UFC

Universidade Federal do Ceará
Centro de Ciências
Departamento de Estatística e Matemática Aplicada
Bacharelado em Estatística

Victor Meneses Navarro Fernandes e Wesley Braga Barbosa

Modelos de Regressão para Dados Longitudinais

Fortaleza
2022

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

1 Introdução

Por definição, nos estudos com característica transversal, as unidades amostrais são mensuradas em um único instante. Logo, a inferência é realizada somente para esta condição. Todavia, quando o fenômeno de interesse envolve tratamentos em diferentes instantes, distâncias ou outra condição ordenada, apresentam-se mais de uma mensuração para uma mesma observação. Assim, os modelos tradicionais lineares de regressão não são adequados. Tendo em vista que um dos pressupostos do modelo linear de regressão é a não correlação dos erros aleatórios. Entretanto, para dados com medidas repetidas nas mesmas condições de avaliações, o erro aleatório possui uma interpretação fundamental. Nesse contexto, o interesse do pesquisador recai no estudo da variação entre as unidades amostrais e na variação entre as condições das avaliações.

Especificamente, dados longitudinais são uma classe de dados com medidas repetidas em que as condições de avaliação são ordenadas e não permutáveis. Diferentemente de séries temporais, as quais observam-se uma ou mais longas séries históricas de uma dada variável, a característica longitudinal refere-se a uma estrutura de repetição de um mesmo indivíduo (unidade amostral) em instantes de tempo diferentes. Uma particularidade associada ao estudo longitudinal refere-se à presença ou não de omissão de dados nas condições de avaliação. O planejamento do estudo é considerado balanceado com relação ao tempo quando não houver omissão de observações em algum instante específico nas condições de avaliação. Outra característica refere-se a regularidade, um estudo longitudinal é dito regular se as condições de avaliações são equidistantes.

Considere o seguinte exemplo extraido de Potthoff e Roy (1964), que consistiu em investigar a mudança na distância, em milímetros, entre o centro da glândula pituitária, também chamada de hipófise, localizada na sela túrcica, e da fissura pterigomaxilar, localizada na fossa pterigomaxila, em jovens entre 8 e 14 anos do sexo feminino e masculino. Os dados contêm 27 unidades amostrais (indivíduos) mensuradas em 4 condições de avaliação (idade) e um fator interunidade amostral, gênero do indivíduo. O estudo é considerado regular e balanceado nas condições de avaliação e desbalanceado em relação ao gênero. Os perfis individuais e médio indicam tendência crescente da distância conforme o indivíduo envelhece conforme a figura abaixo.

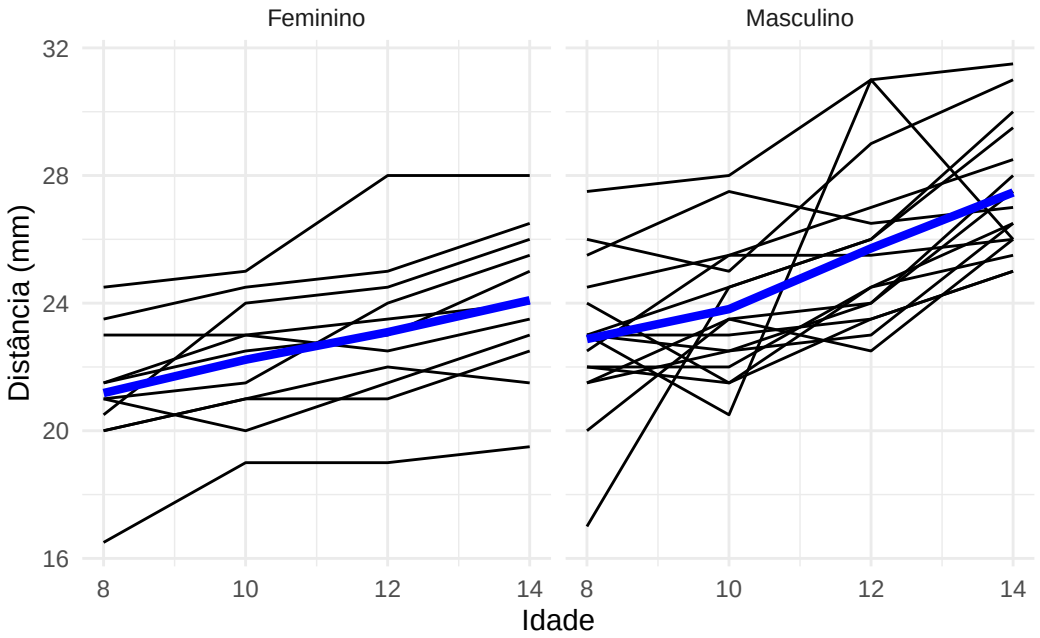

Figura 1: Perfis individuais e médios da distância, em milímetros, entre o centro da glândula pituitária e fissura pterigomaxilar pelo sexo dos participantes.

