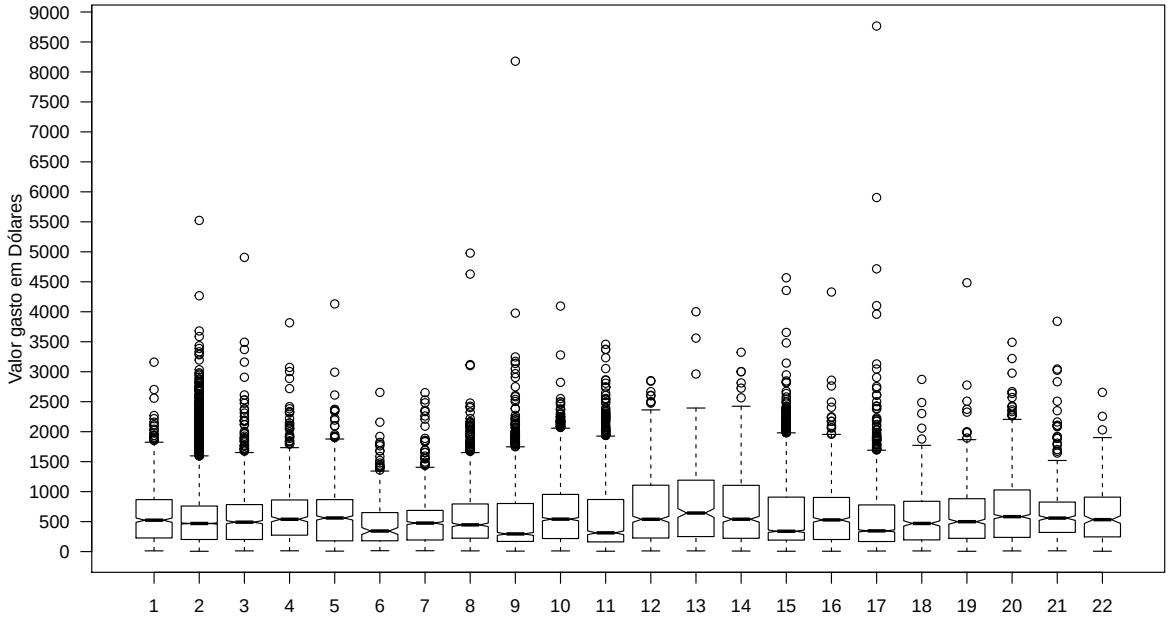


Figura 39 – Total gasto em dólares por internação de câncer de mama separados por Regional de Saúde entre 2008 e 2016 no Paraná.

Como o foco do estudo não é os gastos realizados com as internações por câncer de mama, não foi realizado uma pesquisa mais profunda sobre o assunto. Mesmo assim, entrando em contato com a Secretaria de Saúde da cidade de Maringá e com a Regional de Saúde sobre esse aumento nos gastos, não souberam nos informar o motivo de tal crescimento abrupto nos gastos entre 2012 e 2013.

7.2 O Modelo

7.2.1 Modelo Multinomial

Iniciou-se a análise através de um modelo multinomial somente com efeitos fixos, cuja variável resposta são as ocorrências do câncer de mama segundo a Classificação Internacional de Doenças - CID, e as Macro Regionais juntamente com a Raça e a Faixa Etária como variáveis preditoras. Utilizou-se o CID C50.9, a Macro Regional Leste, a Raça Branca e a Faixa Etária "IV" como *baselines*, pois o CID C50.9 representa o câncer de mama registrado como "sem maiores informações", a Macro Regional Leste é a região que contém a capital do estado - Curitiba, e de maior população, a Raça Branca que é predominante nos dados e na população, e a Faixa Etária "IV" (45 a 64) que, de acordo com o INCA (2016) é a faixa etária de maior predominância de ocorrência de câncer de mama.

O modelo utilizado, como descrito em 6.20, é dado por

$$\text{logit}\left(\frac{CID_{rjlc}}{CID_{rjl9}}\right) = \beta_{0c} + \beta_{1c}\text{Raça}_r + \beta_{2c}\text{MR}_J + \beta_{3c}\text{Idade}_l, \tag{7.1}$$

onde

- β_{0c} representa o valor médio para a categoria c da variável resposta ($c = 1, \cdots, 8$);
- $\beta_{1c}\text{Raça}_r$ representam os efeitos fixos para a variável Raça para a categoria c tal que $r = 1, \cdots, 6$;
- $\beta_{2c}\text{MR}_J$ representam os efeitos fixos para a variável Macro Regional para a categoria c com $j = 1, \cdots, 4$;
- $\beta_{3c}\text{Idade}_l$ representam os efeitos fixos para a variável Faixa Etária para a categoria c tal que $l = 1, \cdots, 5$;
- $c = 1, \cdots, 8$ representam as 8 funções $logit$ que serão utilizadas quando temos o CID C50.9 como *baseline* para a comparação de cada categoria.

Como consequência, podemos escrever esse modelo como probabilidades, conforme descrito em 6.19, da seguinte maneira

$$CID_{rjlc} = \frac{\exp\left\{\beta_{0c} + \beta_{1c}\text{Raça}_r + \beta_{2c}\text{MR}_J + \beta_{3c}\text{Idade}_l\right\}}{1 + \sum_{t=1}^{8} \exp\left\{\beta_{0t} + \beta_{1t}\text{Raça}_r + \beta_{2t}\text{MR}_J + \beta_{3t}\text{Idade}_l\right\}}. \tag{7.2}$$

Devido a grande quantidade de combinações possíveis para as 8 categorias da variável resposta, juntamente com as Raças, as Macro Regionais e as Faixas Etárias, a Tabela 12 apresenta somente os valores das estimativas dos parâmetros, assim como seus

respectivos p-valores, e ainda as *odds ratio* com seus intervalos com 95% de confiança, que apresentaram significância estatística com p-valor abaixo de 0.05. Podemos notar, de maneira geral, que para um mesmo CID, existem fatores de proteção para algumas regionais e fatores de risco para outras, o mesmo acontece com a raça e com as faixas etárias.

Alguns pontos importantes são, para o CID C50.4, o alto fator de risco para as Macro Regionais Noroeste, Norte e Oeste quando comparadas com a Macro Regional Leste, e para o CID C50.5 o alto fator de proteção para as Macro Regionais Noroeste, Norte e Oeste quando comparadas com a Macro Regional Leste. Ou seja, a ocorrência do câncer de mama do tipo C50.4 é mais frequênte nas Macro Regionais Noroeste, Norte e Oeste quando comparadas com a Leste, e a ocorrência do câncer de mama do tipo C50.5 é menos frequênte nas Macro Regionais Noroeste, Norte e Noroeste, quando comparadas com a Leste.

Nota-se também que a Raça Pardo, quando comparada com a Raça Branca, é um fator de risco para os CIDs C50.0, C50.1, C50.4, C50.5 e C50.8. Ou seja, a ocorrência de câncer de mama com esses CIDs para mulheres de Raça Parda são mais frequentes, quando comparadas com mulheres de Raça Branca. Outro fator importante é que para todos os CIDs, as ocorrências do câncer de mama cujas informações de raça não foram registradas, se apresentaram como estatisticamente significantes.

Quando se analisam as faixas etárias, nota-se que a Faixa Etária “II”, que contempla mulheres entre 15 e 24 anos, quando comparadas com o CID C50.9, apresenta fatores de risco em todos os demais CIDs com exceção dos CIDs C50.4 e C50.5. Ou seja, devido ao fato do CID de comparação ser o C50.9 que é representado como “sem maiores informações”, isso indica que as mulheres presentes nesta faixa etária possuem um melhor diagnóstico, sendo indicadas um CID específico. A proporção de ocorrência dessa faixa etária é maior nos demais CIDs do que no CID utilizado como *baseline*.

A Faixa Etária “V” se comporta de maneira contrária à Faixa Etária “II”. Com exceção do CID C50.2, a Faixa Etária “V” apresenta um fator de proteção quando comparado com o CID C50.9. Ou seja, para mulheres acima de 65 anos, o diagnóstico não é tão preciso, e sua especificação do tipo de câncer de mama é reduzida. A proporção de ocorrência dessa faixa etária é menor nos demais CIDs do que no utilizado como *baseline*.

As análises estatísticas foram realizadas com a utilização do *software* `SAS (Versão 9.4)` através da `PROC LOGISTIC`.

Tabela 12 – Estimativas, erros padrão, p-valores e *Odds Ratio* com intervalo de 95% de confiança.

