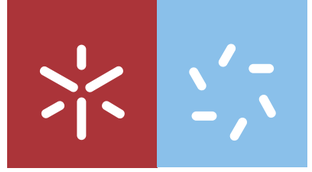
Uma Abordagem Multivariada para Modelos Conjuntos de Dados Longitudinais e de Sobrevivência

Pedro Afonso

UMinho | 2019

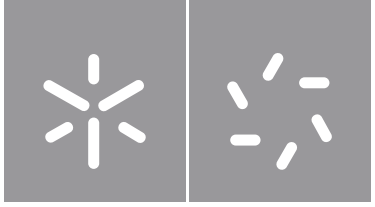


Universidade do Minho
Escola de Ciências

Pedro Manuel Miranda Afonso

Uma Abordagem Multivariada para Modelos Conjuntos de Dados Longitudinais e de Sobrevivência

outubro de 2019



Universidade do Minho
Escola de Ciências

Pedro Manuel Miranda Afonso

Uma Abordam Multivariada para Modelos Conjuntos de Dados Longitudinais e de Sobrevivência

Dissertação de Mestrado
Mestrado em Estatística

Trabalho efetuado sob a orientação da
Professora Doutora Inês Pereira Silva Cunha Sousa

outubro de 2019

Direitos de autor e condições de utilização do trabalho por terceiros

Agradecimentos

Na nossa vida adulta somos tentados a pensar que o nosso sucesso é o fruto das nossas capacidades e do nosso esforço individual. Não é bem assim. Somos, como diria Ortega y Gasset, uma mistura sem regra do Eu e da Circunstância[1]. Esta verdade determina em larga medida o nosso sucesso individual. As próximas palavras são para algumas das pessoas que ao longo dos últimos meses preencheram a minha Circunstância e, por isso, tornaram este trabalho possível (e mais agradável).

As pessoas foram muitas, mas como diria Alceste de Molière[2]: elogiar toda a gente é elogiar ninguém. Gostava de agradecer, sobretudo, à minha orientadora Doutora Inês Sousa. Estou-lhe grato por ter despertado em mim o interesse pela modelação de dados longitudinais, e por me ter acolhido no seu projeto de investigação. Um bom orientador é aquele que nos ensina o método, nos encoraja a seguir as nossas ideias e, o mais importante, nos critica os resultados. Tive a felicidade de encontrar exatamente isso. Se a 'Professora Inês' me ofereceu conhecimento na dúvida e rumo na confusão, foi na família e nos amigos que encontrei apoio incondicional. Foram eles que me consolaram no desastre, que com vincada convicção me persuadiram que tudo acabaria bem. Obrigado aos meus pais e ao meu irmão. Eles, melhor do que ninguém, sabem o quanto de mim e do nosso tempo coloquei neste trabalho. Obrigado ao Luís Ferrás pela companhia e boa disposição com que preencheu o nosso gabinete.

Este trabalho teve o apoio do projeto 028248/SAICT/2017 financiado pelo Programa Operacional Competitividade e Internacionalização (COMPETE2020) na sua componente de Fundo Europeu de Desenvolvimento Regional (FEDER) e pela Fundação para a Ciência e a Tecnologia, I.P. (FCT, I.P.) na sua componente OE.

[1] Ensaio "Meditaciones del Quijote"(1914) de Ortega y Gasset.

[2] Alceste é o protagonista da peça "O Misantropo"(1666) de Molière.

Uma abordagem multivariada para modelos conjuntos de dados longitudinais e de sobrevivência

Resumo: Um grande desafio da análise de dados longitudinais é a presença de observações omissas por abandono de alguns dos participantes. Se o motivo do abandono está relacionado com a resposta longitudinal em análise (e.g., o paciente morre da doença em estudo), os dados observados podem não representar uma amostra aleatória dos dados completos. Esta perda de informação conduz a uma redução da precisão, e se não for tratada adequadamente, pode conduzir a inferências enviesadas e conclusões imprecisas. Por este motivo, quando o abandono ocorre de forma não aleatória, o processo de omissão não pode ser ignorado e deve ser considerado na análise. Neste contexto, a modelação conjunta de dados longitudinais e de tempo-até-evento surge como uma solução.

