Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Biológicas
Departamento de Estatística
Bacharelado em Estatística

Análise de Dados Longitudinais Utilizando a Linguagem SAS

Ludimilla Alves Viana

Ouro Preto-MG
Fevereiro 2024

Ludimilla Alves Viana

Análise de Dados Longitudinais Utilizando a Linguagem SAS

Monografia de Graduação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas e Biológicas da Universidade Federal de Ouro Preto como requisito parcial para a obtenção do grau de bacharel em Estatística.

Orientador

Prof. Dr. Eduardo Bearzoti

Universidade Federal de Ouro Preto – UFOP
Departamento de Estatística – DEEST

Ouro Preto-MG

Fevereiro 2024

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
COLEGIADO DO CURSO DE ESTATISTICA

FOLHA DE APROVAÇÃO

Ludimilla Alves Viana

Análise de dados longitudinais utilizando a linguagem SAS

Monografia apresentada ao Curso de Estatística da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Estatística.

Aprovada em 22 de fevereiro de 2024

Membros da banca

Dr. Eduardo Bearzoti - Orientador (Universidade Federal de Ouro Preto)
Dra. Carolina Silva Pena - Membro (Universidade Federal de Ouro Preto)
Dr. Fernando Luiz Pereira de Oliveira - Membro (Universidade Federal de Ouro Preto)

Prof. Dr. Eduardo Bearzoti, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 22/02/2024

Documento assinado eletronicamente por **Eduardo Bearzoti**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 23/02/2024, às 11:39, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Carolina Silva Pena**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 23/02/2024, às 13:47, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Fernando Luiz Pereira de Oliveira**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 23/02/2024, às 14:33, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0672403** e o código CRC **04044026**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: (31)3559-1312 - www.ufop.br

Agradecimentos

Em especial à minha mãe Maria José que sempre me incentivou nos estudos, às minhas irmãs por estarem comigo em todos os momentos e às minhas amigas que sempre estiveram presentes na minha vida acadêmica. A todos os amigos que fiz em Ouro Preto e também a todos os professores do Departamento de Estatística pela sabedoria que souberam transmitir. E por último ao meu orientador Eduardo pela valiosa contribuição e auxílio nessa reta final do curso.

"Não busque a felicidade fora, mas sim dentro de você, caso contrário nunca a encontrará."

Epicteto

Análise de Dados Longitudinais Utilizando a Linguagem SAS

Autor: Ludimilla Alves Viana
Orientador: Prof. Dr. Eduardo Bearzoti

Resumo

Os chamados Dados Longitudinais consistem, em uma pesquisa científica, de observações que correspondem a diferentes avaliações feitas em um mesmo indivíduo, ou mesmas unidades experimentais. Tais dados permitem decompor a variação residual em componentes entre e dentro de indivíduos. Geralmente estas avaliações correspondem a medições feitas em diferentes momentos do tempo, mas não necessariamente, e daí tais dados também serem por vezes chamados de dados de medidas repetidas. Dados longitudinais surgem em pesquisas de diversas áreas do conhecimento, tais como Ciências da Saúde, Ciências Econômicas e Ciências Agrárias. Este trabalho teve como objetivo uma apresentação e caracterização da análise de dados longitudinais, considerando diferentes possibilidades de ajustamento, e utilizando a linguagem SAS. Embora seja uma linguagem comercial, desde há alguns anos o Instituto SAS disponibilizou uma versão *online*, gratuita, destinada a fins acadêmicos. Neste trabalho, optou-se por ilustrar o ajuste de modelos de dados longitudinais através de dois exemplos básicos, que poderiam ser considerados típicos em uma grande variedade de situações. O primeiro exemplo correspondeu a um conjunto de dados simulados, emulando uma situação de dados pareados. Com este exemplo, é discutida a relação que existe entre o teste t pareado e a análise de dados longitudinais. É apresentado o seu ajustamento com a linguagem SAS, ilustrando duas possibilidades quanto ao modelo estatístico, no que se refere à dependência existente dentro de cada par. Esta dependência pode ser modelada tanto a partir dos resíduos, admitindo uma dependência residual, como também por um modelo misto, admitindo que o efeito dos indivíduos seja de natureza aleatória. O segundo exemplo se referiu a um experimento longitudinal, realizado no LABIIN-UFOP, em que cobaias (camundongos) infectadas por *Trypanosoma cruzi* foram submetidas a duas dietas, uma dieta regular, e uma dieta rica em gordura.

Entre os dias 9 e 29 após a infecção, cada cobaia teve o seu nível de parasitemia avaliado, diariamente. A parasitemia consiste na concentração de parasitas por unidade de volume de sangue. O fato de serem avaliações diárias é que caracteriza o experimento como longitudinal, e o fato de o número de tempos ser bem maior, isto possibilitou a consideração de diferentes estruturas de covariância residual. O Critério de Informação de Akaike indicou que a estrutura autorregressiva de ordem 1 propiciou o melhor ajuste. A linguagem SAS mostrou-se uma ferramenta versátil e eficaz, com uma codificação simples. Uma vez que os códigos também são aqui fornecidos, espera-se que este trabalho possa servir de material de consulta para iniciantes na análise de dados longitudinais que estejam interessados na linguagem SAS.

Palavras-chave: Análise de Dados Longitudinais, Medidas Repetidas, Linguagem SAS.

Longitudinal Data Analysis Using the SAS Language

Author: Ludimilla Alves Viana
Advisor: Dr. Eduardo Bearzoti

Abstract

The so-called Longitudinal Data correspond, in scientific research, to observations that correspond to different assessments made on the same individual, or the same experimental units. Such data allow us to decompose the residual variation into components between and within individuals. Generally, these assessments correspond to measurements taken at different moments in time, but not necessarily, and hence such data are also sometimes called repeated measurements data. Longitudinal data arise in research from different areas of knowledge, such as Health Sciences, Economic Sciences and Agricultural Sciences. This work aimed to present and characterize the analysis of longitudinal data, considering different adjustment possibilities, and using the SAS language. Although it is a commercial language, for some years now the SAS Institute has made a free online version available, intended for academic purposes. In this work, we chose to illustrate the fitting of longitudinal data models through two basic examples, which could be considered typical in a wide variety of situations. The first example corresponded to a set of simulated data, emulating a paired data situation. With this example, the relationship between the paired t test and longitudinal data analysis is discussed. Its adjustment with the SAS language is presented, illustrating two possibilities as statistical models, with regard to the dependence within each pair. This dependence can be modeled both based on residuals, assuming a residual dependence, and also using a mixed model, assuming that the effect of individuals is random in nature. The second example referred to a longitudinal experiment, carried out at LABIIN-UFOP, in which animals (mice) infected by *Trypanosoma cruzi* were subjected to two diets, a regular diet and a high-fat diet. Between days 9 and 29 after infection, each animal had its parasitemia level assessed daily. Parasitemia consists of the concentration of parasites per unit volume of blood. The fact that the assessments were made on a daily basis is what characterizes the experiment as longitudinal, and the fact that the number

of times was much greater, this made it possible to consider different residual covariance structures. The Akaike Information Criterion indicated that the autoregressive structure of order 1 provided the best adjustment. The SAS language proved to be a versatile and effective tool, with simple coding. Since the codes are also provided here, it is hoped that this work can serve as reference material for beginners in longitudinal data analysis who are interested in the SAS language.

Keywords: Longitudinal Data Analysis, Repeated Measures Data, SAS Language.

Lista de figuras

Lista de tabelas

Sumário

1 Introdução

Os chamados *dados longitudinais* correspondem a uma categoria de observações em que há medidas repetidas em uma mesma unidade experimental. Um caso particular de medidas repetidas consiste em um projeto experimental com dados coletados em uma sequência de pontos espaçados no tempo, avaliados em mesmas unidades experimentais, ou parcelas. Tais pontos podem estar igualmente ou desigualmente espaçados no tempo.

Experimentos com medidas repetidas são bastante comuns, sendo utilizados por pesquisadores de diversas áreas, como nas Ciências da Saúde, Ciências Econômicas e Ciências Agrárias, quando se tem o objetivo de avaliar o comportamento de indivíduos, ou unidades observacionais de outra natureza, ao longo do tempo. Estes estudos têm a vantagem de requerer um número menor de unidades amostrais. Além disso, diminuem a variabilidade decorrente de diferenças individuais, e permitem avaliar mudanças que ocorrem dentro e entre as unidades amostrais com mais eficiência. De maneira geral, quaisquer dados medidos repetidamente ao longo do tempo ou espaço são dados de medidas repetidas.

Estudos longitudinais se contrapõem aos chamados *estudos transversais*, em que as avaliações são realizadas em um momento fixado no tempo. Os estudos longitudinais permitem que a variável resposta seja observada em unidades amostrais, considerando, além do tempo, outras covariáveis que podem influenciá-la. Planejamentos longitudinais proveem informações sobre variações individuais nos níveis da variável resposta. É importante notar que alguns parâmetros dos modelos estatísticos subjacentes podem ser estimados de forma mais eficiente sob planejamentos longitudinais do que sob planejamentos transversais, com o mesmo número de observações.

Do ponto de vista estatístico, o que torna os dados longitudinais particulares é o fato de haver uma dependência entre observações repetidas em uma mesma unidade experimental (por exemplo, diferentes avaliações no tempo em um mesmo indivíduo). Isto viola a pressuposição, geralmente feita, de independência entre observações, demandando métodos de estimação que levem esta dependência em conta.