CID		Estim.	Erro Padrão	P-valor	Odds Ratio	Limite Inf.	Limite Sup.
C50.0	**MRNO**	0.3270	0.0591	<.0001	**1.387**	1.235	1.557
	MRN	-0.2619	0.0726	0.0003	0.770	0.668	0.887
	MRO	-0.2368	0.0614	0.0001	0.789	0.700	0.890
	Negro	0.5685	0.1378	<.0001	**1.766**	1.348	2.313
	Pardo	0.5274	0.0765	<.0001	**1.695**	1.459	1.968
	Sem Info	0.9310	0.0693	<.0001	**2.537**	2.215	2.906
	Faixa II	2.3589	0.1780	<.0001	**10.580**	7.464	14.995
	Faixa V	-0.2174	0.0560	0.0001	0.805	0.721	0.898
C50.1	**MRNO**	0.3387	0.0659	<.0001	**1.403**	1.233	1.596
	MRN	0.7194	0.0621	<.0001	**2.053**	1.818	2.319
	Negro	0.5558	0.1432	<.0001	**1.743**	1.317	2.308
	Pardo	0.4567	0.0817	<.0001	**1.579**	1.345	1.853
	Amarelo	0.6244	0.2958	0.0348	**1.867**	1.046	3.334
	Sem Info	1.2183	0.0667	<.0001	**3.381**	2.967	3.854
	Faixa II	2.2313	0.1860	<.0001	**9.312**	6.467	13.408
	Faixa III	0.1560	0.0551	0.0047	**1.169**	1.049	1.302
	Faixa V	-0.5748	0.0645	<.0001	0.563	0.496	0.639
C50.2	**MRNO**	0.2887	0.1247	0.0206	**1.335**	1.045	1.704
	MRN	0.2774	0.1322	0.0359	**1.320**	1.018	1.710
	MRO	-0.3754	0.1382	0.0066	0.687	0.524	0.901
	Sem Info	0.7099	0.1393	<.0001	**2.034**	1.548	2.672
	Faixa II	1.8601	0.3208	<.0001	**6.424**	3.426	12.047
C50.3	**MRNO**	0.7185	0.1209	<.0001	**2.051**	1.619	2.600
	Sem Info	0.3249	0.1628	0.0459	**1.384**	1.006	1.904
	Faixa II	2.3781	0.2716	<.0001	**10.784**	6.333	18.364
	Faixa V	-0.5429	0.1344	<.0001	0.581	0.447	0.756
C50.4	**MRNO**	1.6717	0.0622	<.0001	**5.321**	4.711	6.011
	MRN	1.7818	0.0623	<.0001	**5.941**	5.258	6.712
	MRO	0.3939	0.0722	<.0001	**1.483**	1.287	1.708
	Pardo	0.2948	0.0728	<.0001	**1.343**	1.164	1.549
	Sem Info	0.3521	0.0717	<.0001	**1.422**	1.236	1.637
	Faixa V	-0.2961	0.0555	<.0001	0.744	0.667	0.829
C50.5	**MRNO**	-1.6434	0.1118	<.0001	0.193	0.155	0.241
	MRN	-1.5029	0.1108	<.0001	0.222	0.179	0.276
	MRO	-2.4049	0.1244	<.0001	0.090	0.071	0.115
	Negro	-0.8522	0.3010	0.0046	0.426	0.236	0.769
	Pardo	0.4275	0.0959	<.0001	**1.533**	1.271	1.850
	Sem Info	-1.0215	0.1587	<.0001	0.360	0.264	0.491
	Faixa V	-0.3874	0.0647	<.0001	0.679	0.598	0.771
C50.6	**MRNO**	0.9224	0.0646	<.0001	**2.515**	2.216	2.855
	MRN	-0.3214	0.0951	0.0007	0.725	0.602	0.874
	Indígena	1.5680	0.7142	0.0281	**4.726**	1.166	19.163
	Sem Info	0.8380	0.0791	<.0001	**2.312**	1.980	2.700
	Faixa II	1.9650	0.2074	<.0001	**7.135**	4.752	10.713
	Faixa III	0.2297	0.0609	0.0002	**1.258**	1.117	1.418
	Faixa V	-0.5561	0.0728	<.0001	0.573	0.497	0.661
C50.8	**MRNO**	0.9751	0.0427	<.0001	**2.651**	2.439	2.883
	MRN	1.0608	0.0434	<.0001	**2.889**	2.653	3.145
	MRO	0.8795	0.0371	<.0001	**2.410**	2.241	2.591
	Pardo	0.1266	0.0539	0.0189	**1.135**	1.021	1.261
	Amarelo	-0.5332	0.2475	0.0312	0.587	0.361	0.953
	Sem Info	-0.3432	0.0625	<.0001	0.709	0.628	0.802
	Faixa II	1.9139	0.1639	<.0001	**6.780**	4.917	9.348
	Faixa V	-0.1761	0.0361	<.0001	0.839	0.781	0.900

Devido a análise descritiva apresentar pequenas diferenças nas faixas etárias em relação ao CID, julgou-se importante verificar a necessidade da implementação dessa variável no modelo. Para isso, utilizou-se o seguinte modelo para comparação.

$$\text{logit}\left(\frac{CID_{rjc}}{CID_{rj9}}\right) = \beta_{0c} + \beta_{1c}\text{Raça}_r + \beta_{2c}\text{MR}_j, \tag{7.3}$$

onde

- β_{0c} representa o valor médio para a categoria c da variável resposta ($c = 1, \cdots, 8$);
- $\beta_{1c}\text{Raça}_r$ representam os efeitos fixos para a variável Raça para a categoria c ($r = 1, \cdots, 6$);
- $\beta_{2c}\text{MR}_j$ representam os efeitos fixos para a variável Macro Regional para a categoria c ($j = 1, \cdots, 4$);
- $c = 1, \cdots, 8$ representam as 8 funções $logit$ que serão utilizadas quando temos o CID C50.9 como *baseline* para a comparação de cada categoria.

De acordo com os resultados obtidos através do *software* `SAS (Versão 9.4)` com a `PROC LOGISTIC`, para os testes AIC, além do Teste da Razão de Verossimilhança exibidos na Tabela 13, temos que a implementação do efeito fixo relativo à Faixa Etária é realmente mais adequado para o estudo.

Tabela 13 – Valores para os testes AIC e para a Log-verossimilhança através do *software* `SAS (Versão 9.4)`

	AIC	**-2Log L**
Com Faixa Etária	109889.33	109681.33
Sem Faixa Etária	110456.48	110312.48

7.2.2 Modelo Multinível Multinomial na análise das ocorrências de câncer de mama por CID no estado do Paraná entre 2008 e 2016

O modelo multinível proposto para a caracterização das internações por câncer de mama segundo suas topografias (CID-10), e segundo as regionais de saúde do estado do Paraná, entre 2008 e 2016, utilizou como variável resposta as classificações do câncer de mama quanto a sua topografia, e variáveis preditoras como no modelo proposto em 7.1,

- Raça

- – Branco
- – Pardo
- – Indígena
- – Negro
- – Amarelo
- – Não informado

- Macro Regionais/Regionais de Saúde.
  - – Macro Regional Leste (7 Regionais)
  - – Macro Regional Noroeste (5 Regionais)
  - – Macro Regional Oeste (5 Regionais)
  - – Macro Regional Norte (5 Regionais)

- Faixa Etária.
  - – Faixa I - 0 a 14 anos;
  - – Faixa II - 15 a 24 anos;
  - – Faixa III - 25 a 44 anos;
  - – Faixa IV - 45 a 64 anos;
  - – Faixa V - Mais que 65 anos.

Portanto, com essas informações, o modelo multinível multinomial proposto, conforme 6.39, é dado por

$$\text{logit}\left(\frac{CID_{rjlc}}{CID_{rjl9}}\right) = \beta_{0c} + \mathbf{b}_{0k|j} + \beta_{1c}\text{Raça}_r + \beta_{2c}\text{MR}_j + \beta_{3c}\text{Idade}_l. \tag{7.4}$$

onde

- β_{0c} representa o valor médio para a categoria c da variável resposta ($c = 1, \cdots, 8$);
- $\beta_{1c}\text{Raça}_r$ representam os efeitos fixos para a variável Raça para a categoria c ($r = 1, \cdots, 6$);
- $\beta_{2c}\text{MR}_j$ representam os efeitos fixos para a variável Macro Regional para a categoria c ($j = 1, \cdots, 4$);
- $\beta_{3c}\text{Idade}_l$ representam os efeitos fixos para a variável Faixa Etária para a categoria c ($l = 1, \cdots, 5$);
- $b_{0k|j}$ representa o efeito aleatório multinível que indica a hierarquia entre as Regionais de Saúde e as Macro Regionais ($k = 1, \cdots, 22$)

Consequentemente, temos o modelo em função de sua probabilidade, de acordo com 6.38,

$$CID_{rjlc} = \frac{\exp\left\{\beta_{0c} + b_{0k|j} + \beta_{1c}\text{Raça}_r + \beta_{2c}\text{MR}_j + \beta_{3c}\text{Idade}_l\right\}}{1 + \sum_{t=1}^{8}\exp\left\{\beta_{0t} + b_{0k|j} + \beta_{1t}\text{Raça}_r + \beta_{2t}\text{MR}_j + \beta_{3t}\text{Idade}_l\right\}}. \tag{7.5}$$

.

Para esse modelo foram utilizadas as mesmas *baselines* do modelo anterior 7.1: o CID C50.9, Macro Regional Leste, a Raça Branca e a Faixa Etária "IV".