Neste trabalho foi desenvolvida uma função para o software R, que permite ao utilizador usar uma base de dados completa para gerar novas bases de dados com dados omissos, enquanto controla o mecanismo de omissão e proporção global de abandono dos participantes. Esta função designa-se por `trim()` e será incluída num *package* já existente no software R. Depois, fazendo uso da função desenvolvida e de conjuntos de dados completos simulados, é apresentado um estudo para avaliar de que forma as caraterísticas do conjunto de dados longitudinais e caraterísticas dos dados omissos influenciam inferências baseadas exclusivamente nos dados observados. Mais ainda, este trabalho estende o modelo Gaussiano transformado proposto por Diggle et al. (2008) [1] para descrever a distribuição conjunta das respostas longitudinal e tempo-até-abandono. A contribuição nesta dissertação foi a proposta de novas estruturas de correlação para este modelo, com uma interpretação mais intuitiva, através da inclusão de um parâmetro de associação entre as duas respostas. O algoritmo EM foi aplicado para obter as estimativas de máxima verosimilhança dos parâmetros do modelo conjunto, em alternativa à diferenciação da log-verosimilhança seguida no trabalho inicial. Esta abordagem permitiu-nos obter, na presença de tempo-até-abandono censurados, pela primeira vez, expressões com forma fechada para alguns dos parâmetros do modelo.

Palavras-chave: dados longitudinais, dados omissos, abandono informativo, modelo conjunto paramétrico.

A mutivariate approach to joint modelling of longitudinal and survival data

Abstract: A major challenge in the analysis of longitudinal data is the presence of missing data due to participants dropping out. If the reasons for dropping out are related to the outcome measure (e.g., a patient dies from the disease under study), the observed data may not resemble a random sample of the complete data. This loss of information leads to a reduction in accuracy, and, if not handled properly, the observed data may lead to biased inferences and inaccurate conclusions. Hence, when drop-outs occur non-randomly, the loss of this data cannot be ignored and must be taken into account in the analysis. Within this context, the joint modelling of longitudinal data and time-to-event data comes into play.

In this work, a function is developed for the software R that allows the user to use a complete dataset to generate new datasets with missing observations while controlling the missing mechanism and the overall subject dropout proportion. This function, `trim()`, will be added to an existing R package. We conducted a simulation study, using the developed function and simulated complete datasets, to investigate how the characteristics of both the longitudinal dataset and missing observations influence inferences based solely on the observed data.

Further work builds upon the multivariate Gaussian model proposed by Diggle et al. (2008) [1] to describe the joint distribution of longitudinal and missing processes. The contribution of this dissertation is the proposal of a new correlation structure for this model with a more intuitive interpretation based on an association parameter between two responses. The EM algorithm is used to derive the maximum likelihood estimates of the joint model in lieu of the differentiation of the log-likelihood method used in the initial work. This approach makes possible, for the first time, closed-form expressions for some of the model parameters when censored times are observed.

Keywords: longitudinal data, missing data, informative dropout, parametric joint model.

Declaração de integridade

Eu, Pedro Manuel Miranda Afonso, n.º PG35775, aluno do Mestrado em Estatística na Escola de Ciências da Universidade do Minho, declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducentes à sua elaboração.

Mais declaro que conheço e respeitei o Código de Conduta Ética da Universidade do Minho.

Universidade do Minho, 29 de outubro de 2019

Índice

Abreviaturas

MAR	Omissão aleatória (*Missing At Random*)
MCAR	Omissão completamente aleatória (*Missing Completely at Random*)
MNAR	Omissão não aleatória (*Missing Not At Random*)
EM	Estimação-Maximização (*Expectation Maximization*)
EPM	Erro Percentual Médio
f.d.p.	Função densidade de probabilidade
TGM	Modelo Gaussiano transformado (*Transformed Gaussian Model*)
v.a.	Variável aleatória