Embora relativamente comum, a análise de dados longitudinais nem sempre é abordada, ordinariamente, em cursos de graduação em Estatística. No âmbito da Universidade Federal de Ouro Preto, por exemplo, esta disciplina tem caráter eletivo. Desta maneira, este trabalho teve como objetivo uma apresentação e caracterização da análise de dados longitudinais, considerando diferentes possibilidades de ajustamento, e utilizando a linguagem SAS. Apesar de ser uma linguagem comercial, atualmente é disponibilizada uma versão *online*, gratuita, destinada a fins acadêmicos. A análise de dados longitudinais com esta linguagem é relativamente simples, permitindo especificar diferentes estruturas de dependência entre observações.

2 Referencial Teórico

O desenvolvimento de métodos estatísticos para a análise de dados tem aumentado significativamente nas últimas décadas, em grande parte devido ao desenvolvimento dos recursos e de ferramentas computacionais. Muitos desse estudos estão voltados para a avaliação do comportamento de uma ou mais variáveis resposta ao longo do tempo, ao qual chamamos de *estudos longitudinais*.

Em língua portuguesa, uma descrição detalhada acerca dos estudos longitudinais pode ser encontrada em Singer et al. (2018), da qual alguns aspectos mais relevantes são destacados a seguir.

Estudos longitudinais são métodos de análise de medidas repetidas onde, em geral, tratamentos são atribuídos a unidades experimentais, e os dados são coletados em uma sequência de tempos de cada unidade experimental. Devido à grande utilização de medições repetidas em Ciências da Saúde, as unidades experimentais são frequentemente referenciadas como sujeitos, ou indivíduos. São estudos, assim, onde em geral existem dois fatores a serem considerados, tratamentos e tempo (bem como sua interação).

Nesse sentido, pode-se dizer que os experimentos de medidas repetidas poderiam ser considerados experimentos fatoriais, ou experimentos de parcelas subdivididas no tempo. O tratamento é chamado de fator entre sujeitos, pois corresponderia ao fator de parcela, sendo cada indivíduo uma "parcela" (uma unidade experimental). O tempo é chamado de fator dentro de sujeitos, pois o mesmo sujeito é avaliado em diferentes momentos, correspondendo, assim, ao fator de subparcela (dentro de parcelas).

Em experimentos de medidas repetidas, o interesse se concentra em verificar como os tratamentos diferem, e se estas diferenças mudam com o tempo. A análise de dados de medidas repetidas tem como objetivo comparar médias de tratamentos ou curvas de regressão de tratamento ao longo do tempo, sendo necessário um esforço adicional na análise estatística, para identificar uma estrutura adequada de covariância para os dados.

A maior desvantagem dos estudos longitudinais está relacionada com seu custo, pois

em muitas situações exige-se um grande esforço para garantir a observação das unidades amostrais nos instantes pré-determinados e, em outras situações, o período de observação pode ser muito longo. Em muitos ensaios clínicos, por exemplo, é necessário acompanhar os pacientes com extremo cuidado para que cumpram o protocolo experimental e não abandonem o estudo. Os aspectos técnicos também podem ser considerados como uma desvantagem, pois a análise estatística de dados obtidos sob esse tipo de planejamento é, em geral, mais difícil que a análise de dados obtidos sob esquemas transversais.

Em geral, alguns dos problemas ou inferências de interesse com os quais nos deparamos no contexto de estudos longitudinais são similares àqueles com que nos defrontamos em estudos transversais. Para dados com distribuição normal, eles podem ser classificados como problemas de Análise de Variância (ANOVA) ou Análise de Regressão (linear ou não linear); ver, por exemplo, Montgomery (2012). A diferença básica entre eles reside numa possível dependência (estatística) entre as observações amostrais, presente apenas nos dados provenientes de estudos longitudinais. A consequência prática desse tipo de dependência reflete-se às vezes num fenômeno conhecido como trilhamento (*tracking*), segundo o qual unidades amostrais com níveis de resposta mais altos (ou mais baixos) no início da coleta de observações tendem a manter suas posições relativas ao longo de todo o estudo. O esforço adicional requerido na análise de dados longitudinais, relativamente àquele exigido em estudos transversais, concentra-se praticamente na modelagem dessa estrutura de dependência estatística.

Usar uma ANOVA padrão neste caso não é apropriado porque falha em modelar a correlação entre as medidas repetidas. A análise desses experimentos pode ser realizada, por exemplo, com o programa SAS, usando os procedimentos PROC GLM ou o PROC MIXED. Nos últimos anos, a utilização deste programa se tornou mais acessível, desde a sua disponibilização gratuita, em formato *online* (SAS Institute, 2015). Requere-se do usuário apenas a realização de um cadastro.

Três tipos gerais de análise estatística são mais usados para medidas repetidas. Um método trata os dados de medidas repetidas como vindo de um experimento de parcelas subdivididas. Esse método, geralmente chamado de análise de variância univariada, pode ser implementado na linguagem SAS usando PROC GLM, com a instrução RANDOM. Outro método aplica métodos de análise multivariada e univariada a transformações lineares das medidas repetidas. As transformações lineares podem ser médias, diferenças entre respostas em diferentes pontos de tempo, inclinações de curvas de regressão, etc. Essas técnicas são invocadas pela instrução REPEATED em PROC GLM. O terceiro mé-

todo aplica métodos baseados em um modelo misto com estrutura paramétrica especial nas matrizes de covariância. Este tipo de metodologia tem sido computacionalmente viável apenas nas últimas décadas. É aplicado em PROC MIXED, normalmente usando a instrução RANDOM.

O SAS apresenta grande flexibilidade para o ajuste de modelos mistos, destacando-se a excelente performance do procedimento MIXED. O livro de Littel et al. (1996) pode ser considerado uma obra já clássica, acerca do ajuste de modelos mistos no programa SAS, com numerosos exemplos. O capítulo 3 trata da análise de dados de medidas repetidas, com um exemplo em que são comparados três programas de condicionamento físico (levantamento de peso), sendo que diferentes voluntários foram submetidos a estes programas, sendo avaliados em diferentes momentos no tempo. Trata-se de um exemplo padrão de dados longitudinais, em que há um fator de tratamento (os programas de condicionamento), bem como o tempo.

Além de apresentar importantes considerações práticas sobre a realização da análise de dados longitudinais no programa, a obra de Littel et al. (1996) também traz importantes considerações teóricas sobre o ajustamento de modelos mistos. Alguns dos aspectos mais relevantes são destacados a seguir.

Modelos mistos são utilizados para descrever dados de experimentos cuja estrutura de tratamentos envolve alguns fatores que são fixos e alguns que são aleatórios, ou seja, modelos lineares que contêm efeitos fixos e aleatórios. Um fator de efeito fixo em geral tem seus níveis escolhidos, e a inferência a ser feita se restringe aos níveis considerados no estudo. Já um fator de efeito aleatório tem seus níveis sorteados de uma população de referência. Ou seja, a inferência irá se referir à população de níveis do fator, da qual os níveis considerados no estudo correspondem a uma amostra. No exemplo apresentado por Littel et al. (1996), os programas de condicionamento físico consistem de níveis de um fator de efeito fixo (bem como o tempo), enquanto que os efeitos dos voluntários participantes consistem de níveis de um fator aleatório, pois tais voluntários correspondem a uma amostra dos possíveis usuários de tais programas de condicionamento.

Antes do desenvolvimento de algoritmos e recursos computacionais específicos, os modelos mistos eram tradicionalmente analisados por meio de procedimentos da análise de variância, que essencialmente ignoram a natureza aleatória dos efeitos aleatórios. Ou seja, tais modelos eram tratados efetivamente como se fossem fixos. Um exemplo de modelo fixo seria como aquele em 2.1.

$$y_{ijk} = \mu + \alpha_i + \tau_j + (\alpha\tau)_{ij} + e_{ijk} \tag{2.1}$$

em que y_{ijk} é a variável resposta, μ a constante do modelo, α_i é efeito de um fator de efeito fixo (como o programa de condicionamento), τ_j o efeito de um segundo fator de efeito fixo (como o tempo), e $(\alpha\tau)_{ij}$ a interação entre estes. O modelo 2.1 não seria o mais adequado para analisar os dados do exemplo apresentado por Littel et al. (1996), pois ignora que um mesmo voluntário é avaliado em diferentes tempos, tratando os dados como observações independentes de um ensaio fatorial (por exemplo, Montgomery, 2012).

De qualquer forma, um modelo fixo (ou admitido como fixo), pode ser expresso matricialmente como:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{e} \tag{2.2}$$

sendo $\mathbf{y}$ o vetor de observações, $\mathbf{X}$ a matriz de incidência dos fatores fixos, $\boldsymbol{\theta}$ o vetor de parâmetros de efeitos fixos, e $\mathbf{e}$ o vetor de resíduos.

Se são consideradas atendidas as pressuposições usuais dos modelos fixos (independência, normalidade, e homogeneidade de variâncias), então o vetor de parâmetros $\boldsymbol{\theta}$ pode ser estimado pelo método dos quadrados mínimos ordinário, que resulta na solução do sistema de equações 2.3, conhecido como *sistema de equações normais*.

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\theta}} = \mathbf{X}'\mathbf{y} \tag{2.3}$$

Em 2.3, a notação $'$ se refere à operação de transposição.