As Tabelas 14 e 15 apresentam os valores das estimativas, seus erros padrão e seus p-valores, as *odds ratio* com seu intervalo de 95% confiança, para dois métodos de estimação diferentes. As informações foram obtidas através do software `SAS` versão 9.4 e a utilização da `PROC GLIMMIX` (SCHABENBERGER, 2005).

A Tabela 14 exibe as informações obtidas através do método de estimação Quadratura de Gauss-Hermite - QGH, e a Tabela 15 mostra as informações obtidas através do método de estimação **`Aproximação de Laplace - LAP`**.

Para as informações obtidas com o método de estimação Quadratura de Gauss-Hermite - QGH, foi utilizada a opção `METHOD = QUAD`, e para as informações obtidas com o método de estimação `Aproximação de Laplace - LAP`, foi utilizada a opção `METHOD = LAPLACE`.

Analisando as Tabelas 14 e 15 por CID, temos algumas diferenças entre os valores obtidos.

Deve-se recordar que a quantidade de internações com a raça indígena é muito pequena relativamente com as demais, apresentando somente 11 casos entre os 32541 totais.

- CID C50.0

  Embora os valores das estimativas, assim como os valores das *odds ratio* estejam muito próximas em ambos os métodos, temos que pelo método LAP, foi indicado como significativo a Raça Negra, enquanto o método QGH não indicou tal informação. Para essa classe, todas as variáveis são fatores de risco quando comparadas com a classe de referência.

- CID C50.1

  Para essa classe da variável resposta, ambos os métodos concordaram com a Macro Regional Norte, Raça Parda e Raça "Sem Informação", obtendo valores próximos para as estimativas e para as *odds ratio*. Porém, no método QGH foi indicada como significativa a Raça Indígena, e no método LAP foi indicada como significativa a Raça Negra. Para essa classe, todas as variáveis (com exceção da Raça indígena) foram consideradas fatores de risco para a ocorrência do câncer de mama para esse tipo de CID.

- CID C50.2

Nesta categoria, ambos os métodos indicaram a Raça “Sem Informação” como significativa apresentando valores de estimativas e *odds* similares, porém, o método QGH também apresentou a variável Raça Indígena como significativa, enquanto o método LAP não o fez.

- CID C50.3

  Para este CID, os métodos divergiram sobre as variáveis significativas. Enquanto que o método QGH apresentou a Raça Indígena como significativa, o método LAP apresentou a Raça “Sem Informação”.

- CID C50.4

  Nesta classe, os métodos concordaram com as informações sobre as Macros Regionais Noroeste e Norte, e com a Raça Pardo, indicando valores das estimativas e *odds ratio* similares, porém o método QGH também indicou a Raça indígena como significativa. Para essa classe, com exceção da variável Raca indígena, todas as variáveis são apresentadas como fatores de risco.

- CID C50.5

  Para a categoria CID C50.5, os métodos concordaram com as informação sobre as Macros Regionais Noroeste, Norte e Oeste, e com as Raças Negro, Pardo e “Sem Informação”. Porém, o método QGH também considerou como significativas as variáveis Raças Amarelo e Indígena. Deve-se notar também que essa classe, com exceção à Raça Pardo, todas as demais variáveis foram apresentadas como fatores de proteção quando comparadas com as variáveis de referência, diferenciando essa classe da variável resposta com as demais.

- CID C50.6

  Para essa classe, o método QGH apresentou somente a variável Raça Indígena como significativa, enquanto que o método LAP apresentou, além da variável indígena, as variáveis Macro Regional Noroeste e Raça “Sem Informação” como significativas. Além disso, todas as variáveis são apresentadas como fatores de risco.

- CID C50.8

  Os métodos concordaram, para essa categoria, com as Macros Regionais Noroeste, Oeste e Norte e as Raças pardo e “Sem Informação”. O método QGH apresentou como significativo a variável Raça indígena. Para essa classe, todas as Macro Regionais presentes são apresentadas como fatores de risco, assim como a Raça Pardo. Já a Raça indígena e “Sem Informação” são fatores de proteção.

Tabela 14 – Estimativas, p-valores, *odds ratio* com I.C. 95% - MÉTODO DE ESTIMAÇÃO = QUADRATURA.

CID		Estim.	Erros Padrão	P-valor	Odds Ratio	Limite Inf.	Limite Sup.
C50.0	**Pardo**	0.2201	0.0920	0.0168	**1.246**	1.040	1.493
	Sem Info	0.7681	0.3163	0.0152	**2.156**	1.160	4.007
	Faixa II	2.2396	0.1834	<.0001	**9.390**	6.554	13.453
	Faixa V	-0.2126	0.0436	<.0001	0.808	0.742	0.881
C50.1	**MRN**	0.7624	0.2656	0.0047	**2.143**	1.268	3.623
	Pardo	0.2482	0.0952	0.0091	**1.282**	1.064	1.545
	Indígena	-6.7383	0.7413	<.0001	0.001	<0.001	0.005
	Sem Info	1.1030	0.1898	<.0001	**3.013**	2.077	4.371
	Faixa II	2.1562	0.2350	<.0001	**8.638**	5.450	13692
	Faixa III	0.1599	0.0570	0.0050	**1.173**	1.049	1.312
	Faixa V	-0.5735	0.07095	<.0001	0.564	0.490	0.648
C50.2	**Indígena**	-6.9488	0.7140	<.0001	0.001	<0.001	0.004
	Sem Info	0.5496	0.2242	0.00142	**1.732**	1.116	2.689
	Faixa II	1.8638	0.2211	<.0001	**6.448**	4.181	9.946
C50.3	**Indígena**	-6.7566	0.7488	<.0001	0.001	<0.001	0.005
	Faixa I	-3.7790	0.5950	<.0001	0.023	0.007	0.073
	Faixa II	2.3552	0.3371	<.0001	**10.540**	5.444	20.405
	Faixa V	-0.5166	0.1065	<.0001	0.597	0.484	0.735
C50.4	**MRNO**	1.6437	0.5522	0.0034	**5.174**	1.737	15.410
	MRN	1.39772	0.5929	0.0198	**4.044**	1.253	13.054
	Indígena	-6.5223	0.9209	<.0001	0.001	<0.001	0.009
	Faixa I	-3.8270	0.6757	<.0001	0.022	0.006	0.082
	Faixa V	-0.2987	0.0484	<.0001	0.742	0.675	0.816
C50.5	**MRNO**	-1.3846	0.4764	0.0042	0.250	0.098	0.642
	MRN	-1.5133	0.7501	0.0455	0.220	0.050	0.970
	MRO	-2.2570	0.5781	0.0001	0.105	0.033	0.328
	Negro	-0.9403	0.1747	<.0001	0.391	0.277	0.550
	Pardo	0.3875	0.1192	0.0011	**1.473**	1.166	1.861
	Amarelo	-1.7700	0.6787	0.0091	0.170	0.045	0.644
	Indígena	-7.2005	0.6244	<.0001	<0.001	<0.001	0.003
	Sem Info	-1.1547	0.1866	<.0001	0.315	0.219	0.454
	Faixa II	0.4204	0.2116	0.0470	**1.522**	1.006	2.305
	Faixa V	-0.3819	0.0997	0.0001	0.683	0.561	0.830
C50.6	**MRNO**	0.8326	0.4096	0.0439	**2.299**	1.023	5.166
	Indígena	1.8968	0.0823	<.0001	**6.665**	5.671	7.832
	Faixa II	2.0019	0.1774	<.0001	**7.403**	5.228	10.482
	Faixa III	0.2406	0.0707	0.0007	**1.272**	1.107	1.461
	Faixa V	-0.5554	0.0768	<.0001	0.574	0.494	0.667
C50.8	**MRNO**	0.9310	0.2919	0.0017	**2.537**	1.425	4.517
	MRN	0.7676	0.3845	0.0478	**2.155**	1.008	4.607
	MRO	0.8983	0.3105	0.0044	**2.455**	1.329	4.536
	Pardo	0.1759	0.0887	0.0475	**1.192**	1.002	1.419
	Indígena	-2.0810	1.0328	0.0439	0.125	0.016	0.945
	Sem Info	-0.5009	0.1846	0.0066	0.606	0.422	0.870
	Faixa II	1.9864	0.2143	<.0001	**7.290**	4.790	11.094
	Faixa V	-0.1684	0.0446	0.0002	0.845	0.774	0.922

Tabela 15 – Estimativas, p-valores, *odds ratio* com I.C. 95% - MÉTODO DE ESTIMAÇÃO = LAPLACE.