Figuras

Tabelas

Capítulo 1

Introdução

Dados longitudinais são gerados quando os mesmos indivíduos são observados repetidamente ao longo do tempo relativamente a uma mesma caraterística. Os modelos longitudinais descrevem o processo estocástico dos dado observados. Estes modelos são uma importante ferramenta de investigação porque nos permitem, sobretudo, distinguir a variabilidade dos dados dentro e entre indivíduos [2], e relacionar essa variabilidade com o tempo e outros fatores. Uma vez que a recolha de dados longitudinais é dilatada no tempo, a presença de dados omissos em estudos longitudinais é muito frequente. A modelação correta de dados longitudinais na presença de dados omissos continua a ser um dos maiores desafios da análise deste tipo de dados. Se a omissão ocorre de forma não aleatória, os dados observados não constituem uma amostra aleatória dos dados completos. Esta perda de informação conduz não só a uma redução da precisão mas também, se não for tratada adequadamente, a inferências enviesadas e conclusões imprecisas. É neste contexto que a modelação conjunta de dados longitudinais e de tempo-até-evento surge como uma solução para este problema, modelando conjuntamente o processo estocástico longitudinal e o processo estocástico de omissão.

Este trabalho teve como objetivo desenvolver uma função em software R que permita ao utilizador simular dados com observações omissas a partir de dados longitudinais completos, e com isso conduzir estudos de simulação independentes para avaliar de que forma diferentes caraterísticas dos dados longitudinais e dos dados omissos podem afetar inferências baseadas exclusivamente nos dados observados. Pretendeu-se ainda estender o modelo Gaussiano transformado, desenvolvido por Diggle et al. (2008) [1], para incluir novas estruturas de cor-

relação com uma interpretação mais intuitiva e obter as equações dos estimadores dos seus parâmetros.

Este documento é composto por 5 capítulos. Nas próximas secções do Capítulo 1 é introduzida a terminologia e notação utilizado ao longo do documento. Nas secções seguintes dão apresentados de forma breve conceitos fundamentais da análise de dados longitudinais e de sobrevivência. Essas secções não têm a pretensão de descrever exaustivamente os métodos desenvolvidos em ambos as áreas do conhecimento, apenas o objetivo de introduzir alguns conceitos essenciais à compreensão do trabalho exposto nos restantes capítulos. No Capítulo 2 descreve-se a classificação de dados omissos em estudos longitudinais. Ainda no mesmo capítulo descreve-se o desenvolvimento da função `trim()` no software R, que permite ao utilizador simular dados omissos por abandono em conjuntos de dados longitudinais completos. Depois, é apresentado um estudo de simulação que faz uso da função desenvolvida para avaliar de que forma as caraterísticas dos dados longitudinais e dos dados omissos influenciam inferências baseadas exclusivamente nos dados observados. No Capítulo 3 são apresentadas as principais abordagens descritas na literatura para a modelação conjunta de dados longitudinais e de dados tempo-até-evento, e as suas principais motivações. No Capítulo 4 aborda-se o modelo Gaussiano transformado proposto por Diggle et al. (2008) [1]. São apresentadas novas estruturas de correlação para este modelo conjunto, com uma interpretação mais intuitiva. Apresentamos as equações dos estimadores dos parâmetros para uma estrutura particular. No capítulo 5 é feita uma reflexão crítica sobre o trabalho desenvolvido e apresentam-se direções para o trabalho futuro. Este documento é ainda constituído pelos anexos A, B, e C. No Anexo A é disponibilizado o código integral utilizado no software R para o desenvolvimento da função `trim()` e o estudo simulação descritos no Capítulo 2. Todos os resultados apresentados fizeram uso de uma semente de aleatoriedade disponibilizada ao leitor no código utilizado, permitindo a reprodutibilidade dos mesmos no seu computador. Não sendo possível disponibilizar dos conjuntos de dados em anexo, estes podem ser consultados em `https://bit.ly/2Oe17ga`. No Anexo B são apresentados gráficos completares relativos ao estudo de simulação, mas que são referenciados ao longo do texto do Capítulo 2. No Anexo C são compilados fundamentos de teoria multivariada Gaussiana, resultados de operações sobre matrizes, e a fórmula de Leibniz. Estes conceitos são fundamentais para a

compreensão dos resultados apresentados no Capítulo 4, e referenciados no corpo do texto, sempre que oportuno.