Se, por outro lado, a pressuposição de independência é violada, haverá uma matriz de variâncias e covariâncias residuais referente ao vetor $\mathbf{y}$, matriz essa que pode ser representada por $\mathbf{R}$. Neste caso, o método de estimação mais apropriado não será mais o método de quadrados mínimos ordinário, mas sim o método dos quadrados mínimos generalizado, que consistirá na solução do sistema 2.4.

$$\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\hat{\boldsymbol{\theta}} = \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \tag{2.4}$$

A dificuldade inerente ao sistema 2.4 é a de que, em geral, a matriz $\mathbf{R}$ envolverá um ou mais parâmetros desconhecidos, o que tornará o sistema não-linear, exigindo assim métodos numéricos para a sua resolução. No caso dos modelos mistos (com ou sem dependência residual) em geral também serão necessários métodos numéricos iterativos para o

seu ajustamento, a não ser nos casos mais elementares.

Um exemplo de modelo misto seria aquele apresentado em 2.5.

$$y_{ijk} = \mu + \alpha_i + d_{i(j)} + \tau_k + (\alpha\tau)_{ik} + e_{ijk} \tag{2.5}$$

Este modelo é muito semelhante ao modelo 2.1, com a única diferença de que foi incluído o efeito $d_{i(j)}$, que é o efeito (aleatório) do voluntário j que participa do programa de condicionamento i. Na literatura de modelos mistos, é comum utilizar letras gregas para representar efeitos fixos, e letras latinas para designar efeitos aleatórios, como feito em 2.5. Estes efeitos $d_{i(j)}$ são tidos como vindos de uma população de referência (geralmente admitindo-se distribuição normal), com vetor de médias igual a um vetor de zeros, e matriz de variâncias e covariâncias representada por $\mathbf{G}$. Esta matriz apresentará covariâncias, constantes, referentes a quaisquer pares de observações com o mesmo nível do fator aleatório, por exemplo, duas observações de um mesmo indivíduo. Esta estrutura de covariâncias corresponderá à estrutura de simetria composta, detalhada mais adiante.

Um modelo como o 2.5 pode ser expresso matricialmente como em 2.6.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{Zu} + \mathbf{e} \tag{2.6}$$

Em 2.6, a diferença em relação a 2.2 é a inclusão da matriz $\mathbf{Z}$ de incidência dos parâmetros de efeito aleatório, por sua vez presentes no vetor $\mathbf{u}$.

Para o ajustamento de um tal modelo, o método mais utilizado é o da máxima verossimilhança restrita, tido hoje em dia como um método padrão, na abordagem frequentista. Essencialmente, e já considerando a possibilidade de dependência residual entre observações, a utilização da máxima verossimilhança restrita pode ser representada como sendo a solução do sistema 2.7, conhecido como sistema de *equações do modelo misto de Henderson.*

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\theta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y} \end{bmatrix} \tag{2.7}$$

Da mesma forma que em 2.4, em 2.7 as matrizes $\mathbf{R}$ e $\mathbf{G}$ quase sempre conterão termos desconhecidos, que precisam ser estimados, o que torna o sistema não linear, demandando assim métodos numéricos para sua resolução. Dois dos métodos mais utilizados para tal são o chamado algoritmo "EM" de esperança-maximização (Dempster et al., 1977; Foulley

et al., 2000), e o método de Newton-Raphson (Lindstrom e Bates, 1988). Este último é o método utilizado pelo procedimento PROC MIXED, do programa SAS, conforme algoritmo descrito por Wolfinger et al. (1994).

Conforme mencionado anteriormente, uma das grandes dificuldades da análise de dados longitudinais é a modelagem da estrutura de covariâncias residuais. Em outras palavras, definir a estrutura da matriz **R**. No programa SAS, isto é facilmente implementado através da instrução REPEATED, do procedimento PROC MIXED. Três das estruturas de covariâncias mais utilizadas são:

- Simetria composta (abreviatura inglesa: CS);
- Autorregressiva de ordem 1 (abreviatura inglesa: AR(1));
- Não-estruturada (abreviatura inglesa: UN).

Tais abreviaturas inglesas são justamente os códigos utilizados na instrução REPEATED. A estrutura de simetria composta CS é aquela em que qualquer par de níveis de um fator aleatório apresenta a mesma covariância. Esta é a estrutura considerada, por exemplo, em ensaios de parcelas subdivididas, onde admite-se que a covariância entre quaisquer duas subparcelas dentro de uma mesma parcela seja constante.

Em função disso, no SAS, para ajustar um modelo considerando a estrutura CS, há duas maneiras. Pode-se especificar o efeito de "parcelas" (voluntários, por exemplo) como sendo aleatório, na instrução RANDOM. A segunda maneira consiste em utilizar a instrução REPEATED, com a opção CS.

Entretanto, em dados longitudinais, em geral é mais apropriado considerar que as covariâncias residuais vão diminuindo, à medida que os dois tempos considerados vão ficando mais afastados. Uma maneira de se modelar este comportamento é o de se considerar estruturas típicas de Séries Temporais, como a estrutura autorregressiva de ordem 1 (ou ordens superiores), ou de médias móveis.

Finalmente, a estrutura de covariâncias mais geral seria a não-estruturada, em que se admite uma covariância diferente para cada par de tempos. Embora seja a estrutura mais flexível, geralmente corresponde a um modelo pouco parcimonioso, dado o grande número de parâmetros a serem estimados, o que, inclusive, pode acarretar dificuldades de convergência, no processo numérico de ajustamento.

Apenas a título de exemplificação, considere-se apenas as observações referentes à avaliação de um voluntário em 4 tempos. A matriz de covariâncias residuais referentes a

estas observações são dadas como segue, para estas três estruturas. Para a estrutura CS, teríamos a matriz:

$$\begin{bmatrix} \sigma^2+\sigma_g^2 & \sigma_g^2 & \sigma_g^2 & \sigma_g^2 \\ \sigma_g^2 & \sigma^2+\sigma_g^2 & \sigma_g^2 & \sigma_g^2 \\ \sigma_g^2 & \sigma_g^2 & \sigma^2+\sigma_g^2 & \sigma_g^2 \\ \sigma_g^2 & \sigma_g^2 & \sigma_g^2 & \sigma^2+\sigma_g^2 \end{bmatrix}$$

sendo σ^2 a variância residual e σ_g^2 o componente de variância referente à população de voluntários. Percebe-se que este componente corresponde às covariâncias, constantes para qualquer par de tempos, mais afastados, ou menos afastados.

Para a estrutura AR(1), teríamos a seguinte matriz de covariâncias referentes a estas 4 observações:

$$\sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

sendo ρ um parâmetro de autocorrelação. Sendo um valor entre 0 e 1 (em geral as autocorrelações são positivas), percebe-se que a dependência residual vai diminuindo, à medida que dois tempos vão ficando mais distanciados entre si.

Finalmente, para o caso não estruturado, teríamos a matriz:

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_{44} \end{bmatrix}$$

Note-se que a representação acima considera não apenas covariâncias diferentes, mas ainda variâncias heterogêneas conforme o tempo (havendo diferentes parâmetros na diagonal), sendo bastante geral.

3 Metodologia

Neste trabalho, optou-se por ilustrar o ajuste de modelos de dados longitudinais através de dois exemplos básicos, que poderiam ser considerados típicos em uma grande variedade de situações. Em ambos os casos, o ajustamento foi feito utilizando a linguagem SAS. Embora esta seja uma linguagem comercial, desde alguns anos existe uma versão *online*, gratuita, destinada a finalidades acadêmicas (SAS Institute, 2015). Trata-se da plataforma "SAS® OnDemand for Academics".

O primeiro exemplo se refere a um conjunto de dados pareados, o que pode ser considerado como sendo a situação mais simples de dados longitudinais, pois cada "par" se refere a duas observações realizadas em um mesmo indivíduo, ou em uma mesma unidade experimental. Nas Ciências da Saúde, este pareamento geralmente consiste em uma avaliação antes e uma avaliação depois da aplicação de determinado tratamento, como um medicamento. Por exemplo, para avaliar o efeito de um novo medicamento para normalização da pressão arterial, pode-se medir esta pressão em diferentes indivíduos, antes e depois da aplicação deste novo medicamento. Percebe-se aqui que se tratam de observações temporais (antes e depois), mas dados pareados também podem surgir em outras situações, como, por exemplo, na avaliação da pressão intraocular dos olhos esquerdo e direito de diferentes indivíduos. Independentemente de o pareamento ser temporal ou não, do ponto de vista do modelo estatístico em ambos os casos tem-se uma dependência entre as observações dentro de cada par.

Em dados pareados, em geral espera-se que esta dependência se reflita em uma covariância positiva, devido ao fenômeno de trilhamento, mencionado anteriormente. Assim, este primeiro exemplo correspondeu a um conjunto de dados pareados simulado, tendo a simulação sido feita de maneira que a covariância dentro dos pares fosse positiva.

A motivação deste exemplo correspondeu a uma situação fictícia em que um nutricionista teria desenvolvido um cardápio para redução da pressão arterial, desejando-se verificar se esse cardápio seria eficiente. Foi simulado um grupo de 10 pacientes com um

histórico de pressão alta, que teriam utilizado o cardápio por alguns meses. A pressão sistólica destes pacientes (em mmHg) teria sido medida antes e depois da utilização deste cardápio. Estes dados simulados estão apresentados na Tabela 1.

Tabela 1: Exemplo de dados pareados, simulando uma situação em que 10 pacientes teriam tido sua pressão arterial sistólica avaliada antes e depois da utilização de uma dieta destinada à redução da pressão arterial.