CID	Macro/ Raça	Estim.	Erros Padrão	P-valor	Odds Ratio	Limite Inf.	Limite Sup.
C50.0	**Negro**	0.3917	0.1400	0.0052	**1.479**	1.124	1.947
	Pardo	0.2195	0.07831	0.0051	**1.245**	1.068	1.452
	Sem Info	0.7672	0.0708	<.0001	**2.154**	1.875	2.474
	Faixa II	2.2388	0.1794	<.0001	**9.382**	6.601	13.335
	Faixa V	-0.2129	0.0566	0.0002	0.808	0.723	0.903
C50.1	**MRN**	0.7619	0.3383	0.0258	**2.142**	1.098	4.181
	Negro	0.3609	0.1453	0.0130	**1.435**	1.079	1.907
	Pardo	0.2474	0.08316	0.0029	**1.281**	1.088	1.507
	Sem Info	1.1024	0.06843	<.0001	**3.011**	2.633	3.444
	Faixa II	2.1558	0.1869	<.0001	**8.635**	5.987	12.455
	Faixa III	0.1601	0.0555	0.0039	**1.174**	1.053	1.309
	Faixa V	-0.5733	0.0649	<.0001	0.564	0.496	0.640
C50.2	**Sem Info**	0.5485	0.1424	<.0001	**1.731**	1.309	2.288
	Faixa II	1.8638	0.3215	<.0001	**6.448**	3.434	12.109
C50.3	**Faixa II**	2.3550	0.2759	<.0001	**10.539**	6.136	18.099
	Faixa V	-0.5175	0.1353	<.0001	0.596	0.457	0.777
C50.4	**MRNO**	1.6393	0.5435	0.0030	**5.152**	1.759	15.085
	MRN	1.3976	0.5448	0.0113	**4.045**	1.378	11.875
	Pardo	0.2142	0.0760	0.0048	**1.239**	1.067	1.438
	Faixa II	0.6276	0.2892	0.0300	**1.873**	1.063	3.302
	Faixa V	-0.2981	0.0566	<.0001	0.742	0.664	0.829
C50.5	**MRNO**	-1.3856	0.5978	0.0219	0.250	0.077	0.815
	MRN	-1.5126	0.6053	0.0136	0.220	0.067	0.729
	MRO	-2.2572	0.5997	0.0002	0.105	0.032	0.342
	Negro	-0.9419	0.3018	0.0018	0.390	0.216	0.704
	Pardo	0.3880	0.0974	<.0001	**1.474**	1.218	1.784
	Sem Info	-1.1558	0.1592	<.0001	0.315	0.230	0.430
	Faixa V	-0.3822	0.0653	<.0001	0.682	0.600	0.776
C50.6	**MRNO**	0.8324	0.2818	0.0037	**2.299**	1.317	4.013
	Indígena	1.9047	0.7251	0.0086	**6.717**	1.622	27.822
	Sem Info	0.6630	0.0813	<.0001	**1.941**	1.655	2.276
	Faixa II	2.0043	0.2079	<.0001	**7.421**	4.937	11.154
	Faixa III	0.2408	0.0612	<.0001	**1.272**	1.128	1.435
	Faixa V	-0.5550	0.0731	<.0001	0.574	0.497	0.663
C50.8	**MRNO**	0.9300	0.3287	0.0053	**2.535**	1.323	4.854
	MRN	0.7683	0.3292	0.0210	**2.156**	1.125	4.133
	MRO	0.8965	0.3270	0.0069	**2.451**	1.284	4.678
	Pardo	0.1758	0.0562	0.0018	**1.192**	1.068	1.331
	Sem Info	-0.5014	0.0644	<.0001	0.606	0.534	0.687
	Faixa II	1.9854	0.1655	<.0001	**7.282**	5.264	10.073
	Faixa III	0.0732	0.0373	0.0499	**1.076**	1.001	1.158
	Faixa V	-0.1683	0.0369	<.0001	0.845	0.786	0.909

7.2.3 Teste da Razão de Verossimilhança Restrita

Para a comparação entre o modelo multinomial com o modelo multinomial multinível, foi utilizado o Teste da Razão de Máxima Verossimilhança Restrita como descrito em 6.5.1.1.

De acordo com os valores obtidos pelo `SAS`, temos:

$$\begin{aligned} TRVR &= 2l(\theta_1) - 2l(\theta_0) \\ &= 109681.33 - 105887.2 \\ &= 3794.13 \end{aligned} \tag{7.6}$$

onde θ_1 representa o modelo multinomial e θ_0 representa o modelo multinível, já com o acréscimo do efeito aleatório.

Com isso, temos $P(\chi^2_1 > G = 3794.13)$ e portanto, rejeitamos a hipótese nula. Logo, pelo TRVR, temos que a implementação do efeito aleatório multinível é significativa para o modelo.

7.2.4 Coeficiente de Correlação Intraclasse

A partir das informações obtidas através do *software* `SAS (Versão 9.4)` com `PROC GLIMMIX`, e conforme as ideias vistas em 6.5.1.2, temos os valores do CCI para cada categoria do CID

- C50.0 $= \frac{0.5660}{0.5660+3.29} = 0.15$;
- C50.1 $= \frac{0.3039}{0.3039+3.29} = 0.09$;
- C50.2 $= \frac{0.3738}{0.3738+3.29} = 0.10$;
- C50.3 $= \frac{0.5185}{0.5185+3.29} = 0.14$;
- C50.4 $= \frac{0.8293}{0.8293+3.29} = 0.20$;
- C50.5 $= \frac{0.9654}{0.9654+3.29} = 0.23$;
- C50.6 $= \frac{0.2026}{0.2026+3.29} = 0.06$;
- C50.8 $= \frac{0.3023}{0.3023+3.29} = 0.08$.

A partir de tais valores, podemos notar que, dentro de cada categoria, uma certa porcentagem das internações das Regionais de Saúde dentro de uma mesma Macro Regional apresentam dependência, corroborando com valores encontrados na literatura citadas em 6.5.1.2, verificando a utilização de um modelo que se adeque a essa característica.

7.2.5 Análise de Resíduos

A `PROC GLIMMIX` somente disponibiliza os valores preditos para análise e assim, utilizando como valores observados a proporção de ocorrências de cada categoria do CID, pelo total de ocorrências no banco de dados, e juntamente com os valores preditos, temos a Figura 40 realizada através do *software* `R (Versão 3.3.1)`, que exibe a dispersão dos valores preditos pelos valores resíduais.

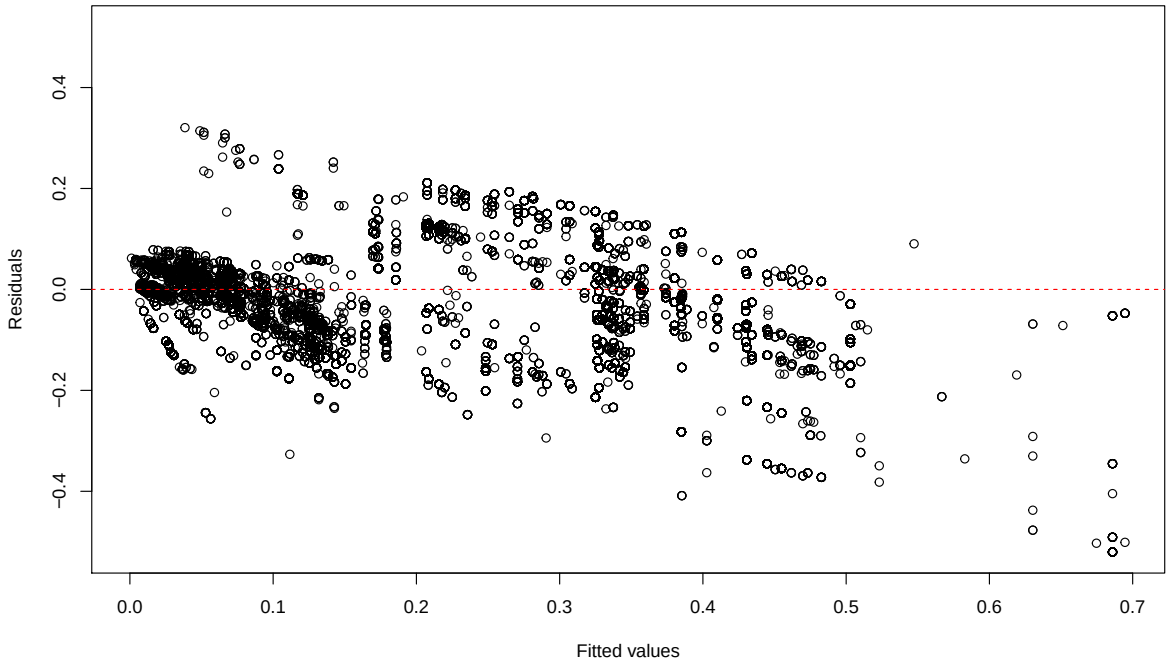


Figura 40 – Valores Preditos Vs Valores Residuais.

Nota-se a partir deste gráfico, a variabilidade dos resíduos, com sua dispersão entre -0.5 e 0.4, e uma leve tendência a valores negativos para os resíduos. A alta aglomeração dos valores predítos entre 0 e 0.1, corrobora com a proporção de ocorrências entre as categorias da variável resposta.

Através dos valores obtidos com o *software* `SAS`, e transpondo esses valores para o *software* `R (Versão 3.3.1)`, obtivemos a Figura 41, que apresenta o gráfico dos valores teóricos pelos valores amostrais. A partir desta figura, nota-se o bom ajuste do modelo com os dados apresentando uma leve variação nas caudas.