1.1 Terminologia e notação

Neste documento letras maiúsculas são utilizadas para descrever variáveis aleatórias (v.a.), e letras minúsculas para representar as respetivas realizações. Distinguem-se matrizes/vetores de escalares apresentando-se os primeiros a negrito e os segundos a estilo normal. No entanto, vetores são representados com letras minúsculas, e matrizes com letras maiúsculas. Pontualmente apresenta-se em subscrito as dimensões dos vetores e/ou matrizes a fim facilitar a compreensão do leitor. Neste documento, a matriz transposta de $\boldsymbol{M}$ é representada por $\boldsymbol{M}^\top$. O elemento na linha i e na coluna j da matriz $\boldsymbol{X}$ representa-se por x_{ij}. O traço de uma matriz representa-se por $\text{Tr}[\cdot]$. O valor esperado e a variância de uma variável aleatória (v.a.) são representados por $E[\cdot]$ e $Var[\cdot]$, respetivamente. A covariância entre duas v.a. é representada por $Cov(\cdot,\cdot)$. O símbolo $\perp\!\!\!\perp$ denota que duas v.a. são independentes entre si.

Alguns exemplos:

1. x_i é uma realização da variável aleatória univariada X_i, $X_i \sim N(\mu, \sigma^2)$;
2. $\boldsymbol{x}_i$ é uma realização da variável aleatória multivariada $\boldsymbol{X}_i$, $\boldsymbol{X} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$;
3. Se $\boldsymbol{A}$ for uma matriz simétrica verifica-se a relação $\boldsymbol{A}^\top = \boldsymbol{A}$;
4. $\underset{r\times c}{\boldsymbol{B}}$ é uma matriz com r linhas e c colunas;
5. $x_{12} = 7$ é o elemento da matriz $\boldsymbol{X}$ na linha 1 e na coluna 2, atendendo que $\boldsymbol{X} = \begin{bmatrix} 6 & 7 \\ 8 & 9 \end{bmatrix}$;
6. Sejam $\boldsymbol{a} = (a_1, \cdots, a_n)^\top$ e $\boldsymbol{b} = (b_1, \cdots, b_n)^\top$ dois vetores de dimensão n, $\boldsymbol{a}^\top \cdot \boldsymbol{b} = \sum_{i=1}^{n} a_i \cdot b_i$.
7. $Var[X + Y] = Var[X] + Var[Y] = ((E[X^2] + E[X]^2) + Var[Y]$, se $X \perp\!\!\!\perp Y$.

Uma descrição mais específica da notação utilizada será apresentada caso a caso ao longo do documento, sempre que pertinente.

1.2 Modelos para análise de dados longitudinais

Dados longitudinais são dados que descrevem a evolução no tempo de uma mesma característica no mesmo conjunto de unidades amostrais (e.g., indivíduos, animais, amostras de laboratório). Os dados podem ser recolhidos de forma prospetiva, i.e., as unidades amostrais são observadas repetidas vezes ao longo do tempo relativamente a uma mesma caraterística. Alternativamente, a informação pode ser obtida retrospetivamente, onde informações que descrevem os indivíduos são recolhidas do seu passado [2]. Nas ciências sociais este tipo de dados são habitualmente designados por dados de painel.

Neste documento designar-se-á as unidades amostradas por indivíduos. A observação registada no indivíduo i no tempo j é designada por y_{ij}. Cada indivíduo é observado em diferentes ocasiões t_j. As n_i observações recolhidas no indivíduo i são representadas pelo vetor $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{in_i})^\top$. O vetor $\boldsymbol{y}_i$ é uma realização da v.a. multivariada $\boldsymbol{Y}_i$. O i-ésimo indivíduo em cada instante j tem a si associado, a par da resposta y_{ij}, um vetor de covariáveis $\boldsymbol{x}_{ij}$ com dimensão p. O valor das variáveis pode variar no tempo. Os vetores de covariáveis do mesmo indivíduo podem ser organizados na forma matricial,

$$\underset{n_i \times p}{\boldsymbol{X}_i} = \begin{bmatrix} \boldsymbol{x}_{i1}^\top \\ \vdots \\ \boldsymbol{x}_{it_{n_i}}^\top \end{bmatrix},$$

conhecida por matriz desenho. Num estudo longitudinal em que participam m indivíduos observados em n_i ocasiões, o número total de observações é representado por N, com $N = \sum_i^m n_i$. O conjunto das variáveis respostas para todos os m indivíduos a participar no estudo é dado pelo vetor $\boldsymbol{Y} = (\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_m)^\top$ de dimensão N.