Paciente	Antes	Depois
1	164	157
2	167	156
3	143	148
4	149	145
5	134	135
6	150	135
7	163	151
8	135	140
9	156	152
10	151	138

Em seguida, foi considerado um segundo exemplo, real, considerando agora um estudo longitudinal envolvendo um número bem maior de tempos de avaliação. Este estudo foi realizado pelo LABIIN[1] (Laboratório de Imunobiologia da Inflamação), do Departamento de Ciências Biológicas da Universidade Federal de Ouro Preto, tendo sido os dados gentilmente cedidos pelos autores. Este trabalho de pesquisa resultou em uma publicação (Figueiredo *et al.*, 2018), onde informações detalhadas sobre o estudo podem ser encontradas. O LABIIN tem como linha principal de pesquisa o estudo da resposta inflamatória associada a doenças parasitárias, como o protozoário *Trypanosoma cruzi*, considerado neste trabalho de pesquisa em questão, utilizando cobaias.

Neste trabalho, Figueiredo *et al.* (2018) tiveram como ponto de partida o fato de que dietas ricas em gordura podem desencadear doenças metabólicas e cardiovasculares. Dois grupos de cobaias (camundongos machos da linhagem C57BL/6) foram formados, sendo que o primeiro recebeu uma dieta regular, e o segundo grupo recebeu uma dieta rica em gordura. Cada grupo foi tratado com a respectiva dieta por um período de 8 semanas, quando então foram infectados com a linhagem VL-10 de *Trypanosoma cruzi*.

Após essa infecção, cada respectiva dieta foi mantida ainda por um período de cerca de um mês, durante o qual os camundongos foram avaliados quanto ao nível de parasitemia. A parasitemia consiste na quantidade de parasitas presentes na corrente sanguínea. Esta

[1] `https://sites.ufop.br/labiin/`

avaliação foi feita coletando-se sangue da veia da cauda de cada camundongo, diariamente, entre os dias 9 e 29 após a infecção, tendo sido expressa como o número de parasitas $\times 10^3$ por 0,1 mL de sangue. Em seguida, os camundongos foram sacrificados, para a realização de outras avaliações.

Esta avaliação diária nos mesmos camundongos é que caracteriza este estudo como sendo longitudinal. O conjunto de dados em questão é ligeiramente desbalanceado, no sentido de que o número de cobaias efetivamente acabou sendo diferente em cada grupo (7 camundongos no grupo da dieta regular, e 10 camundongos no grupo da dieta rica em gordura), bem como o fato de que nem todas as cobaias sobreviveram até o $29\,\underline{0}$ dia após a infecção. No entanto, uma das grandes vantagens da linguagem SAS é a facilidade de se lidar com dados desbalanceados.

Este segundo exemplo, possuindo um número bem maior de tempos, permitiu considerar diferentes estruturas de covariância no ajustamento, utilizando o procedimento PROC MIXED da linguagem SAS. Estas diferentes estruturas foram comparadas quanto à qualidade do ajustamento, conforme o chamado Critério de Informação de Akaike (ver, por exemplo, Littel *et al.*, 1996), comumente abreviado pela sigla inglesa AIC. Este critério pode ser definido como em 3.1. Nesta definição, $\ln L(\boldsymbol{\theta}|\mathbf{y})$ é a log-verossimilhança (estimada) do modelo ajustado, e P é o número de parâmetros considerados no ajustamento.

$$\mathrm{AIC} = -2\left(\ln L(\boldsymbol{\theta}|\mathbf{y}) - P\right) \tag{3.1}$$

Assim, conforme 3.1, quanto menor o valor de AIC, melhor o ajustamento. Por vezes, o AIC é definido na literatura sem o sinal negativo e, neste caso, inverte-se a interpretação. O critério de Akaike admite como um bom ajustamento aqueles com alto $\ln L(\boldsymbol{\theta}|\mathbf{y})$, mas com uma penalização quanto ao número de parâmetros. Quanto maior o número de parâmetros, maior o AIC, o que não é desejado.

4 Resultados e Discussão

Neste Capítulo, são apresentadas as análises estatísticas, e sua discussão, correspondentes a dois exemplos de conjunto de dados longitudinais.

4.1 Exemplo 1 (Dados Pareados)

O primeiro exemplo correspondeu a um conjunto de dados simulados, emulando uma situação de dados pareados. Dados pareados são relativamente frequentes nas chamadas Ciências da Saúde, sendo assim comumente abordados em livros-texto de Bioestatística (ver, por exemplo, Siqueira e Tibúrcio, 2011). A análise padrão para este tipo de dado (admitindo-se distribuição normal) consiste no chamado *teste t para dados pareados*. A seguir, é apresentada esta análise padrão, bem como uma análise ignorando o pareamento, discutindo suas relações com a Análise de Dados Longitudinais.

A análise padrão de dados pareados consiste na realização de um teste t das diferenças observadas em cada par. Estas diferenças são tratadas como uma nova variável, sendo testada a hipótese de nulidade H_0 de que a média desta nova variável é igual a zero.

Se os pares corresponderem a dois tempos de observação em um mesmo indivíduo, estas diferenças podem ser calculadas tanto considerando o valor do tempo posterior menos o valor do tempo anterior, como o contrário. O único cuidado a ser observado é que, se o teste for unilateral, a hipótese alternativa deve ser elaborada de maneira coerente com a maneira com que as diferenças foram calculadas.

A motivação deste primeiro exemplo, contendo dados simulados, foi a do interesse em investigar se uma determinada dieta seria eficiente em reduzir a pressão arterial sistólica em pacientes com histórico de pressão elevada. Estes dados foram apresentados no Capítulo anterior, mas estão reproduzidos novamente na Tabela 2, com uma coluna adicional, contendo as diferenças (na Tabela designadas por D) entre os valores do tempo anterior (X_1) e os do tempo posterior (X_2).

Tabela 2: Exemplo de dados pareados em dois tempos, um tempo anterior X_1 e um tempo posterior X_2, bem como as diferenças $D = X_1 - X_2$.

Voluntário	X_1	X_2	$D = X_1 - X_2$
1	164	157	7
2	167	156	11
3	143	148	-5
4	149	145	4
5	134	135	-1
6	150	135	15
7	163	151	12
8	135	140	-5
9	156	152	4
10	151	138	13

Nesta Tabela 2, como as diferenças foram calculadas considerando os valores do tempo anterior menos os do tempo posterior, faz sentido elaborar uma hipótese alternativa especificando que a média das diferenças seja maior que zero. Ou seja, se o cardápio foi eficiente, espera-se que tenha reduzido a pressão no tempo posterior ao uso da dieta, de maneira a promover uma diferença média positiva. Assim, designando esta diferença média como $\mu_{_D}$, tem-se que o teste de hipóteses adequado seria um teste unilateral à direita:

$$\begin{cases} H_0: & \mu_{_D} \leq 0 \\ H_1: & \mu_{_D} > 0 \end{cases}$$

Caso as diferenças tivessem sido calculadas tomando-se os valores do tempo posterior menos os do tempo anterior, o teste adequado seria unilateral à esquerda.

O chamado teste t para dados pareados nada mais é do que um teste t usual, trabalhando-se com as diferenças D. Assim, a estimativa da diferença média, considerando as diferenças apresentadas na Tabela 2, é:

$$\hat{\mu}_{_D} = \bar{d} = \frac{55}{10} = 5,5$$

A variância amostral destas diferenças é $s_d^2 = 54,2778$, com desvio padrão $s_d = \sqrt{s_d^2} = 7,3673$. Com isso, a estatística de teste do teste t pareado (evidenciando o fato de que o

teste está sendo feito em relação ao valor 0) é dada conforme 4.1.

$$t_c = \frac{\bar{d} - 0}{\frac{s_d}{\sqrt{n}}} = \frac{5,5 - 0}{\frac{7,3673}{\sqrt{10}}} = 2,361 \tag{4.1}$$

Tendo-se escolhido um nível de significância $\alpha = 0,05$, tem-se que o valor de t tabelado, com 9 graus de liberdade (pois são dez valores de diferença), é igual a $t_{_{0,05}} = 1,833$. Assim, como $2,361 > 1,833$, rejeita-se H_0 a 5% de probabilidade. Se o conjunto de dados fosse real, concluiríamos que o cardápio foi eficiente.

O teste t para dados pareados pode ser considerado como a abordagem mais elementar de Análise de Dados Longitudinais. Como a análise se baseia nas diferenças entre os dois tempos, implicitamente leva em conta a dependência entre estes. Isto pode ser evidenciado na expressão da variância da média das diferenças $\bar{d}$, apresentada em 4.2.

$$\begin{aligned} \mathrm{V}(\bar{d}) = \mathrm{V}\left[\frac{\sum_{i=1}^{n}(X_{1i} - X_{2i})}{n}\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{V}(X_{1i} - X_{2i}) = \\ = \frac{1}{n^2}\sum_{i=1}^{n}[\mathrm{V}(X_1) + \mathrm{V}(X_2) - 2\mathrm{Cov}(X_1, X_2)] = \frac{\mathrm{V}(X_1) + \mathrm{V}(X_2) - 2\mathrm{Cov}(X_1, X_2)}{n} \end{aligned} \tag{4.2}$$

Em um teste t convencional (dados não pareados), admite-se que os dados de cada um dos dois grupos correspondem a duas amostras independentes. Se considerássemos os dois tempos como independentes, a inferência seria feita a partir da diferença entre as duas médias amostrais (e não da média das diferenças), com variância dada conforme 4.3. Nesta expressão, está-se admitindo que as duas amostras têm o mesmo tamanho n.

$$\mathrm{V}(\bar{X}_1 - \bar{X}_2) = \frac{\mathrm{V}(X_1) + \mathrm{V}(X_2)}{n} \tag{4.3}$$

Comparando-se 4.2 e 4.3, verifica-se que de fato o teste t para dados pareados leva em conta a dependência (estando presente a covariância), enquanto que o teste t convencional a ignora. Em geral, espera-se que esta covariância seja positiva. Dessa forma, embora $\bar{d}$ e $(\bar{X}_1 - \bar{X}_2)$ sejam numericamente iguais, suas variâncias não o são, e, caso de fato a covariância seja positiva, isto resultará numa menor variância para $\bar{d}$, o que em geral leva a um teste t mais poderoso, mesmo tendo um número de graus de liberdade menor.