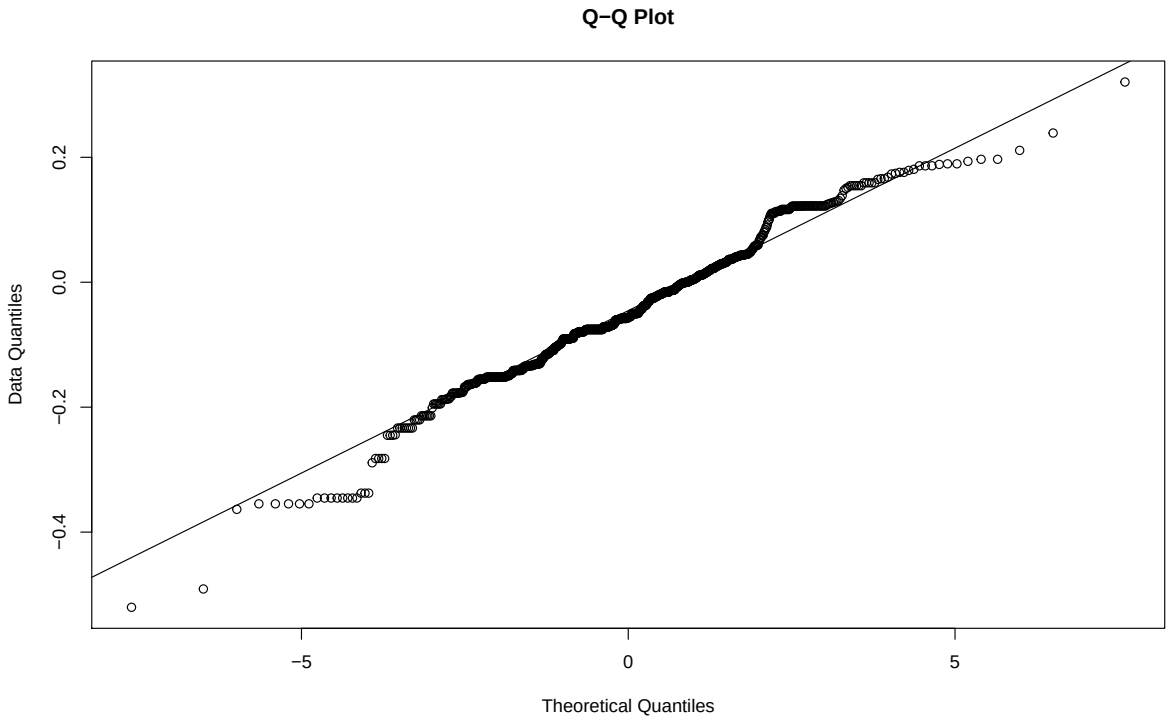


Figura 41 – Valores Teóricos Vs Valores amostrais.

7.3 Discussão e Recomendações

Devido às características básicas do estudo do tipo ecológico, na qual a unidade de análise é um grupo de indivíduos agregados em função de fatores geográficos, a presença de semelhanças entre observações dentro do mesmo grupo ou região devem ser consideradas, pois apresentam uma estrutura de dependência, e essa característica deve ser considerada no modelo estatístico.

Na comparação realizada na Seção 6.5.1.1, verificou-se o melhor ajuste do modelo multinível quando comparado com um modelo de somente efeitos fixos, e com os valores encontrados para os coeficientes de correlação intraclasse na Seção 6.5.1.2, verificou-se o grau de semelhança entre as internações ocorridas em Regionais de Saúde presentes em uma mesma Macro Regional.

Devemos nos atentar que o presente estudo trata-se de uma abordagem preliminar acerca dos detalhes das ocorrências de câncer de mama separados pelo CID, e que devido à falta de material literário abordando esse assunto, maiores informações e estudos mais aprofundados são necessários.

De acordo com o teste Tipo III de efeitos fixos, obtido através do *software* `SAS`, o qual apresenta testes de hipótese para a significância global de cada um dos efeitos fixos especificados no modelo, temos que todos os efeitos fixos presentes, *Macro Regionais,*

Raça e *Faixa Etárias* são significativos de maneira geral. Nos próximos parágrafos, serão exibidos de maneira mais específica e detalhadas as informações obtidas sobre os efeitos fixos.

Levando em consideração a localidade de residência das mulheres internadas, verificamos que as mulheres que residem na **Macro Regional Noroeste**, cujas principais Regionais são Maringá e Umuarama, possuem maiores chances de serem internadas devido a ocorrência do tipos C50.4 e C50.8 de câncer de mama, e uma menor chance de serem internadas devido ao tipo C50.5 de câncer de mama do que as residentes da Macro Regional Leste.

Para a **Macro Regional Oeste** cujas principais Regionais são Cascavel e Foz do Iguaçu, as mulheres residentes nessa Macro Regional apresentam maiores chances de serem internadas devido a ocorrência do tipo C50.8 de câncer de mama, e menores chances de serem internadas devido ao tipo C50.5 de câncer de mama quando comparadas com as mulheres residentes na Macro Regional Leste.

Quando é analisado a **Macro Regional Norte**, cujas principais Regionais são Londrina e Apucarana, as mulheres residentes nessa Macro Regional apresentam maiores chances de serem internadas devido a ocorrência dos tipos C50.1, C50.4 e C50.8, e menores chances devido ao tipo C50.5 de câncer de mama quando comparadas as mulheres residentes na Macro Regional Leste.

Notou-se que para os tipos C50.0, C50.2 e C50.3, não houve diferenças significativas entre as Macro Regionais quando comparadas com a Macro Regional Leste.

Analisando os resultados sobre a Raça, também obteve-se diferenças entre as categorias da variável resposta.

A Raça **Negra** somente foi significativa para o tipo C50.5, com um valor de *odds ratio* menor do que 1. Ou seja, mulheres de Raça Negra possuem menor chances de serem internadas devido ao tipo C50.5 de câncer de mama do que mulheres da Raça Branca. Para essa Raça não foram detectadas diferenças significativas nas demais categorias da variável resposta.

Para a Raça **Amarela**, obteve-se resposta de que uma mulher da Raça Amarela possui menor chance de ser internada pelo tipo C50.5 de câncer de mama quando comparada a uma mulher de Raça Branca. Para as demais categorias, não foram apresentados diferenças significativas.

A Raça **Parda** apresentou diferenças em 5 as 8 categorias estudadas. Os resultados obtidos foram de que mulheres da Raça Pardo possuem maiores chances de serem internadas devido aos tipos C50.0, C50.1, C50.4, C50.5 e C50.8 de câncer de mama quando comparadas a mulheres da Raça Branca. Em nenhuma categoria a Raça Pardo apresentou valores que significassem uma menor chance de internações.

Para a Raça **Indígena**, houve divergência nas estimativas entre os métodos de estimação utilizados. O método Quadratura de Gauss-Hermite indicou como significativa as estimações para essa Raça. Já o método de Laplace, contrariou esta significância. Em todo o banco de dados, houve somente 11 internações com essa característica. Utilizando o método de estimação da Quadratura de Gauss-Hermite, temos que mulheres da Raça Indígena, quando comparadas com mulheres de Raça Branca, possuem menor chance de ser internadas pelos tipos C50.1, C50.2, C50.3, C50.4, C50.5, C50.6 e C50.8 de câncer de mama. Somente para o tipo C50.0 de câncer de mama, não houve diferenças significativas.

Na análise dos resultado sobre as Faixas Etárias surgiram algumas informações quando comparadas com a Faixa “IV”, que é indicada pela Organização Mundial de Saúde e pelo Instituto Nacional de Câncer, como a faixa com predominância dos casos.

A **Faixa Etária “I”**, que contém as mulheres com idade menor que 14 anos, apresentou significância estatística para os CIDs C50.3 e C50.4, onde ambos se apresentaram como fator de proteção. Ou seja, as ocorrências de câncer de mama em mulheres menores de 14 anos possuem menores frequências nos CIDs citados, quando comparada com a categoria de referência.

A **Faixa Etária “II”** apresenta valores de *odds ratio* maiores do que 1 para os CIDs C50.0, C50.1, C50.2, C50.3, C50.5, C50.6 e C50.8. Com exceção da categoria C50.5, as *odds ratio* apresentadas são valores muito altos. Como esses valores são de comparação com a categoria C50.9, que indica somente o câncer de mama de uma maneira geral, e sem especificações, esses altos valores podem indicar que, para essa faixa etária, os diagnósticos são mais precisos, resultando em uma melhor distribuição dos casos de câncer de mama entre suas categorias.

Para a **Faixa Etária “III”**, os resultados apresentados indicam que as ocorrências de câncer de mama para os CIDs C50.1 e C50.6, são mais frequêntes quando comparadas à categoria de referência.

Finalmente, os resultados para a **Faixa Etária “V”** indicam que para todas as categorias do CID, com exceção do CID C50.2, acontece o oposto do resultado da Faixa Etária “II”. Neste caso, os valores apresentados indicam que as ocorrências dessa faixa etária para essas categorias são menores do que as ocorrências na categoria referência, que indica o caso geral do câncer.