Erroneamente, pode-se ajustar uma regressão para dados longitudinais sem identificar as unidades de investigação em tempos distintos. Na figura 2, nota-se o confundimento incorrido quando não consideramos a relação intra-unidades no estudo.

Desse modo, surge a necessidade de ajustar um modelo que possibilite definir uma estrutura de covariâncias para os componentes mencionados. A escolha da matriz de covariâncias deve ser suficientemente flexível para incluir no mínimo três fontes diferentes de variação aleatória, aquela devida aos efeitos aleatórios, correlação serial e ao erro de mensuração (Diggle et al, 2002). Para guiar a escolha dessa estrutura, recomenda-se a análise dos gráficos de perfis médios e de perfis individuais, da função de autocorrelação e do variograma amostral das observações. Ademais, no apêndice **A**, apresentamos brevemente algumas estruturas de covariância.

A análise de dados longitudinais envolve essencialmente modelos de MANOVA, marginais e mistos. Os modelos de MANOVA são adequados quando admite-se uma estrutura de covariâncias não-estruturada e não existe omissão de dados. Quando o interesse recai sobre a modelagem do perfil médio, modelos

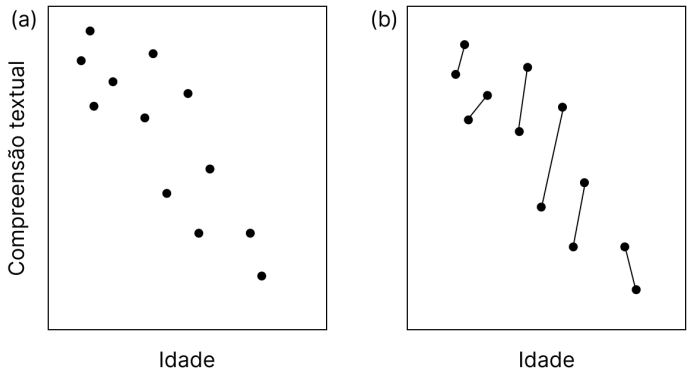

Figura 2: Dados hipotéticos da relação entre idade e compreensão textual (Adaptado de Diggle et al, 2002)

marginais são mais apropriados. Frequentemente, busca-se realizar inferências para uma população geradora e não somente nos elementos contidos na amostra. Nos modelos mistos, consideramos um modelo probabilístico subjacente às unidades de investigação, denominados de efeitos aleatórios. Além disso

Embora modelos de regressão para dados longitudinais apresentam a possibilidade de estudar a estrutura de covariância do fenômeno de interesse, executá-lo torna-se mais dispendioso. Além disso, outra problemática envolve não ocorrer a coleta de dados em determinados instantes do estudo devido a impossibilidade de contactar o indivíduo ou decorrentes de outros fatores que as inviabilizam.

O presente relatório possui a seguinte estrutura, na seção 2 apresentamos as principais propostas de modelos de regressão para dados longitudinais, o modelo linear multivariado normal, marginal e o misto. Na seção 3, abordamos o inferência em modelos mistos tendo como base Verbeke e Molenbergh (2000). Ademais, a fim de exemplificar a aplicação dos modelos mencionados na segunda seção, será analisado um conjunto de dados referente a um ensaio clínico odontológico (Hedge & Kuch, 1999), envolvendo a descrição e ajuste.

2 Modelos para dados longitudinais

A principal característica que difere os modelos para dados longitudinais dos modelos lineares gerais para dados transversais consiste na capacidade de incorporar a possível correlação entre as observações em uma mesma condição de avaliação e em uma mesma unidade amostral. Em estudos longitudinais, frequentemente, surge a necessidade de avaliar a existência de influência do tempo em uma variável resposta de um dado grupo e/ou interações entre outros fatores, como em ensaios clínicos para verificar eficácia de medicamentos, ou, ajustar polinômios para descrever o crescimento ou decrescimento da variável ao longo do tempo. Para avaliar estas e outras questões, varios modelos podem ser empregados. Nesta seção apresentamos e discutimos três modelos lineares multivariados, justificando sua utilização e expondo suas limitações.