Isto pode ser ilustrado com este primeiro exemplo. Conforme mencionado no Capítulo de Metodologia, este conjunto de dados foi simulado de maneira que houvesse uma covariância positiva entre os dois tempos. De fato, com tais dados, a covariância amostral está apresentada em 4.4.

$$\text{Cov}(X_1, X_2) = \frac{\sum_{i=1}^{n} X_{1i}X_{2i} - \frac{\sum_{i=1}^{n} X_{1i} \sum_{i=1}^{n} X_{2i}}{n}}{n-1} = \frac{220.972 - \frac{1.512 \times 1.457}{10}}{10-1} = 74,84 \tag{4.4}$$

Uma vez que aqui a covariância foi positiva, este exemplo pode ilustrar a perda em precisão/poder ao analisar os dados ignorando o pareamento (e, consequentemente, a dependência entre os tempos). Ao se ignorar o pareamento, o teste t convencional basicamente trata os dois tempos como sendo dois grupos de pessoas independentes, e a inferência consiste em testar se as médias desses dois grupos podem ser consideradas iguais, ou não. Para realizar esta inferência, pode-se utilizar as estimativas apresentadas na Tabela 3.

Tabela 3: Médias e variâncias amostrais em cada tempo, calculadas em um exemplo de dados pareados simulados.

Tempo 1	Tempo 2
$\bar{X}_1 = 151,2$	$\bar{X}_2 = 145,7$
$s_1^2 = 134,1778$	$s_2^2 = 69,7889$

Para realizar o teste de igualdade entre médias, existem duas possibilidades, conforme se admite que as variâncias dentro de cada grupo (tempo) possam ser consideradas homogêneas ou não (ver, por exemplo, Oliveira *et al.*, 2014). Considerando que a variância de ambos os grupos seja a mesma, esta pode ser estimada a partir de uma média ponderada entre as duas variâncias amostrais, tendo como fator de ponderação os números de graus de liberdade de cada amostra. Como aqui o tamanho das duas amostras é o mesmo (uma vez que na realidade existe um pareamento), esta variância comum é estimada conforme 4.5.

$$s_c^2 = \frac{(10-1)134,1778 + (10-1)69,7889}{10+10-2} = 101,9833 \tag{4.5}$$

Em posse do valor de s_c^2 em 4.5, podemos fazer o teste ignorando pareamento. As

hipóteses em questão podem ser formalizadas como sendo:

$$\begin{cases} H_0: & (\mu_1 - \mu_2) \leq 0 \\ H_1: & (\mu_1 - \mu_2) > 0 \end{cases}$$

sendo μ_1 e μ_2 as médias dos tempos 1 e 2. Neste teste, temos o valor do "t calculado" apresentado em 4.6.

$$t_c = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{\sqrt{s_c^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(151,2 - 145,7) - 0}{\sqrt{101,9833 \left(\frac{1}{10} + \frac{1}{10}\right)}} = 1,218 \tag{4.6}$$

Considerando novamente um nível de significância de $\alpha = 0,05$, tem-se que o valor de t tabelado, com número de graus de liberdade $(n_1 + n_2 - 2) = (10 + 10 - 2) = 18$, é igual a $t_{0,05} = 1,734$. Como $1,218 < 1,734$, *não se rejeita* H_0, ao contrário do aconteceu com o teste t para dados pareados.

Ou seja, ignorando o pareamento, não foi possível identificar que o cardápio é eficiente. Pode-se assim ilustrar que, em geral, dados pareados tendem a ser mais precisos que dados não-pareados. Este comportamento é esperado quase sempre quando temos uma covariância positiva ao longo dos pares. Isto em geral ocorre, apesar do fato de o teste ignorando o pareamento apresentar um número maior de graus de liberdade (o dobro).

Estas duas análises podem ser facilmente realizadas utilizando a linguagem SAS. Antes de mais nada, é preciso definir o conjunto de dados na linguagem, o que pode ser feito (para o exemplo em questão) conforme o código a seguir.

```
data pareados1;
input tempo1 tempo2;
d = tempo1 - tempo2;
cards;
164 157
167 156
143 148
149 145
134 135
150 135
163 151
135 140
156 152
151 138
;
```

O comando `data` cria o conjunto de dados de nome *pareados1*. O comando `input` especifica quais as variáveis a serem lidas, e na linha de baixo é criada uma nova variável, d, dada pela diferença entre os dois tempos. Finalmente, o comando `cards` especifica que os dados são listados em seguida.

Para realizar a análise das diferenças, pode-se utilizar o procedimento PROC REG, construído para o ajuste de modelos de regressão. Para tal, escolhe-se um modelo contendo apenas o intercepto. O teste referente a este parâmetro corresponde ao teste da hipótese de que μ_D seja diferente de zero.

O código para este ajustamento (o qual pode ser especificado logo após a definição do conjunto de dados) é dado por:

```
proc reg;
model d = ;
run;
```

Um recorte da saída de análise da linguagem SAS, para este ajustamento, está apresentado na Figura 1.

The REG Procedure
Model: MODEL1
Dependent Variable: d

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	0	0	.	.	.
Error	9	488.50000	54.27778		
Corrected Total	9	488.50000			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.50000	2.32976	2.36	0.0425

Figura 1: Recorte da saída do programa SAS, referente à análise de dados pareados utilizando diferenças, usando o procedimento PROC REG.

No quadro da ANOVA da Figura 1, o modelo está apresentado com zero graus de liberdade porque se trata de um modelo ajustado para o intercepto. Na mesma Figura, observa-se que o valor da estimativa do intercepto é justamente o valor encontrado anteriormente de $\bar{d}$, ou seja, 5,5. O valor da estatística "t calculado" também foi exatamente o mesmo que aquele em 4.1 (2,36). O valor-p, é importante notar, se refere ao teste bilateral,

que no presente exemplo não faz sentido. Para obter o valor-p do teste unilateral à direita, basta dividir o valor apresentado na saída por 2.

Em seguida, pode-se mostrar que o teste t para dados pareados corresponde a uma análise clássica de Análise de Dados Longitudinais, seja considerando dependência residual, seja considerando um modelo misto, em que os voluntários têm seus efeitos considerados como de natureza aleatória.

Para o ajuste de um modelo com dependência residual, deve-se construir um conjunto de dados no formato comumente chamado, em Análise de Dados Longitudinais, como formato 'longo', em que cada linha corresponde a uma observação. Anteriormente, na definição do conjunto de dados *pareados1*, isto foi feito de maneira diferente, ou seja, utilizando o chamado formato 'largo', em que cada tempo ocupa uma coluna.

Aqui, podemos definir o conjunto em um formato 'longo', conforme a seguir:

```
data pareados2;
input paciente $ tempo $ y;
cards;
1 Antes 164
1 Depois 157
2 Antes 167
2 Depois 156
3 Antes 143
3 Depois 148
4 Antes 149
4 Depois 145
5 Antes 134
5 Depois 135
6 Antes 150
6 Depois 135
7 Antes 163
7 Depois 151
8 Antes 135
8 Depois 140
9 Antes 156
9 Depois 152
10 Antes 151
10 Depois 138
;
```

No código acima, perceba-se que para a definição das variáveis não-numéricas (categóricas), como *paciente* e *tempo*, é preciso alocar um cifrão à frente do nome de tais variáveis, no comando `input`. Em seguida, para o ajustamento do modelo com dependência residual, pode-se utilizar o PROC MIXED, utilizando os comandos a seguir.

```
proc mixed;
class paciente tempo;
model y = tempo;
repeated / type=cs sub=paciente;
run;
```

No código acima, é interessante destacar um aspecto. Na instrução `repeated`, através da opção `type=cs`, está sendo especificado que a estrutura de covariância residual é a de simetria composta, para as observações de um mesmo paciente (`sub=paciente`). Na realidade, a escolha do tipo de estrutura de covariância no presente exemplo é arbitrária, uma vez que se tratam de apenas dois tempos (antes e depois), resultando em um único parâmetro de covariância a ser estimado. Ou seja, poderiam ter sido escolhidas outras estruturas, que resultariam na mesma estimativa desta única covariância sendo estimada.

O resultado deste ajustamento está apresentado na Figura 2.

The Mixed Procedure

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
CS	paciente	74.8444
Residual		27.1389

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
tempo	1	9	5.57	0.0425

Differences of Least Squares Means

Effect	tempo	_tempo	Estimate	Standard Error	DF	t Value	Pr > |t|	Adjustment	Adj P
tempo	Antes	Depois	5.5000	2.3298	9	2.36	0.0425	Tukey-Kramer	0.0425

Figura 2: Recorte da saída do programa SAS, referente à análise de dados pareados utilizando utilizando o comando `repeated`, do procedimento PROC MIXED.

No quadro da ANOVA da Figura 2, é apresentada a estatística de teste F (igual a 5,57), que corresponde ao quadrado da estatística t, apresentada no último quadro (igual a 2,36). Perceba-se que esta estatística t é exatamente igual àquela calculada em 4.1. Além disso, o valor do parâmetro de covariância, apresentado no primeiro quadro da saída na Figura 2, igual a 74,84, foi exatamente igual à covariância amostral calculada anteriormente, em 4.4.