7.4 Limitações

Uma importante limitação estatística deste trabalho é proveniente de aspectos computacionais. Embora vários *softwares* sejam capazes de trabalhar com modelos lineares generalizados mistos, essa quantidade diminui quanto a distribuição utilizada é a multinomial.

Para as distribuições Binomial e Poisson, os programas são capazes de realizar análises de maneira mais eficientes e práticas, porém com a distribuição multinomial, e ainda com a utilização de muitas categorias nas variáveis respostas, análises computacionais são mais restritas.

Devido a grande quantidade de categorias da variável resposta no estudo (9), e também da grande quantidade de classes para cada uma das variáveis preditoras, a quantidade de possibilidades para cada modelo se torna computacionalmente trabalhoso. No total, considerando as oito funções `logit`, temos 21120 casos, pois são 8 funções, e para cada uma temos 6 categorias para Raça, 5 para a Faixa Etária, 4 para as Macro Regionais e 22 para as Regionais de Saúde. Para cada `logit` são 2640 variações.

Com esses fatos e valores em consideração, uma extensa e trabalhosa revisão de literatura foi necessária para acomodar as ferramentas necessárias para a análise do estudo, e através de trocas de e-mail com os professores Alan Agresti, do departamento de estatística da Universidade da Florida nos Estados Unidos, e Geert Molenberghs, do Centro de Bioestatística e Bioinformática da Universidade de Leuven na Bélgica, essa falta de metodologia e aspectos computacionais ligado a modelos multiníveis com distribuição multinomial ficou ainda mais evidente.

As análises de resíduos para dados com comportamento hierárquicos, especialmente para dados não Gaussianos ainda é de difícil acesso ou menor disponibilidade, e é uma lacuna presente nos *softwares* utilizados.

Cabe ressaltar ainda, que durante a realização deste estudo, encontrou-se problemas na construção do banco de dados. A falta de atenção dos profissionais da saúde responsáveis pelo preenchimento dessas informações torna o trabalho mais árduo, e com lacunas na garantia de qualidade dos dados. Entre os empecilhos, citamos a grande quantidade de internações sem informação sobre raça, embora esse valor tenha diminuído ente 2008 com 784 casos a 2016 com 56 internações, e a grande quantidade de internações sob o CID C50.9, que se trata de uma maneira geral de descrever o câncer de mama, quando este é passível de uma descrição melhor detalhada. E, como consequência, é possível que um mesmo paciente possa ter sido internado mais de uma vez. Um problema que é possível de acontecer com essa informação, é de que a mesma paciente que na primeira internação foi informado possuir um certo tipo de câncer de mama (CID), seja novamente internado e, nessa segunda internação, ela seja informado outro tipo de CID. Por exemplo, uma paciente foi internada e apresentada com o CID C50.9, que é a categoria geral de câncer de mama, sem outras especificações. Porém, em uma segunda internação da mesma paciente, o profissional de saúde responsável tenha decidido que o diagnóstico de câncer de mama é o CID C50.5.

7.5 Trabalhos Futuros

- Apresentar novas análises referentes a estes dados;
- Realização de análise estatística espacial;
- Realizar análise de diagnósticos;
- Abordagem alternativa para a análise desse banco de dados, sem a utilização de modelos multiníveis, mas sim utilizando métodos de Survey, conforme sugerido pelo professor Geert Molenberghs.

Referências

AGRESTI, A. **An introduction to categorical data analysis**. [S.l.]: Wiley New York, 1996. v. 135.

AITKIN, M.; ANDERSON, D.; HINDE, J. Statistical modelling of data on teaching styles. **Journal of the Royal Statistical Society. Series A (General)**, JSTOR, p. 419–461, 1981.

AITKIN, M.; LONGFORD, N. Statistical modelling issues in school effectiveness studies. **Journal of the Royal Statistical Society. Series A (General)**, JSTOR, p. 1–43, 1986.

ANDERSON, D. A.; AITKIN, M. Variance component models with binary response: interviewer variability. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 203–210, 1985.

BENNETT, N.; JORDAN, J.; LONG, G.; WADE, B. Teaching styles and pupil progress. JSTOR, 1976.

BOLKER, B. M.; BROOKS, M. E.; CLARK, C. J.; GEANGE, S. W.; POULSEN, J. R.; STEVENS, M. H. H.; WHITE, J.-S. S. Generalized linear mixed models: a practical guide for ecology and evolution. **Trends in ecology & evolution**, Elsevier, v. 24, n. 3, p. 127–135, 2009.

BOOTH, J. G.; HOBERT, J. P. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 61, n. 1, p. 265–285, 1999.

BOYLE, M. H.; LIPMAN, E. L. **Do Places Matter?: A Multilevel Analysis of Geographic Variations in Child Behaviour in Canada**. [S.l.]: Human Resources Development Canada, Applied Research Branch, 1998.

BRESLOW, N. E.; CLAYTON, D. G. Approximate inference in generalized linear mixed models. **Journal of the American statistical Association**, Taylor & Francis Group, v. 88, n. 421, p. 9–25, 1993.

CAPANU, M.; GÖNEN, M.; BEGG, C. B. An assessment of estimation methods for generalized linear mixed models with binary outcomes. **Statistics in medicine**, Wiley Online Library, v. 32, n. 26, p. 4550–4566, 2013.

CORDEIRO, G. M.; DEMÉTRIO, C. G. Modelos lineares generalizados e extensões. **Sao Paulo**, 2008.

DOBSON, A. J.; BARNETT, A. **An introduction to generalized linear models**. [S.l.]: CRC press, 2008.

DRAPER, N. R.; SMITH, H. **Applied regression analysis**. [S.l.]: John Wiley & Sons, 2014.

DUNCAN, C.; JONES, K.; MOON, G. Do places matter? a multi-level analysis of regional variations in health-related behaviour in Britain. **Social science & medicine**, Elsevier, v. 37, n. 6, p. 725–733, 1993.

DUNCAN, C.; JONES, K.; MOON, G. Context, composition and heterogeneity: using multilevel models in health research. **Social science & medicine**, Elsevier, v. 46, n. 1, p. 97–117, 1998.

ELKIN, E. B.; HUDIS, C.; BEGG, C. B.; SCHRAG, D. The effect of changes in tumor size on breast carcinoma survival in the us: 1975–1999. **Cancer**, Wiley Online Library, v. 104, n. 6, p. 1149–1157, 2005.

EZZET, F.; WHITEHEAD, J. A random effects model for ordinal responses from a crossover trial. **Statistics in medicine**, Wiley Online Library, v. 10, n. 6, p. 901–907, 1991.

FIKRET, I. Generalized linear mixed models: An introduction for tree breeders and pathologists. In: **Fourth International Workshop on the Genetics of Host-Parasite Interactions in Forestry**. [S.l.: s.n.], 2011. v. 31.

FITZMAURICE, G.; DAVIDIAN, M.; VERBEKE, G.; MOLENBERGHS, G. **Longitudinal data analysis**. [S.l.]: CRC Press, 2008.

FITZMAURICE, G. M.; LAIRD, N. M.; WARE, J. H. **Applied longitudinal analysis**. [S.l.]: John Wiley & Sons, 2012. v. 998.

FROST, J. **Why You Need to Check Your Residual Plots for Regression Analysis: Or, To Err is Human, To Err Randomly is Statistically Divine**. 2012. Disponível em: <http://blog.minitab.com/blog/adventures-in-statistics/why-you-need-to-check-your-residual-plots-for-regression-analysis>.

GILTHORPE MARK S, C. S. J. The application of multilevel, multivariate modelling to orthodontic research data. **Community dental health**, v. 17, n. 4, p. 236–242, 2000.

GOLDSTEIN, H. Multilevel mixed linear model analysis using iterative generalized least squares. **Biometrika**, Biometrika Trust, v. 73, n. 1, p. 43–56, 1986.

GOLDSTEIN, H. **Multilevel statistical models**. [S.l.]: John Wiley & Sons, 2011. v. 922.

GRILLI, L.; RAMPICHINI, C. A multilevel multinomial logit model for the analysis of graduates' skills. **Statistical Methods and Applications**, Springer, v. 16, n. 3, p. 381–393, 2007.

HANCOCK, G. R.; MUELLER, R. O. **The reviewer's guide to quantitative methods in the social sciences**. [S.l.]: Routledge, 2010.

HARTZEL, J. .; AGRESTI, A. Multinomial logit random effects models. **Statistical Modelling**, SAGE Publications, v. 1, n. 2, p. 81–102, 2001.

HARVILLE, D. A. Maximum likelihood approaches to variance component estimation and to related problems. **Journal of the American Statistical Association**, Taylor & Francis, v. 72, n. 358, p. 320–338, 1977.

HARVILLE, D. A.; MEE, R. W. A mixed-model procedure for analyzing ordered categorical data. **Biometrics**, JSTOR, p. 393–408, 1984.

HEDEKER, D. Generalized linear mixed models. **Encyclopedia of statistics in behavioral science**, Wiley Online Library, 2005.