2.1 Definições iniciais

Dados longitudinais apresentam-se em dois formatos de tabelas, longo e amplo. No formato amplo, comumente denominado de formato multivariado, a i-ésima linha corresponde ao perfil de resposta individual, ou somente vetor de variáveis respostas, $\mathbf{y}_i = (y_{i1}, ..., y_{in_j})$, que corresponde as mensurações da mesma unidade amostral em diferentes condições de avaliações $j = (1, 2, ..., m)$. No que tange ao manuseio computacional, o formato longo é mais apropriado para construção de matrizes de covariância e correlações amostrais, por exemplo. A inclusão de variáveis explicativas em estudos longitudinais aparece tanto como covariáveis numéricas ou representando fatores de um tratamento, podendo ou não variar ao longo das condições de avaliações. Por simplicidade, abordaremos exclusivamente o caso balanceado, $\sum_{j=1}^{m} n_j = N$, e sem covariáveis indexadas ao tempo.

2.2 Modelo normal linear multivariado

No geral, podemos utilizar modelos de análise de variância multivariada (MANOVA) quando o interesse recai na comparação entre os perfis médios de resposta em diferentes níveis de um fator, feminino e masculino, por exemplo. O modelo normal linear tem a seguinte forma funcional:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

Onde

- **Y** é uma matriz $n \times m$, em que a i-ésima linha corresponde ao perfil de resposta individual, $\mathbf{y}_{i \cdot k} = (y_{i1k}, ..., y_{imk})$ da i-ésima unidade amostral $(i = 1, ..., n_j)$ pertencente ao k-ésimo grupo $(k = 1, ..., p)$;
- **X** é uma matriz $n \times p$ conhecida e não-aleatória, $\mathbf{X} = \oplus_{j=1}^{p} \mathbf{1}_{n_j}$;
- **B** é uma matriz $p \times m$, em que a k-ésima linha corresponde ao perfil médio de respostas nos m níveis da condição de avaliação, $\mathbb{E}(\mathbf{y}_{\cdot\cdot k}^{T}) = \boldsymbol{\mu}_k = (\mu_{k1}, ..., \mu_{km})^T$;
- **E** é uma matriz $n \times m$ de erros aleatórios, $\mathbf{e}_i \sim \mathcal{N}_m(\mathbf{0}, \boldsymbol{\Sigma})$ independentes.

Admitimos matriz de covariâncias intra-unidades não-estruturada e homoscedástica em relação aos grupos e assumimos a não existência de omissão de dados. A hipótese de igualdade entre as matrizes de covariância pode ser verificada por meio do teste de Box.

O estimador de máxima verossimilhança para **B** no modelo linear multivariado é dado por:

$$\hat{\mathbf{B}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

A decomposição de matriz de covariâncias total

$$\boldsymbol{\Omega}_{\mathsf{T}} = \hat{\mathbf{E}}^T\hat{\mathbf{E}} + \hat{\mathbf{Y}}^T\hat{\mathbf{Y}} - n\bar{\mathbf{y}}\bar{\mathbf{y}}^T = \boldsymbol{\Omega}_{\mathsf{Reg}} + \boldsymbol{\Omega}_{\mathsf{Res}}$$

Em que $\hat{\mathbf{E}}$ é a matriz de resíduos, $\bar{\mathbf{y}}$ um vetor $m \times 1$ de médias da variável resposta e $\hat{\mathbf{Y}}$ uma matriz de valores ajustados. Sejam $\lambda_1 > ... > \lambda_m$ os autovalores diferentes de zero da matriz $\boldsymbol{\Omega}_{\mathsf{Reg}}\boldsymbol{\Omega}_{\mathsf{Res}}^{-1}$, para realizar procedimentos inferências, utiliza-se as seguintes estatísticas de teste:

- Traço de Pillai-Barlett $T_{PB} = \sum_{j=1}^{m} \frac{\lambda_j}{1-\lambda_j} = \text{tr}\left[\boldsymbol{\Omega}_{\mathsf{Reg}}(\boldsymbol{\Omega}_{\mathsf{Reg}} + \boldsymbol{\Omega}_{\mathsf{Res}})^{-1}\right]$;
- Traço de Hotelling-Lawley $T_{HL} = \sum_{j=1}^{m} \lambda_j$;
- Lambda de Wilk $\Lambda = \prod_{j=1}^{m} \frac{1}{1+\lambda_j}$;

- Máxima raiz de Roy λ_1

Para cada uma dessas estatísticas existem aproximações pela distribuição F (Rao, 1973).

Essencialmente, testes de hipóteses em modelos lineares multivariados podem ser expressos como:

$$\mathrm{H}_0 : \mathbf{CBU} = \mathbf{M} \tag{1}$$

Onde **C** é uma matriz $(p-1) \times p$ utilizada para definir as comparações das respostas esperadas entre os grupos, **U** uma matriz $m \times (m-1)$ utilizada para definir as comparações das respostas esperadas entre as condições de avaliação e **M** é uma matriz $p \times m$ de constantes preedeterminadas.