Uma análise equivalente a esta é baseada no ajuste de um modelo misto, tendo o efeito dos pacientes tidos como aleatórios. Com esta natureza aleatória, naturalmente

surge uma covariância entre os tempos Antes e Depois. O ajuste de um tal modelo é feito pelos comandos a seguir.

```
proc mixed;
class paciente tempo;
model y = tempo;
random paciente;
run;
```

Os resultados deste ajustamento estão na Figura 3.

The Mixed Procedure

Covariance Parameter Estimates

Cov Parm	Estimate
paciente	74.8444
Residual	27.1389

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
tempo	1	9	5.57	0.0425

Differences of Least Squares Means

Effect	tempo	_tempo	Estimate	Standard Error	DF	t Value	Pr > |t|	Adjustment	Adj P
tempo	Antes	Depois	5.5000	2.3298	9	2.36	0.0425	Tukey-Kramer	0.0425

Figura 3: Recorte da saída do programa SAS, referente à análise de dados pareados utilizando utilizando o comando `random`, do procedimento PROC MIXED.

No quadro da ANOVA da Figura 3, observa-se que a estatística de teste t apresentou o mesmo valor daquele correspondente na Figura 2 (2,36), bem como do valor calculado para a estatística t para dados pareados, apresentado em 4.1. Além disso, o mesmo valor de covariância observado na Figura 2 foi observado aqui (74,84), por sua vez igual ao valor obtido em 4.4. Há aqui apenas uma diferença conceitual. No presente modelo, esta covariância corresponde ao componente de variância associado à variação aleatória entre pacientes, não podendo ser negativo. Caso esta covariância tivesse sido negativa, este componente de variância acabaria sendo estimado como sendo igual a zero. Isto comprometeria a estimação da variância residual, e alteraria os testes de hipóteses. Na abordagem anterior (dependência nos resíduos), a estimativa poderia ser negativa, e assim seria uma estratégia mais apropriada. Mas, conforme dito anteriormente, em geral espera-se uma covariância positiva, em dados pareados.

Finalmente, pode-se ilustrar a análise ignorando o pareamento utilizando a linguagem SAS, mediante os comandos a seguir:

```
proc mixed;
class tempo;
model y = tempo;
run;
```

Os resultados deste ajustamento estão na Figura 4.

The Mixed Procedure

Covariance Parameter Estimates	
Cov Parm	Estimate
Residual	101.98

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
tempo	1	18	1.48	0.2390

Differences of Least Squares Means									
Effect	tempo	_tempo	Estimate	Standard Error	DF	t Value	Pr > |t|	Adjustment	Adj P
tempo	Antes	Depois	5.5000	4.5163	18	1.22	0.2390	Tukey	0.2390

Figura 4: Recorte da saida do programa SAS, referente à análise de um conjunto de dados pareados simulados, mas ignorando o pareamento.

No quadro da ANOVA da Figura 4, a variância estimada é igual a 101,98, como visto em 4.5, com número de graus de liberdade igual a 18. O valor t também foi igual a 1,218 (na saída, o valor está arredondado para duas casas decimais, 1,22) conforme mencionado anteriormente em 4.6.

4.2 Exemplo 2 (Experimento Longitudinal)

O segundo exemplo se refere a um experimento longitudinal, onde foram observados valores de parasitemia em camundongos ao longo do tempo, expostos a duas dietas: uma comum (denominada aqui como "C") e outra rica em gordura (denominada "H") .O código SAS para a leitura dos dados está ilustrado abaixo (mostrando aqui apenas parte dos dados):

```
data trypanosoma;
input Animal $ Dieta $ Dia $ Parasitemia;
cards;
1 C 9 2
1 C 10 14.2
1 C 11 26.4
  ⋮
7 C 28 72
7 C 29 74
101 H 9 26.4
101 H 10 79.2
  ⋮
110 H 29 348
;
```

Os camundongos da dieta rica em gordura ("H") foram identificados somando-se à sua numeração o valor 100, para diferenciar (por exemplo) o camundongo 1 da dieta "C" do camundongo 1 da dieta "H". Contudo, este cuidado não é rigorosamente necessário, tendo em vista os modelos que foram ajustados, conforme discutido mais adiante.

Inicialmente, foi feita uma representação gráfica dos perfis dos camundongos ao longo do tempo, para cada tipo de dieta, objetivando-se uma visualização dos dados. Os perfis da dieta regular estão apresentados na Figura 5.

Conforme a Figura 5, é possível observar que os camundongos da dieta C apresentaram uma tendência de aumento da parasitemia, ao longo dos dias, com um valor médio oscilando em torno de 100×10^3 parasitas por 0,1 mL de sangue, a partir do vigésimo dia após a infecção, aproximadamente. Já os perfis da dieta rica em gordura estão apresentados na Figura 6.

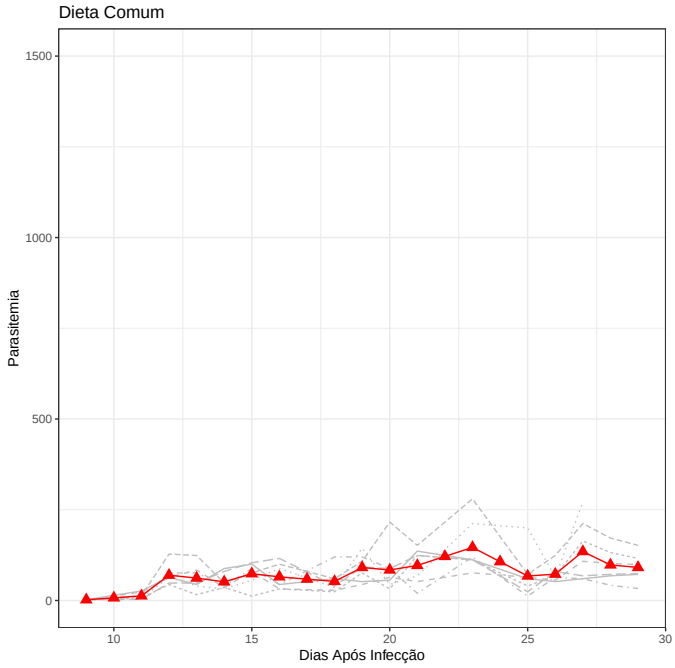

Figura 5: Perfis da parasitemia de 7 camundongos que receberam uma dieta regular, em função dos dias após a infecção. A linha vermelha representa a parasitemia média em cada dia.

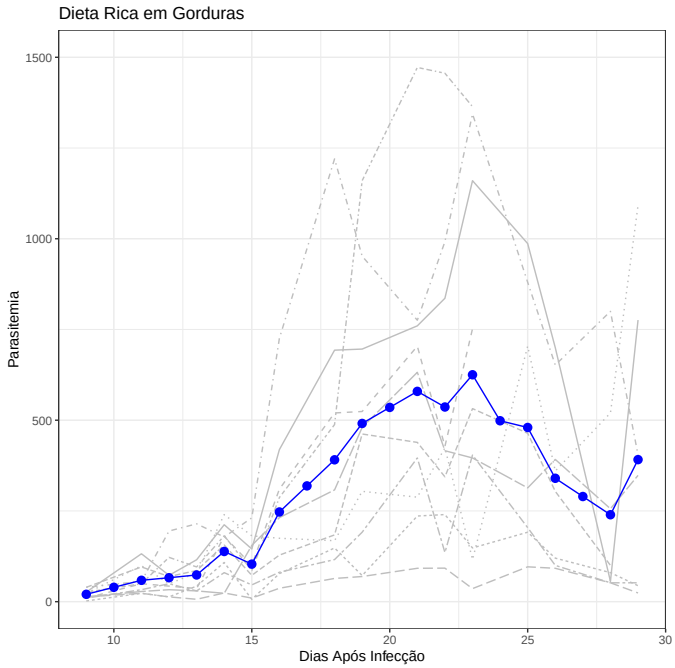

Figura 6: Perfis da parasitemia de 10 camundongos que receberam uma dieta rica em gordura, em função dos dias após a infecção. A linha azul representa a parasitemia média em cada dia.

Na Figura 6, é possível observar que os perfis dos camundongos da dieta H apresentaram níveis de parasitemia bem mais elevados que os da dieta C (perceba-se que alguns perfis, na Figura 6, chegaram a ultrapassar 1000×10^3 parasitas por 0,1 mL de sangue). A parasitemia média para a dieta H (representada por uma curva azul, na Figura 6) ultrapassou 500×10^3 parasitas por 0,1 mL de sangue, após o vigésimo dia. Comparando-se as Figuras 5 e 6, verifica-se que, em coerência com a hipótese de trabalho de Figueiredo *et al.* (2018), os dados sugerem que de fato uma dieta rica em gordura induz a uma maior parasitemia. Também é interessante notar que em ambas as dietas a parasitemia média teve um valor máximo aos 23 dias após a infecção, tendo havido, nos dias seguintes, uma tendência de estabilização, ou mesmo redução.

Os dados deste segundo exemplo foram ajustados a diferentes modelos estatísticos que levassem em conta a dependência existente entre observações de um mesmo camundongo, ao longo do tempo. Da mesma maneira que no primeiro exemplo, foi considerado tanto um modelo misto, admitindo o efeito de animais como de natureza aleatória, bem como modelos com dependência residual, admitindo diferentes estruturas de covariância residual, ao longo do tempo.