HENDERSON, C. R. Estimation of genetic parameters. In: INTERNATIONAL BIOMETRIC SOC 1441 I ST, NW, SUITE 700, WASHINGTON, DC 20005-2210. **Biometrics**. [S.l.], 1950. v. 6, n. 2, p. 186–187.

INCA. **Instituto Nacional de Câncer José Alencar Gomes da Silva**. [S.l.]: Estimativa, 2016.

JANSEN, J. On the statistical analysis of ordinal data when extravariation is present. **Applied Statistics**, JSTOR, p. 75–84, 1990.

JONES, K. Everywhere is nowhere: multilevel perspectives on the importance of place. **The University of Portsmouth Inaugural Lectures**, 1993.

JR, D. W. H.; LEMESHOW, S. **Applied logistic regression**. [S.l.]: John Wiley & Sons, 2004.

KLUTHCOVSKY, A. C. G. C.; PALOZI, F. T. N.; CARNEIRO, F. H.; STRONA, R. Female breast cancer mortality in Brazil and its regions. **Revista da Associação Médica Brasileira**, SciELO Brasil, v. 60, n. 4, p. 387–393, 2014.

KUSS, O.; MCLERRAN, D. A note on the estimation of the multinomial logistic model with correlated responses in SAS. **Computer methods and programs in biomedicine**, Elsevier, v. 87, n. 3, p. 262–269, 2007.

LAIRD, N. M.; WARE, J. H. Random-effects models for longitudinal data. **Biometrics**, JSTOR, p. 963–974, 1982.

LINDSTROM, M. J.; BATES, D. M. Nonlinear mixed effects models for repeated measures data. **Biometrics**, JSTOR, p. 673–687, 1990.

LIU, Q.; PIERCE, D. A. A note on Gauss—Hermite quadrature. **Biometrika**, Biometrika Trust, v. 81, n. 3, p. 624–629, 1994.

LONGLEY, P. A.; BATTY, M. **Spatial analysis: modelling in a GIS environment**. [S.l.]: John Wiley & Sons, 1996.

MASON, W. M.; WONG, G. Y.; ENTWISLE, B. Contextual analysis through the multilevel linear model. **Sociological methodology**, JSTOR, v. 1984, p. 72–103, 1983.

MATOS, J. C. de; CARVALHO, M. D. de B.; PELLOSO, S. M.; UCHIMURA, T. T.; MATHIAS, T. A. de F. Mortalidade por câncer de mama em mulheres do município de Maringá, Paraná, Brasil. **Revista Gaúcha de Enfermagem**, v. 30, n. 3, p. 445, 2009.

MCCULLAGH, P.; NELDER, J. A. **Generalized linear models**. [S.l.]: CRC press, 1989. v. 37.

MCCULLOCH, C.; SEARLE, S. **Generalized, Linear, and Mixed Models, Vol. 1**. 1. ed. [S.l.]: Wiley-Interscience, 2001. (Wiley Series in Probability and Statistics). ISBN 9780471193647,0-471-19364-X.

MCCULLOCH, C. E. Maximum likelihood algorithms for generalized linear mixed models. **Journal of the American statistical Association**, Taylor & Francis, v. 92, n. 437, p. 162–170, 1997.

MCCULLOCH, C. E.; NEUHAUS, J. M. **Generalized linear mixed models**. [S.l.]: Wiley Online Library, 2001.

MDALA, I.; HAFFAJEE, A. D.; SOCRANSKY, S. S.; BLASIO, B. F. D.; THORESEN, M.; OLSEN, I.; GOODSON, J. M. Multilevel analysis of clinical parameters in chronic periodontitis after root planing/scaling, surgery, and systemic and local antibiotics: 2-year results. **Journal of oral microbiology**, v. 4, 2012.

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. **Encyclopedia of statistical sciences**, Wiley Online Library, 1972.

ORGANIZATION, W. H. Provisional guidelines on standard international age classification. **Department of International Economic and Social Affairs**, Series M, p. 74, 1982.

ORGANIZATION, W. H. **The world health report 2001 - Mental Health: New Understanding, New Hope**. Genf, Schweiz, 2016. Disponível em: <http://www.who.int/whr/2001/en/index.html>.

PINHEIRO, J.; BATES, D. **Mixed-effects models in S and S-PLUS**. [S.l.]: Springer Science & Business Media, 2006.

PINHEIRO, J.; BATES, D.; DEBROY, S.; SARKAR, D.; R Core Team. **nlme: Linear and Nonlinear Mixed Effects Models**. [S.l.], 2016. R package version 3.1-124. Disponível em: <http://CRAN.R-project.org/package=nlme>.

PINHEIRO, J. C.; CHAO, E. C. Efficient laplacian and adaptive gaussian quadrature algorithms for multilevel generalized linear mixed models. **Journal of Computational and Graphical Statistics**, Taylor & Francis, v. 15, n. 1, p. 58–81, 2006.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2015. Disponível em: <https://www.R-project.org/>.

ROBINSON, G. K. That BLUP is a good thing: the estimation of random effects. **Statistical science**, JSTOR, p. 15–32, 1991.

RODRIGUES, J. D.; CRUZ, M. S.; PAIXÃO, A. N. Uma análise da prevenção do câncer de mama no Brasil. **Revista Ciência & Saúde Coletiva**, v. 20, n. 10, 2015.

RUIZ, C. A.; JUNIOR, R. F. Thoughts on breast cancer in Brazil. **Revista da Associação Médica Brasileira**, SciELO Brasil, v. 61, n. 1, p. 1–2, 2015.

SCHABENBERGER, O. Introducing the GLIMMIX procedure for generalized linear mixed models. **SUGI 30 Proceedings**, Citeseer, p. 196–30, 2005.

SEARLE, S. R.; CASELLA, G.; MCCULLOCH, C. E. **Variance components**. [S.l.]: John Wiley & Sons, 2009. v. 391.

SESA/PR. Secretaria de estado da saúde do Paraná. 2016.

SINGER, J. M.; ANDRADE, D. F. Análise de dados longitudinais. **Simpósio Nacional de Probabilidade e Estatística**, Embrapa São Paulo, v. 7, 1986.

SNIJDERS, T.; BOSKER, R. **Multilevel analysis: An introduction to basic and applied multilevel analysis**. [S.l.]: London: Sage, 1999.

TAMURA, K. A. **Modelo Logístico Multinível: um enfoque em métodos de estimação e predição**. Tese (Dissertação de Mestrado) — Universidade de São Paulo, 2007.

TJPR. Tribunal de justiça do estado do Paraná. 2009.

VERBEKE, G.; MOLENBERGHS, G. **Models for Discrete Longitudinal Data**. [S.l.]: Springer Series in Statistics, 2005.

WEST, B. T.; WELCH, K. B.; GALECKI, A. T. **Linear mixed models: a practical guide using statistical software**. [S.l.]: CRC Press, 2014.

WOLFINGER, R.; O'CONNELL, M. Generalized linear mixed models a pseudo-likelihood approach. **Journal of statistical Computation and Simulation**, Taylor & Francis, v. 48, n. 3-4, p. 233–243, 1993.

WÜNSCH, F. V.; MONCAU, J. Mortalidade por câncer no Brasil 1980-1995: padrões regionais e tendências temporais. **Rev Assoc Med Bras**, SciELO Brasil, v. 48, n. 3, p. 250–7, 2002.

Capítulo 8

Apêndice A

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Estudo comparativo do poder de afastamento gengival vertical entre o cloreto de alumínio e o cloridrato de nafazolina

Pesquisador: Sérgio Sábio

Área Temática:

Versão: 2

CAAE: 53153116.6.0000.0104

Instituição Proponente: CCS - Centro de Ciências da Saúde

Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 1.515.263

Apresentação do Projeto:

Trata-se de projeto de pesquisa proposto por pesquisador vinculado à Universidade Estadual de Maringá.

Objetivo da Pesquisa:

Avaliar o efeito do uso do cloridrato de nafazolina (presente em soluções oftálmicas) em comparação com o cloreto de alumínio (elemento de referência) no afastamento gengival vertical.

Avaliação dos Riscos e Benefícios:

Avalia-se que os possíveis riscos a que estarão submetidos os sujeitos da pesquisa serão suportados pelos benefícios apontados.