Equação 1 também pode ser expressada como uma hipótese linear geral, tal qual no modelo linear univariado:

$$\mathbf{L}\beta = \mathbf{M}$$

Em que $\mathbf{L} = \mathbf{C} \otimes \mathbf{U}^T$ e $\beta = \mathrm{vec}(\mathbf{B})$.

2.3 Modelo linear misto

A metodologia apresentada anteriormente desconsidera como analisar situações desbalanceadas e como modelar a matriz de covariâncias apropriadamente. Pois, cabe mencionar que o modelo normal linear multivariado não é capaz de manejar a omissão de dados e admite somente forma não-estrutuada para matriz de covariâncias.

Dessa maneira, a fim de alcançar o mesmo objetivo, nos modelos mistos, consideram-se variáveis latentes, comumente denominados efeitos aleatórios. Ou seja, quando os efeitos de interesse admitem um modelo probabilístico subjacente, estes, então, são aleatórios e modelam características individuais e são utilizamos como base para fazer inferências sobre a população geradora do fenômeno.

O modelo linear misto apresenta a seguinte forma funcional

$$\mathbf{y_i} = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i \; i = 1, ..., m \tag{2}$$

Onde

- $\mathbf{y}_i = (y_{i1}, ..., y_{in_i})^{\mathsf{T}}$ é o vetor $(n_i \times 1)$ do perfil individual de resposta da i-ésima unidade amostral;
- $\boldsymbol{\beta}$ é o vetor $(p \times 1)$ dos efeitos fixos;
- $\mathbf{X}_i = (\mathbf{x}_{i1}, ..., \mathbf{x}_{ip})$ é a matriz $(n_i \times p)$ de especificação dos efeitos fixos da i-ésima unidade amostral e da j-ésima variável explicativa $(j = 1, ..., p)$;
- $\mathbf{b}_i$ é o vetor $(q \times 1)$ dos efeitos aleatórios;
- $\mathbf{Z}_i$ é a matriz $(n_i \times q)$ de especificação dos efeitos aleatórios;
- $\mathbf{e}_i$ é o vetor $(n_i \times 1)$ dos erros aleatórios;

Assumimos que as matrizes de especificações $\mathbf{X}_i$ e $\mathbf{Z}_i$ são conhecidas e de posto completo.

Além disso que

$$\mathbf{b}_1, ..., \mathbf{b}_k \overset{\text{iid}}{\sim} \mathcal{N}_q(\mathbf{0}, \mathbf{G}) \text{ e } \mathbf{e}_i \overset{\text{ind}}{\sim} \mathcal{N}_{n_i}(\mathbf{0}, \mathbf{R}_i),\ i = 1, ..., m$$

Em que

- $\mathbf{b}_i$ e $\mathbf{e}_i$ independentes;
- $\mathbf{G}$ e $\mathbf{R}$ são matrizes de variância-covariância de dimensões $(q \times q)$ e $(n_i \times n_i)$, respectivamente.

Assim, a estrutura de dependência é modelada através das matrizes $\mathbf{G}$ e $\mathbf{R}$. Especificamente, . Além disso, ambas matrizes são funções de parâmetros desconhecidos, denominados componentes de variância $\boldsymbol{\alpha}$, ou seja, $\mathbf{G} = \mathbf{G}(\boldsymbol{\alpha})$ e $\mathbf{R} = \mathbf{R}(\boldsymbol{\alpha})$.Vale ressaltar que $\boldsymbol{\alpha}$ independe funcionalmente de β. Note que não podemos denominar independência estocástica, pois não são variáveis aleatórias.

Nesse contexto, a escolha da estrutura de covariância depende do conhecimento prévio sobre o fenômeno modelado e do processo de obtenção das observações. Além disso, pode-se utilizar de técnicas de análise exploratória para direcionar essa decisão, o variograma amostral ou gráfico de autocorrelação, por exemplo.

O modelo linear misto também pode ser entendido como um modelo linear em dois estágios. No primeiro estágio, o interesse recai primariamente na variação

intra-unidade, que equivale a ajustar sucessivamente o modelo condicionalmente ao conhecimento de $\mathbf{b}_i$. Subsequentemente, no segundo estágio, considera-se a variabilidade inter-unidade, incorporando o termo $\mathbf{b}_i$.

2.4 Modelo linear marginal

Essencialmente, no modelo linear marginal, modelamos o perfil médio da variável resposta. Note que isto corresponde a considerar o modelo anterior sem o efeito aleatório, ou seja

$$f(\mathbf{y}_i) = \int f(\mathbf{y}_i \mid \mathbf{b}_i) f(\mathbf{b}_i)\, d\mathbf{b}_i \tag{3}$$

Em que $\mathbf{y}_i|\mathbf{b}_i \sim \mathcal{N}_q(\mathbf{X}_i\beta, \mathbf{R}_i)$ e $\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{G})$, logo, $y_i \sim \mathcal{N}(\mathbf{X}_i\beta, \mathbf{V}_i)$, onde $\mathbf{V}_i = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i^T + \mathbf{R}_i$ é a matriz de covariância marginal.