Num estudo longitudinal, a variável resposta pode ser contínua ou discreta. Nos casos em que a variável é continua, e.g., pressão sanguínea, habitualmente aplicam-se modelos assentes no pressuposto de que a variável resposta segue uma distribuição Gaussiana. Por vezes é necessário aplicar uma transformação Box-Cox. No caso discreto, a variável pode ser descrever, por exemplo, a presença ou ausência de um sintoma ao longo do tempo (variável resposta binária) [3]. Ou ainda descrever um processo de contagem [4, 5], e.g., o número de vezes que o sintoma se manifesta. Nesta introdução abordaremos apenas a modelação de

dados longitudinais Gaussianos.

Análises simplistas de dados longitudinais passam por resumir a sequência de medidas de cada indivíduo num único valor sumário. Algumas destas medidas sumário são, por exemplo, a área sob a curva, ou o declive ao longo do tempo. Embora apelativas pela sua simplicidade estas abordagens implicam uma perda de informação, uma vez que diferentes perfis individuais podem produzir a mesma medida sumário. Contudo, este tipo de análises mantém-se em uso para análises exploratórias dos dados [6].

Estudos longitudinais constituem uma importante ferramenta de investigação uma vez que fornecem conhecimento sobre a evolução e persistência da caraterística de interesse no tempo e dos fatores que contribuem para esse desenvolvimento. Ao contrário do que acontece num estudo transversal, a existência de medições repetidas ao longo do tempo sobre o mesmo indivíduo permite separar o efeito do coorte do efeito da idade [2]. Cada indivíduo funciona como o seu próprio controlo, sendo as alterações na caraterística de interesse observadas em cada indivíduo ao longo do tempo estimadas independentemente de qualquer variação da caraterística entre indivíduos. O desenho do estudo permite, por isso, captar as alterações dentro de cada indivíduo; e ao separar os erros nas medições dos erros aleatórios aumentar a potência estatística, e reduzir o enviesamento [7]. Assim, os estudos longitudinais possibilitam uma compreensão mais profunda sobre as relações causa-efeito das variáveis observadas, do que seria possível por meio de um estudo transversal.

Estudos longitudinais distinguem-se da análise de séries temporais pelo número de unidades amostrais observadas. No estudo de séries temporais, o interesse recai na observação de uma unidade amostral durante um período longo; enquanto que no segundo várias unidades amostrais são acompanhadas ao longo de um período de tempo, habitualmente mais curto.

Num estudo longitudinal, os indivíduos não necessitam, necessariamente, de ser observados nos mesmos instantes (espaçamento irregular entre observações) e/ou o mesmo número de vezes. Diz-se que estamos na presença de um estudo longitudinal *balanceado* (ou *equilibrado*) quando os indivíduos possuem o mesmo número de observações e se estas foram feitas nas mesmas ocasiões; e *não balanceado* (ou *não equilibrado*) caso contrário. Laird & Ware (1982) [8] propuseram uma classe flexível de modelos para dados longitudinais, os modelos lineares mistos, capazes de lidar com dados longitudinais não equilibrados. No entanto, nem

todos os métodos são capazes de o fazer, como é o caso da ANOVA univariada de medições repetidas [6].

Uma vez que os dados longitudinais constituem medições repetidas no tempo sobre os mesmos indivíduos, pelo que existe provavelmente uma correlação serial entre as observações de um mesmo indivíduo. Esta caraterística implica que uma regressão linear simples por assumir que as observações são independentes não é apropriada para analisar dados longitudinais. Por este motivo a estrutura de correlação desempenha um papel preponderante na estimação dos parâmetros do modelo a ajustar aos dados [2]. Modelos para a análise de dados longitudinais devem não ter apenas em conta a dependência entre as covariáveis e a variável resposta, mas também a dependência entre as respostas do mesmo indivíduo. Contudo, respostas de diferentes indivíduos são consideradas independentes.