Inicialmente, foi considerado um ajustamento com dependência residual, admitindo-se uma estrutura de covariância de simetria composta. O código SAS para este ajustamento está apresentado abaixo:

```
proc mixed;
class Animal Dieta Dia;
model Parasitemia = Dieta Dia Dieta*Dia ;
repeated / type=cs sub=Animal(Dieta);
run;
```

A Figura 7 mostra o recorte da saída do programa SAS, referente à análise com estrutura de covariância de simetria composta.

Conforme comentado anteriormente, a estrutura de covariância de simetria composta impõe que haja uma mesma covariância entre quaisquer pares de tempos. Conforme se vê na Figura 7, esta covariância foi estimada como sendo igual a 24.439 unidades de parasitemia ao quadrado. Além disso, conforme se observa-se pelos Testes F, houve diferenças significativas tanto para os fatores principais (Dieta e Dia), bem como para a interação entre esses fatores. Este último resultado significativo sugere que a comparação entre Dietas deva ser feita separadamente, para cada Dia.

The Mixed Procedure

Covariance Parameter Estimates		
Cov Parm	**Subject**	**Estimate**
CS	Animal(Dieta)	24439
Residual		24825

Fit Statistics	
-2 Res Log Likelihood	4020.0
AIC (Smaller is Better)	4024.0
AICC (Smaller is Better)	4024.1
BIC (Smaller is Better)	4025.7

Type 3 Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Dieta	1	15	9.50	0.0076
Dia	20	285	8.94	<.0001
Dieta*Dia	20	285	5.20	<.0001

Figura 7: Recorte da saida do programa SAS, referente à análise com estrutura de covariância de simetria composta, usando o procedimento MIXED.

O código SAS para o ajustamento considerando um modelo misto, admitindo o efeito de cada camundongo (dentro de cada dieta) como sendo de natureza aleatória, está apresentado abaixo:

```
proc mixed;
class Animal Dieta Dia;
model Parasitemia = Dieta Dia Dieta*Dia ;
random Animal(Dieta);
run;
```

Perceba-se, por este código, que foi ajustado um modelo misto, contendo efeitos fixos e aleatórios. Os fatores de efeito fixo corresponderam a: Dieta, Dia, bem como a interação entre Dieta e Dia. Na linguagem SAS, os efeitos fixos, no PROC MIXED, são especificados sob o comando `model`. O fator de efeito aleatório é especificado sob o comando `random`, e correspondeu ao fator Animal, hierarquizado dentro do fator Dieta. Na linguagem SAS, a hierarquização é especificada, conforme se pode ver acima, utilizando parênteses. O fator dentro dos parênteses (Dieta) representa o fator dentro o qual a hierarquização é considerada.

Como Animal é um fator hierarquizado dentro de cada Dieta, a rigor não há a necessidade de utilizar identificações de camundongo diferentes em cada Dieta, conforme foi feito

aqui na leitura dos dados. Em outras palavras, tanto o primeiro camundongo da Dieta C, como o primeiro camundongo da dieta H, poderiam ter sido identificados como (por exemplo) camundongo "1", que, no ajustamento, a linguagem reconheceria que se tratam de camundongos diferentes.

A Figura 8 mostra o recorte da saída do programa SAS, referente à análise considerando o modelo misto.

The Mixed Procedure

Covariance Parameter Estimates	
Cov Parm	Estimate
Animal(Dieta)	24439
Residual	24825

Fit Statistics	
-2 Res Log Likelihood	4020.0
AIC (Smaller is Better)	4024.0
AICC (Smaller is Better)	4024.1
BIC (Smaller is Better)	4025.7

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Dieta	1	15	9.50	0.0076
Dia	20	285	8.94	<.0001
Dieta*Dia	20	285	5.20	<.0001

Figura 8: Recorte da saida do programa SAS, referente à análise considerando um modelo misto, admitindo o efeito de cada camundongo como sendo de natureza aleatória, usando o procedimento MIXED.

Comparando-se as Figuras 7 e 8, verifica-se que são análises essencialmente iguais. A única diferença, nas saídas, é a identificação do parâmetro estimado de covariância. Na Figura 7 este parâmetro é identificado como "CS" (da sigla inglesa, *compound symmetry*), enquanto que na Figura 8 este parâmetro é identificado como "Animal(Dieta)", que correspondeu ao fator Animal, dentro de cada Dieta. À parte esta identificação, os resultados são os mesmos, com o mesmo valor de covariância estimada (24.439 unidades de parasitemia ao quadrado), e mesmos Testes F.

Ou seja, esta covariância entre tempos pode ser decorrente tanto de uma atribuição do efeito de Animal como aleatório, como de uma especificação de que os resíduos das observações de um mesmo Animal são dependentes, com covariância constante.

Esta análise, considerando qualquer uma dessas duas parametrizações, corresponde,

em Planejamento de Experimentos, à análise de uma estrutura de parcelas subdivididas, utilizando o delineamento inteiramente casualizado (por exemplo, Montgomery, 2012). Na ANOVA de um ensaio em parcelas subdivididas, existem dois erros, sendo o primeiro para testar a hipótese de igualdade entre os níveis do fator de parcela, e o segundo erro é utilizado para testar o fator de subparcela, bem como a interação entre o fator de parcela e aquele de subparcela. Em um delineamento inteiramente casualizado, o primeiro erro corresponde à variação de repetições dentro (ou seja, hierarquizado) do fator de parcela. No presente exemplo, esta variação corresponde à variação entre animais dentro de cada dieta. Por isso é que nos testes F de ambas as Figuras 7 e 8, referente ao teste para o fator Dieta, apresenta 15 graus de liberdade no denominador. Como foram 7 camundongos submetidos à Dieta C, e 10 camundongos para a Dieta H, tem-se um total de $(7-1)+(10-1)=15$ graus de liberdade, como em um ensaio em parcelas subdividas desbalanceado.

Em um estudo longitudinal, a grande desvantagem dos dois enfoques anteriores é a de que admitem que a covariância entre quaisquer pares de tempos é constante. No entanto, é frequentemente razoável admitir que a covariância tenda a diminuir, à medida que os dois tempos em questão vão ficando mais afastados. No primeiro exemplo, esta consideração não precisou ser levada em conta, uma vez que eram apenas dois tempos considerados (antes e depois). No segundo exemplo, contudo, como o número de tempos é bem maior, faz sentido avaliar outras estruturas de covariância residual, como a autorregressiva de ordem 1. O código SAS para o ajuste de uma tal estrutura é dado como:

```
proc mixed;
class Animal Dieta Dia;
model Parasitemia = Dieta Dia Dieta*Dia ;
repeated / type=ar(1) sub=Animal(Dieta);
run;
```

A Figura 9 mostra o recorte da saída do programa SAS, referente à análise referente à análise com estrutura de covariância autorregressiva de ordem 1.

Pode-se observar na Figura 9 que o parâmetro de autocorrelação foi relativamente elevado (0,8755), o qual corresponde à correlação residual entre dois tempos espaçados por um dia. A correlação entre dois tempos espaçados por dois dias é dada como sendo $0,8755^2 = 0,7665$, entre dois tempos espaçados de três dias $0,8755^3 = 0,6711$, e assim por diante.

Esta outra estrutura de covariância altera os valores-p dos teste F (Figura 9), embora as conclusões dos testes não tenham se alterado, em relação àqueles das Figuras 7 e 8.

The Mixed Procedure

Covariance Parameter Estimates		
Cov Parm	**Subject**	**Estimate**
AR(1)	Animal(Dieta)	0.8755
Residual		50860

Fit Statistics	
-2 Res Log Likelihood	3775.2
AIC (Smaller is Better)	3779.2
AICC (Smaller is Better)	3779.2
BIC (Smaller is Better)	3780.9

Type 3 Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Dieta	1	15	10.58	0.0054
Dia	20	285	2.37	0.0010
Dieta*Dia	20	285	2.03	0.0063

Figura 9: Recorte da saida do programa SAS, referente à análise com estrutura de covariância autorregressiva de ordem 1, usando o procedimento MIXED.

Para verificar se a estrutura autorregressiva promoveu um melhor ajustamento, em relação à simetria composta, pode-se utilizar o Critério de Informação de Akaike (AIC). O AIC é uma medida que é utilizada para avaliar a qualidade de um modelo estatístico, levando em consideração tanto a capacidade de ajuste do modelo aos dados quanto a sua complexidade. Aqui, quanto menor o valor do AIC, melhor o ajustamento. Os critérios de informação de Akaike (AIC) para os modelos destas duas estruturas de covariância, estão apresentados na Tabela 4.

Tabela 4: Desempenho de dois modelos estatísticos, conforme o Critério de Informação de Akaike, admitindo duas estruturas de covariância residual, ajustados aos dados do Exemplo 2.

Modelo (conforme a estrutura de covariância)	AIC
Simetria Composta	4024,0
Autorregressiva de Ordem 1	3779,2

Na Tabela 4, pode-se observar que o menor valor de AIC ocorreu com o modelo autorregressivo de ordem 1 (3779,2), em relação ao da estrutura de simetria composta (4024,0). Com isso, pôde-se verificar que, de fato, foi mais adequado considerar que as covariâncias entre pares de tempos de um mesmo camundongo tendam a diminuir, à medida que os pares de pontos vão ficando mais afastados.