Embora o modelo marginal possa ser obtido através do modelo misto, eles não são equivalentes. Tendo em vista que, no modelo marginal, somente as matrizes $\mathbf{V}_i$ precisam ser positiva-(semi)definidas, o que não necessariamente implicará em $\mathbf{G}$ e $\mathbf{R}_i$ também serem. Ademais, as mesmas possibilidades de estruturas de covariância podem ser escolhidas para a matriz $\mathbf{V}_i(\boldsymbol{\alpha})$, em que $\boldsymbol{\alpha}$ refere-se aos componentes de variância.

Funcionalmente, o modelo é dada por:

$$\mathbf{y_i} = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{e}_i,\ i = 1, ..., m \tag{4}$$

Onde

- $\mathbf{y}_i = (y_{i1}, ..., y_{in_i})^{\mathsf{T}}$ é o vetor $(n_i \times 1)$ do perfil individual de resposta da i-ésima unidade amostral;
- $\boldsymbol{\beta}$ é o vetor $(p \times 1)$ dos efeitos fixos;
- $\mathbf{X}_i = (\mathbf{x}_{i1}, ..., \mathbf{x}_{ip})$ é a matriz $(n_i \times p)$ de especificação dos efeitos fixos, conhecida e de posto completo, da i-ésima unidade amostral e da j-ésima variável explicativa $(j = 1, ..., p)$;
- $\mathbf{e}_i$ é o vetor $(n_i \times 1)$ dos erros aleatórios, $\mathbf{e}_i \sim \mathcal{N}_{n_j}(\mathbf{0}, \mathbf{V}_i)$;

3 Inferência

3.1 Estimação

Considere $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T\right)^T$ o vetor s-dimensional de parâmetros do modelo marginal com espaço paramétrico $\Theta = \Theta_{\boldsymbol{\beta}} \times \Theta_{\boldsymbol{\alpha}}$, em que $\Theta_{\boldsymbol{\beta}} = \mathbb{R}^p$ e $\Theta_{\boldsymbol{\alpha}}$ constituido dos valores de $\boldsymbol{\alpha}$ que tornam **G** e **R** positiva-(semi)definidas.

A inferência do modelo é baseado nos estimadores obtidos através da maximização da função de verossimilhança marginal

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{N} \left\{ (2\pi)^{-n_i/2} \mid \mathbf{V}_i(\boldsymbol{\alpha}) \mid^{-1/2} \exp\left(-\frac{1}{2} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T \mathbf{V}_i^{-1}(\boldsymbol{\alpha})(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) \right) \right\} \tag{5}$$

Onde $N = \sum_{i=1}^{m} n_i$. O estimador de máxima verossimilhança (**EMV**) para $\boldsymbol{\beta}$ considerando $\boldsymbol{\alpha}$ conhecido é dado por:

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}) = \left(\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i \right)^{-1} \sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{y}_i$$

Que coincide ao obtido pela estimação de mínimos quadrados generalizados. O **EMV** para $\boldsymbol{\alpha}$ pode ser obtido pela verossimilhança perfilada, substituindo $\boldsymbol{\beta}$ por $\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha})$ na Equação 5.

Todavia, vale recordar que ao estimar σ^2 pelo método da máxima verossimilhança no modelo normal quando μ é desconhecido, obtemos um estimador viesado. Por outro lado, quando μ é conhecido, o **EMV** é não viesado. Dito isso, uma alternativa para a obtenção de estimadores não viesados para σ^2 é considerar uma transformação do vetor $\mathbf{y}_i$ que sua distribuição não dependa de μ.

Por simplicidade, considere $\mathbf{y} \sim \mathcal{N}(\mu\mathbf{1}, \sigma^2 I_n)$, **A** uma matriz $N \times (N-1)$ com $N-1$ colunas linearmente independentes e que $\mathbf{A1} = \mathbf{0}$. Assim $\mathbf{U} = \mathbf{A^T y} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{A^T A})$ e maximizando a respectiva verossimilhança, obtemos

$$\hat{\sigma}^2 = \frac{\mathbf{yA}(\mathbf{A^T A})^{-1}\mathbf{A^T y}}{N-1}$$

O estimador resultante é denominado de estimador de máxima verossimilhança restrita. Além disso, note que qualquer transformação **U** satisfazendo as restrições supracitadas geram o mesmo estimador para σ^2.