Diggle et al. (2002) [2] divide os modelos longitudinais em três categorias:

Modelo de transição: O modelo de transição modela conjuntamente o valor esperado e a dependência temporal, condicionando a resposta a um subconjunto particular de outras respostas. Nestes modelos a correlação entre as respostas de um mesmo indivíduo é explicada pelo facto do valor de Y_{ij} depender dos valores anteriormente observados $(Y_{i1}, \cdots, Y_{i(j-1)})^\top$, i.e., as respostas anteriores são tratadas como covariáveis.

Modelo marginal: Com um modelo marginal pretende-se fazer inferência sobre o valor médio populacional. O valor esperado marginal é modelado em função de covariáveis mas não é condicionado a outras variáveis resposta ou a efeitos aleatórios, na forma

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i, \quad \boldsymbol{Z}_i \sim MVN(\boldsymbol{0}, \boldsymbol{V}_i), \tag{1.1}$$

com $i = 1, \ldots, m$, onde $\boldsymbol{Y}_i$ é o vetor das variáveis respostas para o i-ésimo indivíduo, $\boldsymbol{\beta}$ é o vetor dos efeitos fixos (parâmetros desconhecidos), e $\boldsymbol{X}_i$ é a matriz desenho dos efeitos fixos. A matriz $\boldsymbol{V}_i$ é a matriz de co(variâncias) que descreve a estrutura de correlação entre as observações do mesmo indivíduo. Exemplos de estruturas utilizadas para modelar $\boldsymbol{V}_i$ são: não estruturada, simetria composta, auto-regressiva, correlação espacial Exponencial ou Gaussiana. Algumas destas estruturas são apresentadas na Tabela 1.1.

Tabela 1.1: Possíveis estruturas para a matriz de co(variâncias) $\boldsymbol{V}_i$ no modelo (1.1).

Estruturas de correlação		
Não estruturada	Simetria composta	Auto-regressiva (ordem 1)
$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ & \sigma_2^2 & \sigma_{21} & \sigma_{23} & \sigma_{24} \\ & & \sigma_3^2 & \sigma_{31} & \sigma_{32} \\ & & & \sigma_4^2 & \sigma_{41} \\ & & & & \sigma_5^2 \end{bmatrix}$	$\sigma^2 \cdot \begin{bmatrix} 1 & \rho & \rho & \rho & \rho \\ & 1 & \rho & \rho & \rho \\ & & 1 & \rho & \rho \\ & & & 1 & \rho \\ & & & & 1 \end{bmatrix}$	$\sigma^2 \cdot \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ & 1 & \rho & \rho^2 & \rho^3 \\ & & 1 & \rho & \rho^2 \\ & & & 1 & \rho \\ & & & & 1 \end{bmatrix}$

Como Fitzmaurice et al. (2008) [6] refere, apesar deste tipo de modelos terem em conta a dependência entre as observações repetidas eles não oferecem nenhuma explicação para a origem dessa dependência.

Modelo de efeitos aleatórios: O modelo de efeitos aleatórios tem como objetivo fazer inferência sobre o valor médio individual e não populacional. Estes modelos permitem por isso descrever a resposta média de cada indivíduo e a sua dependência com as covariáveis. Estes modelos incorporam além das covariáveis, efeitos aleatórios que descrevem a variabilidade do indivíduo em relação ao valor médio populacional,

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{M}_i\boldsymbol{U}_i + \boldsymbol{Z}_i, \tag{1.2}$$

onde

$$\begin{cases} \boldsymbol{Z}_i \sim MVN(0, \boldsymbol{\Sigma}_i) \\ \boldsymbol{U}_i \sim MVN(0, \boldsymbol{D}) \\ \boldsymbol{Z}_i \perp\!\!\!\perp \boldsymbol{U}_i \end{cases},$$

com $i = 1, \ldots, m$. $\boldsymbol{Y}_i$ é o vetor das variáveis respostas para o i-ésimo indivíduo, $\boldsymbol{\beta}$ e