Desta maneira, é razoável basear a inferência no modelo autorregressivo de ordem 1. Como a interação entre os fatores Dieta e Dia foi significativa, é interessante realizar o seu desdobramento, ou seja, comparar as duas Dietas em cada Dia, mediante um teste F. Aqui não há necessidade de realizar um teste *post hoc* de comparações múltiplas (como o Teste de Tukey), uma vez que são apenas duas dietas, e assim o próprio teste F as diferencia, ou não. O código SAS, incluindo a solicitação de desdobramento da interação, está apresentado abaixo.

```
proc mixed;
class Animal Dieta Dia;
model Parasitemia = Dieta Dia Dieta*Dia ;
repeated / type=ar(1) sub=Animal(Dieta);
lsmeans Dieta*Dia / adjust=tukey slice=Dia; ;
run;
```

O desdobramento é feito pelo comando `lsmeans`, tendo como opção `slice=Dia`. Esta opção é a que faz com que o desdobramento seja feito comparando as duas dietas em cada dia. A opção `adjust=tukey` foi utilizada meramente para que se obtivessem as médias de quadrados mínimos de cada dieta, em cada dia. Em conjuntos de dados desbalanceados, as médias de quadrados mínimos (obtidas em função de um dado modelo estatístico) não necessariamente correspondem a médias aritméticas.

O desdobramento da interação Dieta $\times$ Dia está apresentado na Figura 10.

Observa-se, na Figura 10, que as duas dietas apresentaram níveis de parasitemia estatisticamente diferentes a partir do dia 17, embora, considerando um nível de significância de 5%, a diferença entre as duas dietas não tenha sido estatisticamente significativa nos dias 27 e 28. Talvez esta não-significância tenha ocorrido apenas por uma falta de poder estatístico, uma vez que, mesmo nestes dias 27 e 28, a parasitemia para a dieta rica em gordura manteve-se superior. Isto pode ser observado na Figura 11, que mostra as curvas médias de parasitemia para ambas as dietas. Nesta mesma Figura são destacados, através de linhas pontilhadas, aqueles dias em que as duas dietas foram significativamente diferentes.

Finalmente, a última estrutura de covariância que poderia ser considerada seria aquela designada como "não-estruturada", em que uma covariância diferente é estimada para cada intervalo diferente entre dois tempos, em relação ao número de dias. No presente exemplo, como o intervalo de tempo variou desde o dia 9 até o dia 29 (21 tempos), haveria um total de $\frac{21(21+1)}{2} = 231$ parâmetros de covariância a serem estimados, lembrando que,

Tests of Effect Slices					
Effect	Dia	Num DF	Den DF	F Value	Pr > F
Dieta*Dia	9	1	285	0.03	0.8691
Dieta*Dia	10	1	285	0.08	0.7710
Dieta*Dia	11	1	285	0.17	0.6781
Dieta*Dia	12	1	285	0.00	0.9734
Dieta*Dia	13	1	285	0.01	0.9150
Dieta*Dia	14	1	285	0.61	0.4352
Dieta*Dia	15	1	285	0.07	0.7894
Dieta*Dia	16	1	285	2.68	0.1026
Dieta*Dia	17	1	285	5.48	0.0199
Dieta*Dia	18	1	285	9.27	0.0026
Dieta*Dia	19	1	285	12.97	0.0004
Dieta*Dia	20	1	285	16.49	<.0001
Dieta*Dia	21	1	285	18.88	<.0001
Dieta*Dia	22	1	285	13.94	0.0002
Dieta*Dia	23	1	285	18.57	<.0001
Dieta*Dia	24	1	285	18.70	<.0001
Dieta*Dia	25	1	285	19.07	<.0001
Dieta*Dia	26	1	285	8.86	0.0032
Dieta*Dia	27	1	285	3.62	0.0581
Dieta*Dia	28	1	285	2.31	0.1294
Dieta*Dia	29	1	285	7.01	0.0085

Figura 10: Desdobramento da interação Dieta × Dia, comparando as duas dietas em cada dia

nesta estrutura geral, também são consideradas variâncias heterogêneas, estimando-se uma variância para cada tempo.

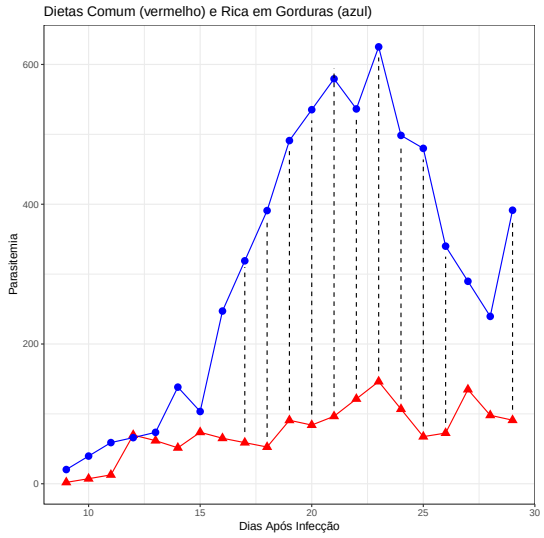

Figura 11: Médias de parasitemia em cada dia, para ambas as dietas. Em vermelho, tem-se as médias para a dieta comum, e em azul as médias para a dieta rica em gordura. As linhas pontilhadas representam diferenças que foram estatisticamente diferentes de zero, com $\alpha = 0,05$.

Talvez devido a este número excessivo de parâmetros, não houve convergência em relação ao ajustamento utilizando a categoria não-estruturada de dependência residual, e assim o resultado desse ajustamento não é apresentado aqui.

De qualquer maneira, este exemplo ilustra como o ajustamento de dados de um experimento longitudinal pode ser feito de maneira simples com a linguagem SAS, levando em conta que a dependência residual entre dois tempos vai se reduzindo, à medida que tais tempos vão ficando mais espaçados, através do uso da estrutura autorregressiva de ordem 1.

5 Considerações Finais

A necessidade da análise de dados longitudinais é muito frequente em variadas áreas do conhecimento. Neste trabalho, procurou-se ilustrar, discutir, e apresentar aspectos práticos e teóricos no ajuste de modelos apropriados para dados longitudinais, considerando dois exemplos elementares que ilustram este tipo de dado.

Em particular, optou-se por explorar os recursos da linguagem SAS, nem sempre abordada em cursos de graduação em Estatística. Esta linguagem oferece versatilidade e simplicidade no ajuste de modelos longitudinais, sendo uma ferramente adicional da qual o estatístico pode dispor.

Como comentários finais, pode-se citar dois exemplos desta simplicidade de codificação com a linguagem. O uso de parênteses e asteriscos, para designar se um fator está hierarquizado ou cruzado com algum outro fator, oferece uma sintaxe muito simples e natural de especificação de modelos estatísticos.

Além disso, é interessante destacar, no `PROC MIXED` (embora isso não tenha sido feito no presente trabalho) é possível utilizar as instruções `random` e `repeated` ao mesmo tempo. Ou seja, ajustar um modelo que ao mesmo tempo seja misto, e que também apresenta dependência residual, recurso esse que nem sempre está prontamente disponível em outras linguagens. Esta é uma situação que pode surgir quando se considera um número maior de fatores, sendo alguns de natureza aleatória.

Em suma, nossa expectativa é a de que este material possa servir de consulta a iniciantes em análise de dados longitudinais, eventualmente interessados na linguagem SAS.

6 Referências Bibliográficas

DEMPSTER, A.P.; LAIRD, N.M.; RUBIN, D.B. Maximum likelihood from incomplete data via the EM algorithm. **Ser. B**. n.39, p.1-8, 1977.

FIGUEIREDO, V.P.; LOPES Jr, E.S.; LOPES, L.R.; SIMÕES, N.F.; PENITENTE, A.R.; BEARZOTI, E.; VIEIRA, P.M.A.V.; SHULZ, R.; TALVANI, A. High fat diet modulates inflammatory parameters in the heart and liver during acute *Trypanosoma cruzi* infection. **International Immunopharmacology**. n.64, p.192-200, 2018.

FOULLEY, J.L.; JAFFRÉZIC, F.; ROBERT-GRANIÉ, C. EM-REML estimation of covariance parameters in Gaussian mixed models for longitudinal data analysis. **Genet. Sel. Evol.**. n.32, p.129-141, 2000.

LINDSTROM, M.J.; BATES, D.M. Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. **Journal of the American Statistical Association**. n.83, p.1014-1022, 1988.

LITTEL, R.C.; MILLIKEN, G.A.; STROUP, W.W.; WOLFINGER, R.D. **SAS® System for Mixed Models**. Cary, NC: SAS Institute Inc., 1996.

MONTGOMERY, D.C. **Design and Analysis of Experiments**. 8.ed. Wiley, 2012.

OLIVEIRA, M.S.; BEARZOTI, E.; BOAS, F.L.V.; NOGUEIRA, D.A.; NICOLAU, L.A.; OLIVEIRA, H.S.S.O. **Introdução à Estatística**. Lavras, Ed. UFLA, 2014. 461p.

SAS INSTITUTE Inc. **SAS® OnDemand for Academics: User's Guide**. Cary, NC: SAS Institute Inc., 2015.

SINGER, J.M.; NOBRE, J.S.; ROCHA, F.M.M. **Análise de Dados Longitudinais**: versão parcial preliminar. São Paulo, 2018. Disponível em: `https://www.ime.usp.br/~jmsinger/MAE0610/Singer&Nobre&Rocha2018jun.pdf`. Acesso em: nov 2023.

SIQUEIRA, A.L.; TIBÚRCIO, J.D. **Estatística na Área da Saúde**: conceitos, metodologia, aplicações e prática computacional. Belo Horizonte, Coopmed, 2011. 538p.

WOLFINGER, R.D.; TOBIAS, R.D.; SALL, J. Computing Gaussian likelihoods and their derivatives for general linear mixed models. **SIAM Journal on Scientific Computing**. n.15, v.6, p.1294-1310, 1994.