Considerando o modelo linear misto na forma compacta, temos que $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\alpha}))$, em que $\mathbf{V}(\boldsymbol{\alpha}) = \text{diag}(\mathbf{V}_i)$. O **EMVR** para $\boldsymbol{\alpha}$ é obtido através da maximização da função de verossimilhança de $\mathbf{U} = \mathbf{A}^\mathbf{T}\mathbf{y}$, onde **A** é uma matriz $N \times (N-1)$ com $N-1$ colunas linearmente independentes e com $\mathbf{AX} = \mathbf{0}$. Assim, $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}^\mathbf{T}\mathbf{V}(\boldsymbol{\alpha})\mathbf{A})$ que não depende de $\boldsymbol{\beta}$.

Portanto, temos que

$$\mathcal{L}(\boldsymbol{\alpha}) = (2\pi)^{-(n-p)/2} \mid \sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i \mid^{1/2} \mid \sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i \mid^{-1/2} \prod_{i=1}^{N} \mid \mathbf{V}_i \mid^{-1/2}$$

$$\exp\left(-\frac{1}{2} \sum_{i=1}^{N} (\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})^T \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}) \right)$$

Note que

$$\mathcal{L}(\boldsymbol{\alpha}) = C \mid \sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i \mid^{-1/2} \mathcal{L}_{\text{MV}}(\boldsymbol{\theta}) \tag{6}$$

Onde C é uma constante que não depende de $\boldsymbol{\alpha}$. Ainda, como o segundo termo da Equação 6 não depende de $\boldsymbol{\beta}$, temos que

$$\mathcal{L}_{\text{MVR}}(\boldsymbol{\theta}) = \mid \sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i \mid^{-1/2} \mathcal{L}_{\text{MV}}(\boldsymbol{\theta})$$

Ambos estimadores, **MV** e **MVR**, detêm propriedades como consistência, normalidade assintótica e eficiência. No entanto, o método por **MVR** não estima os efeitos fixos, tendo em vista que **U** não depende de $\boldsymbol{\beta}$. Além disso, os erros quadráticos médios dos estimadores diferem significativamente na presença de muitos efeitos fixos (Harville, 1977). Ademais, para o procedimento de estimação dos componentes de variância, utilizam-se métodos numéricos como o algoritmo EM ou Newton-Raphson. Comumente, as estimativas dos parâmetros de variância para modelos complexos não convergem ou resultam em estimativas fora do espaço paramétrico. Nesse contexto, geralmente, quando

se utiliza o método Newton-Raphson, é possível ajustar os valores iniciais para tratar o problema da convergência, ou então, utilizar o algoritmo Escore de Fisher. Entretanto, não é uma tarefa trivial escolher os valores iniciais para os componentes de variância.

4 Exemplo de aplicação

Para ilustrar o uso de modelos de regressão para dados longitudinais,usaremos os dados de Hadgu e Koch (1999), que se referem a um ensaio clínico odontológico em que 109 adultos voluntários com pré-existência de placa dentária foram distribuidos aleatoriamente para receberem um enxaguante bucal tipo A (34 indivíduos), tipo B (36 indivíduos) e controle (39 individúos). O intuito do estudo foi avaliar o efeito dos enxaguantes bucais tipo A e B em relação ao controle para a inibição do desenvolvimento de placas dentárias. Para tanto, as placas dentárias de cada indivíduo foram avaliadas e classificadas segundo um escore do início do tratamento, após 3 meses e após 6 meses. Por simplicidade, foram omitidas quatro observações faltantes para as quais não foi possível obter o valor do escore.

Na Figura 3, os perfis individuais indicam tendência decrescente ao longo do período de escovação. Percebe-se considerável aumento da variação após 3 meses de uso dos enxaguantes, sobretudo no tipo A e, subsequentemente, no controle. Cabe ressaltar que o enxaguante bucal controle mantêm-se mais homogêneo em relação aos demais, evidenciado pelo seu coeficiente de variação na tabela de medidas resumo. No perfil médio, não nota-se diferença aparente entre os enxaguantes durante o início do tratamento. Todavia, o valor médio do escore de placa dentária decresce após 3 meses para todos enxaguantes, evidenciando queda mais acentuada para os do tipo A e B.