$\boldsymbol{X}_i$ são o vetor e a matriz desenho dos efeitos fixos, respetivamente. $\boldsymbol{U}_i$ é o vetor de efeitos aleatórios de dimensão q, e $\boldsymbol{M}_i$ é a matriz desenho, com dimensão $(q \times q)$, dos efeitos aleatórios. $\boldsymbol{Z}_i$ é o vetor de erros aleatórios dentro do indivíduo. Enquanto $\boldsymbol{\beta}$ descreve mudanças da resposta média na população, $\boldsymbol{\beta} + \boldsymbol{U}_i$ descreve a trajetória das respostas individuais. A inclusão de efeitos aleatórios permite descrever as alterações dentro de cada indivíduo ao longo do tempo, flexibilizando a representação da estrutura de co(variâncias) e dando-lhe significado. Condicional a $\boldsymbol{U}_i$, Y_{ij} são independentes.

O modelo descrito em (1.2) também é conhecido por modelo linear misto para um único nível de agrupamento proposto por Laird & Ware (1982) [8].

Na análise de dados longitudinais Gaussianos, a utilização de um modelo com efeitos aleatórios ou de um modelo marginal não implica alterações na interpretação dos coeficientes de regressão $\boldsymbol{\beta}$, uma vez que todo o modelo com efeitos aleatórios implica um modelo marginal, i.e.,

$$\boldsymbol{Y}_i \sim MVN\left(\boldsymbol{X}_i\boldsymbol{\beta},\ \boldsymbol{V}_i = \boldsymbol{M}_i\boldsymbol{D}\boldsymbol{M}_i^{\top} + \boldsymbol{\Sigma}_i\right),$$

No entanto, os dois modelos não são equivalentes porque o modelo marginal não define explicitamente a estrutura de efeitos aleatórios. O modelo de efeitos aleatórios implica um modelo marginal, mas o inverso não se verifica. Modelos com diferentes estruturas de efeitos aleatórios, definidos pela matriz $\boldsymbol{D}$, podem implicar o mesmo modelo marginal [9]. Assim, a decisão da escolha entre os dois tipos modelos deve ser guiada pelas questões que motivam o estudo. Por exemplo, podemos estar interessados em selecionar entre as várias opções terapêuticas disponíveis aquela que trará maior benefício a um paciente particular. Por outro lado, o interesse pode recair em identificar o tratamento que melhor reduz a morbilidade populacional. No primeiro exemplo, a escolha deve recair sobre um modelo de efeitos aleatórios, e no segundo sobre um modelo marginal.

Nesta secção introduziram-se conceitos fundamentais da análise de longitudinal essenciais para a compreensão do trabalho exposto nos próximos capítulos. Para um estudo mais abrangente e detalhado dos conceitos mencionados recomenda-se a leitura das seguintes obras: Diggle et al. (2002) [2], Verbeke & Molenberghs (2000) [9], e Pinheiro & Bates (2006) [10].

1.3 Modelos para análise de dados de sobrevivência

Dados de sobrevivência, ou dados tempo-até-evento, descrevem o tempo decorrido desde um ponto de referência até à ocorrência de um evento de interesse num indivíduo. O evento de interesse pode ser, por exemplo, a morte do indivíduo durante um ensaio clínico, o fim de um período de desemprego, ou a falha de um equipamento eletrónico. O tempo decorrido entre o instante inicial e a ocorrência do evento designa-se, habitualmente, por tempo de sobrevivência ou tempo de vida. Os modelos de regressão para dados de sobrevivência permitem-nos estudar a distribuição de tempo de vida de um grupo de indivíduos, e determinar a relação entre a distribuição do tempo de vida e um conjunto de fatores (covariáveis) que se supõe afetarem o tempo de sobrevivência.

Uma caraterística importante deste tipo de dados é a possibilidade de existência de dados censurados. Esta idiossincrasia condiciona a aplicação de muitos métodos estatísticos. Um tempo-até-evento diz-se censurado quando o tempo até à ocorrência do evento não é conhecido com exatidão. Os métodos para análise de sobrevivência, comparativamente aos métodos estatísticos clássicos, são capazes de lidar com a presença de dados censurados. Apesar de o tempo-até-evento não ser conhecido com exatidão, esses dados não são ignorados e continuam a ser incluídos na análise sem que haja perda de informação.