Comentários e Considerações sobre a Pesquisa:

O trabalho será realizado em um consultório particular da cidade de Londrina-PR.O tamanho da amostra foi calculado estatisticamente. Para a confiança de 95%, poder do teste de 80%, e tamanho do efeito 0,25 (diferença estatística signifcativa mínima entre antes e depois do tratamento), o tamanho amostral foi calculado em 24 pessoas para cada grupo. Assim, a amostra será composta por 24 indivíduos. Os dentes 13, 21 e 23 de cada paciente serão selecionados para os procedimentos (N=72). De forma randomizada cada dente receberá um dos tratamentos experimentais (N=24): O Grupo 1, receberá fios retratores impregnados com cloreto de alumínio;

Grupo 2 com cloridrato de nafazolina; e Grupo 3 (controle) sem qualquer tipo de substância química. Os procedimentos serão realizados em uma sessão clínica. Será realizada a avaliação da quantidade do afastamento gengival vertical provocado por cada um dos tratamentos.Inicialmente será realizado o isolamento relativo com roletes de algodão nas áreas correspondentes aos dentes a serem avaliados. Após a limpeza com fio dental e bolinha de algodão embebida em clorexidina a 2% será feito o enxágüe e a secagem dos dentes. Uma camada de protetor gengival fotopolimerizável Top dam® será aplicada à superfície dental dos elementos 13, 21 e 23 ao nível da margem cervical do sulco gengival, para registrar a posição inicial da mesma. Em seguida, utilizando a técnica do duplo fio (Shillinburg, 1998), os fios retratores serão posicionados nos dentes experimentais. Serão utilizados os fios retratores n° 000 e nº 1 (Ultrapak®). Os fios nº 1 dos grupos experimentais 1 e 2 serão embebidos em cloridrato de nafazolina (Legrand®) e cloreto de alumínio (Hemostop® - Grupo 2), respectivamente, por 7 minutos antes de serem aplicados nos dentes. A colocação dos fios retratores nº 1 obedecerá a randomização dos tratamentos. Inicialmente os fios retratores n° 000 serão posicionados no interior do sulco gengival da face vestibular de cada um dos dentes. Em seguida, os fios retratores nº 1 serão posicionados sobre o primeiro fio.Após um período de quatro minutos, os fios retratores serão retirados do sulco gengival, a área será seca com jatos de ar e a moldagem será realizada utilizando silicone por adição (polivinilsiloxana - 3D – Angelus, Londrina – Brasil). Após a tomada de presa do silicone por adição, a moldeira será removida da boca. Decorridas duas horas desta moldagem, o molde será vazado em gesso especial tipo IV. Os modelos serão então recortados em pequenos blocos, e a partir destes, 72 imagens dos modelos (uma imagem para cada dente tratado com afastamento gengival) serão capturadas por uma câmera, acoplada a uma lupa (Olympus SZ-STS). As imagens serão analisadas através do programa Image pro-plus (versão 4.5) para medir a distância entre o protetor gengival Top dam® (que marca a posição inicial da gengiva) e o nível da gengiva. Essas medidas serão realizadas por um único examinador cego aos procedimentos.

Considerações sobre os Termos de apresentação obrigatória:

Apresenta Folha de Rosto devidamente preenchida e assinada pelo responsável institucional. O cronograma de execução é compatível com a proposta enviada. Descreve gastos sob a responsabilidade do pesquisador. O Termo de Consentimento Livre e Esclarecido contempla as garantias mínimas preconizadas. Apresenta as autorizações necessárias. Sanadas as pendências apontadas.

Recomendações:

Conclusões ou Pendências e Lista de Inadequações:

O Comitê Permanente de Ética em Pesquisa Envolvendo Seres Humanos da Universidade Estadual de Maringá é de parecer favorável à aprovação do protocolo de pesquisa apresentado.

Considerações Finais a critério do CEP:

Face ao exposto e considerando a normativa ética vigente, este Comitê se manifesta pela aprovação do protocolo de pesquisa em tela.

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_201830.pdf	24/03/2016 16:53:58		Aceito
Outros	Carta_resposta_digitalizada.pdf	24/03/2016 16:52:37	Sérgio Sábio	Aceito
Outros	Carta_proprietario_consultorio.pdf	11/02/2016 16:51:32	Sérgio Sábio	Aceito
Outros	Questionario_Waenya.pdf	15/01/2016 11:31:41	Waenya Fernandez de Carvalho	Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	TCLE_Waenya.pdf	15/01/2016 11:25:11	Waenya Fernandez de Carvalho	Aceito
Projeto Detalhado / Brochura Investigador	Projeto_de_pesquisa_AGV.pdf	15/01/2016 11:21:51	Waenya Fernandez de Carvalho	Aceito
Folha de Rosto	Folha_de_Rosto_assinada.pdf	15/01/2016 11:20:52	Waenya Fernandez de Carvalho	Aceito

Situação do Parecer:

Aprovado

Necessita Apreciação da CONEP:

Não

MARINGA, 26 de Abril de 2016

Assinado por:
Ricardo Cesar Gardiolo
(Coordenador)

Endereço: Av. Colombo, 5790, UEM-PPG
Bairro: Jardim Universitário **CEP:** 87.020-900
UF: PR **Município:** MARINGA
Telefone: (44)3011-4597 **Fax:** (44)3011-4444 **E-mail:** copep@uem.br

Capítulo 9

Apêndice B

Segundo (TAMURA, 2007), a integral descrita em 6.40 pode ser resolvida utilizando integração numérica ou quadraturas.

A quadratura é o caso mais simples da avaliação de uma integral dada por

$$I = \int_a^b f(x)dx, \tag{9.1}$$

que é equivalente a resolver $I \equiv y(b)$, a equação diferencial $\frac{dy}{dx} = f(x)$ com $y(a) = 0$, com a e b sendo números reais.

O método de quadratura é baseado na avaliação da função f numa sequência de pontos, pertencentes ao intervalo de integração denotados por $x_0, x_1, \cdots, x_N, x_{N+1}$ que são espaçados por uma constante h, tal que

$$x_i = x_0 + ih,$$

com $i = 0, 1, \cdots, N, N+1$. Logo, a função $f(x)$ tem valores conhecidos de $x_i's$

Essa função pode ser multiplicada por um peso w_i que pode ser interpretada como uma média ponderada dessa função na forma

$$\int_{-\infty}^{\infty} f(x_i)w_i dx. \tag{9.2}$$

A função integral dada em 9.2 é aproximada pela soma de seus valores funcionais do conjunto de pontos igualmente espaçados e multiplicados por certos coeficientes (pesos).

Para definirmos o que é uma quadratura Gaussiana, seja $w(x)$ uma função de pesos fixados em um intervalo $[a, b]$ (finito ou infinito), definimos *produto interno* de duas funções

$f(x)$ e $g(x)$ sob um intervalo $[a, b]$ com respeito ao peso $w(x)$ como

$$\int_a^b w(x)f(x)g(x)dx = (f, g). \tag{9.3}$$

As funções f e g são chamadas *ortogonais* sob o intervalo $[a, b]$ com respeito ao peso $w(x)$ se

$$\int_a^b w(x)f(x)g(x)dx = 0. \tag{9.4}$$

Podemos, então, definir uma sequência de polinômios $p_0(x), p_1(x), \cdots, p_n(x)$ que são ortogonais e no qual cada $p_n(x)$ tenha o exato grau n sob o peso $w(x)$, isto é

$$(p_m, p_n) = \int_a^b w(x)p_m(x)p_n(x)dx = 0, \tag{9.5}$$

com $m \neq n$.

Agora, multiplicando cada $p_n(x)$ por uma constante apropriada, pode-se produzir um conjunto de polinômios $p*_n(x)$ que são *ortonormais*, isto é:

$$(p*_m, p*_n) = \int_a^b w(x)p*_m(x)p*_n(x)dx = \delta_{mn} = \begin{cases} 0, \text{ se } m \neq n, \\ 1, \text{ se m = n.} \end{cases} \tag{9.6}$$

Se os n pontos distintos de $x_0, x_1, \cdots, x_n$ do intervalo $[a, b]$ são especificados a priori, então é necessário definir os coeficientes $w_0, w_1, \cdots, w_n$ tal que a regra

$$\int_{-\infty}^{\infty} w(x)f(x)dx = \sum_{i=1}^{n} w_i f(x_i) \tag{9.7}$$

será exata para todo polinômio da classe P_{n-1}, isto é, todas as combinações lineares da potência n com $1, x, x^2, \cdots, x^{n-1}$.

Portanto, a quadratura de Gauss-Hermite pode ser reescrita em termos da densidade normal como

$$\int_{\Re} f(x)\phi(x : \mu, \sigma)dx \approx \sum_{i=1}^{N} w_i f(z_i) \tag{9.8}$$

em que $\phi(x : \mu, \sigma)$ é uma densidade normal arbitrária.

Os pontos para avaliação são $z_i = \mu + \sqrt{2}\sigma x_i$ em que os pesos são modificados de w_i para $\frac{w_i}{\sqrt{\pi}}$, com $i = 1, \cdots, N$ número de pontos de quadratura. A essa aproximação dá-se o nome Quadratura Clássica. Dependendo da maneira com a qual as abscissas são amostradas, é necessário considerar muitos pontos para aproximar a integral.

Usando $\phi(x : \widehat{\mu}, \widehat{\sigma})$, a aproximação é dada por

$$\int g(x)dx \approx \sqrt{2}\widehat{\sigma} \sum_{i=1}^{N} w *_i g(\widehat{\mu} + \sqrt{2}\widehat{\sigma}x_i), \tag{9.9}$$

em que $w*_i = w_i + \exp(x^2)$.

Quando a equação 9.9 é aplicada em somente um ponto, o resultado é a aproximação de Laplace, conforme (LIU; PIERCE, 1994).