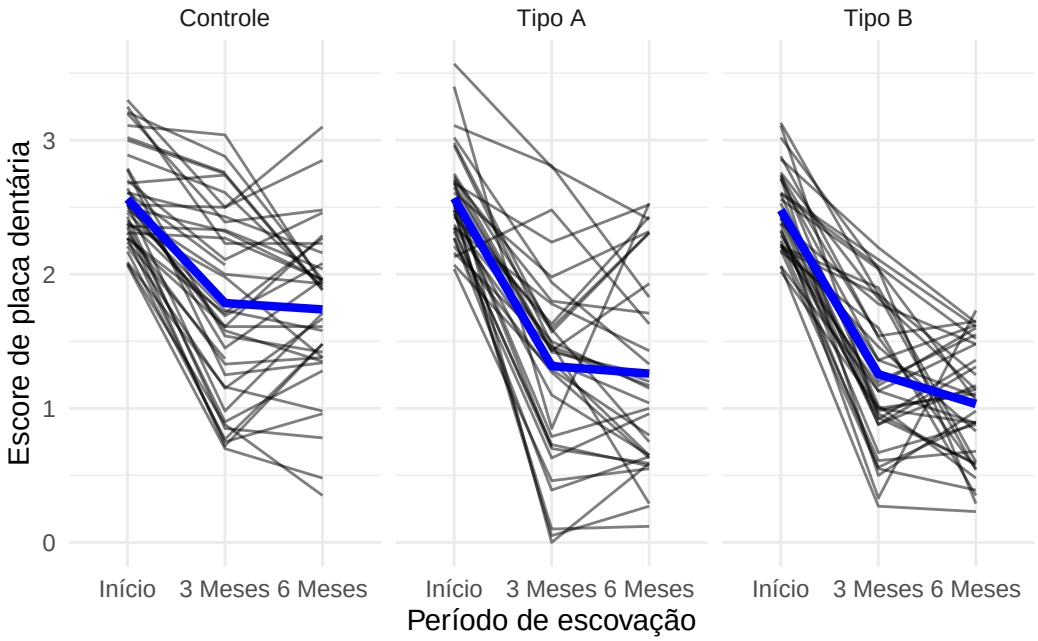

Figura 3: Perfis individuais e médios do escore de placa dentária para os enxaguantes bucais do tipo controle, A e B.

Ademais, os dados apresentam moderada dependência intraunidade, como revela o coeficiente de correlação entre seus valores defesados nos períodos de escovação na matriz de diagramas de dispersão na Figura 4. Vale destacar a relação linear positiva, $\hat{\rho} = 0,627$, entre o escore de placa dentária do período de 3 e 6 meses de escovação. Além disso, o enxaguante bucal controle exibiu a maior correlação intraunidade, por outro lado, o enxaguante bucal do tipo B conteve a menor. A inspeção da matriz de covariâncias e correlações amostrais sugere uma estrutura autoregressiva. Todavia, um exame detalhado da covariância dentro dos tratamentos indica a presença de heterogeneidade. Para verificar a hipótese de igualdade de matrizes de covariância dos tratamentos utilizamos o teste M de Box. O valor observado aproximado é $\chi^2 = 21,437$ com 12 graus de liberdade e valor descritivo $\hat{\alpha} = 0,04433$. Embora, sob o nível de significância de 5% possamos rejeitar a hipótese de igualdade das matrizes, o teste supracitado é sensível as suposições distribucionais. Portanto, por simplicidade, assumimos a homogeneidade das matrizes de covariância dos diferentes tratamentos e prosseguimos com a matriz de covariância combinada $\hat{\boldsymbol{\Omega}}_c$.

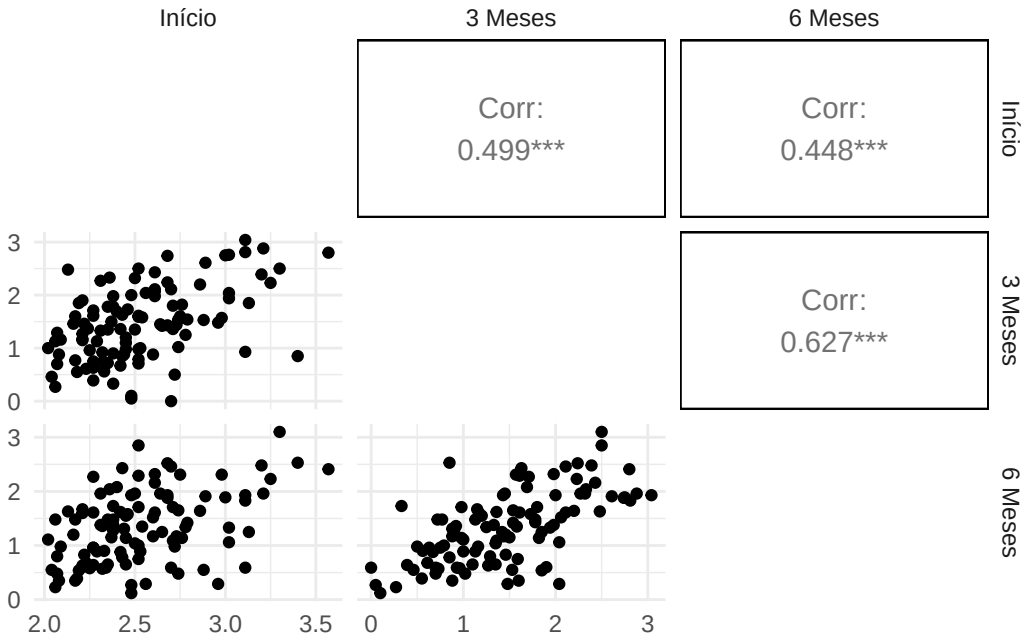

Figura 4: Matriz de diagramas de dispersão e correlações amostrais entre os valores defasados do escore de placa dentária de 105 indivíduos.

$$\hat{\mathbf{\Omega}}_c = \begin{bmatrix} 0,11245831 & 0,1074651 & 0,09129761 \\ 0,10746508 & 0,4338047 & 0,22354651 \\ 0,09129761 & 0,2235465 & 0,36342472 \end{bmatrix}$$

Embora pareça razoável supor que as curvas do segundo grau com efeitos aleatórios para interceptos, coeficientes angulares e quadráticos sejam candidatos adequados considerando a análise dos perfis individuais e médios. Entretanto, iniciaremos o ajuste de curvas de crescimento linear com efeitos fixos para o tipo de enxaguante bucal, para o periodo de escovação e a interação destes, assim como um efeito aleatório para o intercepto conforme Equação 7.

$$y_{ijk} = \beta_0 + \beta_1 \times \text{PERÍODO}_{ij} + \beta_2 \times \text{ENXAGUANTE}_{ij}^{(1)} + \beta_3 \times \text{ENXAGUANTE}_{ij}^{(2)}$$

$$+\beta_4 \times \text{ENXAGUANTE}_{ij}^{(3)} + \beta_5 \times \text{PERÍODO}_{ij} \times \text{ENXAGUANTE}_{ij}^{(1)}$$

$$+\beta_6 \times \text{PERÍODO}_{ij} \times \text{ENXAGUANTE}_{ij}^{(2)} + \beta_7 \times \text{PERÍODO}_{ij} \times \text{ENXAGUANTE}_{ij}^{(3)} + \tag{7}$$

Os resultados do ajuste do modelo supracitado está apresentado na e sugerem que

5 Considerações finais

Os modelos lineares multivariados com erros aleatórios correlacionados são uma importante abordagem para os modelos de regressão, principalmente por considerar a dependência inter-unidade e entre-unidade experimental.

Para o pesquisador, o modelo pode explicar de maneira realista o fenômeno de interesse, uma vez que consideramos que os dados estão sujeitos a correlação entre os seus valores desafados. A abordagem dos erros correlacionados também pode ser aplicada em variados contextos, aprimorando as diferentes formas de análise de dados.

Os modelos marginais e mistos proporcionam diferentes formas de estimação paramétrica usando tanto uma abordagem por máxima verossimilhança "tradicional" e restrita ou residual. A função de máxima verossimilhança restrita é extremamente importante para o desenvolvimento de medidas de diagnóstico e elaboração de testes de hipóteses. Por fim, vale salientar que o modelo de regressão misto possui algumas similaridades com o modelo linear geral. Testes de hipóteses para os efeitos fixos são realizados de maneira análoga ao modelo usual.

5.1 Trabalhos futuros

Como próximos trabalhos, podemos considerar modelos lineares generalizados mistos, admitindo um maior número de distribuições para a variável resposta. Além disso, toda a abordagem vista diz respeito a modelagem sob a suposição de linearidade. No entanto, a natureza da relação pode nos indicar o uso de outras funções de regressão, o que nos leva a classe de modelos não lineares mistos.

6 Referências

1. DIGGLE, P. J. HEAGERTY, P. J. LIANG, K. Y. ZEGER, S. L. Analsysis of Longitudinal Data. Oxford, New York, 2002.

2. POTTHOFF, R. F. ROY, S. N. A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika (1964), 51.

3. SINGER, J. M. NOBRE, J. S. ROCHA, F. M. Análise de dados longitudinais, Versão parcial preliminar, 2019.

4. VERBEKE, G. MOLENBERGHS, G. Linear mixed models for longitudinal data. Springer. 2000.

5. HADGU, A. KOCH, G. Application of generalized estimating equations to a dental randomized clinical trial. Journal of Biopharmaceutical Statistics, 9(1), 161–178 (1999)

6. HAND. J. D. TAYLOR. C. C. Multivariate Analysis of Variance and Repeated Measures: A Practical Approach for Behavioural Scientists. Chapman and Hall, London, 1987.

7. RAO, C. R. Linear Statistical Inference and its Applications. Wiley, 1973.

8. DEMIDENKO, E. Mixed models: theory and applications with R. Wiley, 2013.