Seja T_i uma v.a. absolutamente contínua e não negativa, que representa o tempo-até-evento do indivíduo i pertencente a uma dada população homogénea. Por população homogénea entende-se um grupo de indivíduos que não se distinguem entre si por fatores capazes de influenciar a sua sobrevivência. A probabilidade de o i-ésimo indivíduo sobreviver para além do tempo t_i é representado pela sua função de sobrevivência

$$S(t_i) = P(T_i > t_i) = \int_{t_i}^{\infty} f(u)\, du, \quad t_i \geq 0. \tag{1.3}$$

A função de sobrevivência é monótona decrescente, contínua, com $S(0) = 1$ e $\lim_{t_i \to +\infty} S(t_i) = 0$.

A função densidade de probabilidade da v.a. T_i representa-se por

$$f(t_i) = \lim_{dt \to 0^+} \frac{P(t_i \leq T_i < t_i + dt)}{dt}. \tag{1.4}$$

Um função particularmente útil no contexto da análise de dados tempo-até-evento é a função de risco. Esta função descreve a evolução, em função do tempo, da probabilidade instantânea de se observar o evento de interesse num indivíduo. Por outras palavras, é a probabilidade instantânea que o evento ocorra no tempo t_i dado que o individuo i sobreviveu até ao tempo t_i. A função de risco, também designada por função intensidade, força de mortalidade, ou taxa de falha é dada por

$$h(t_i) = \lim_{dt \to 0^+} \frac{P(t_i \leq T_i < t_i + dt \mid T_i \geq t_i)}{dt}, \tag{1.5}$$

com $h(t_i) \geq 0$ e sem limite superior. A função de risco pode apresentar diferentes formas, podendo ser não monótona ou monótona. As formas mais comuns são monótona crescente ou decrescente, constante, curva banheira ou unimodal [11]. A forma traduz o modo como o risco de evento se modifica ao longo do tempo. Independentemente da forma da função de risco, a função de sobrevivência é sempre decrescente.

A função de risco cumulativo, $H_i(t_i)$, representa-se por

$$H(t_i) = \int_0^{t_i} h(u)\, du, \quad t_i \geq 0, \tag{1.6}$$

e é uma função monótona crescente não negativa.

A distribuição de T pode ser univocamente especificada através de qualquer uma das funções apresentadas. Apresentam-se abaixo algumas relações úteis entre as definições (1.3- 1.6) introduzidas anteriormente, usadas recorrentemente na análise de dados tempo-até-evento:

$$f(t_i) = -\frac{\partial S(t_i)}{\partial t_i},$$

$$h(t_i) = \frac{f(t_i)}{S(t_i)},$$

$$S(t_i) = \exp\left(-H(t_i)\right) \Leftrightarrow H(t_i) = -\log\left(S(t_i)\right),$$

$$f(t_i) = h(t_i)\left(-H(t_i)\right).$$

As provas das relações anteriores podem ser consultadas em [12].

Quando se pretende modelar o tempo-até-evento em populações homogéneas é comum utilizar uma das seguintes distribuições contínuas univariadas para descrever a v.a. T: exponencial, gama, Weibull, Gompertz, log-normal, ou log-logística. Pela sua relevância para o trabalho exposto nos próximos capítulos, introduz-se de seguida a distribuição exponencial e a distribuição log-normal:

Distribuição exponencial: Seja T uma v.a. que com distribuição exponencial de parâmetro $\lambda > 0$, a função densidade de probabilidade, a função de risco e a função de sobrevivência são dadas, respetivamente por,

$$\begin{aligned} f(t) &= \lambda \exp\left(-\lambda \cdot t\right), \\ h(t) &= \lambda, \\ S(t) &= \exp\left(-\lambda \cdot t\right), \end{aligned}$$

com $t \geq 0$, e $\lambda > 0$. Na Figura 1.1 apresenta-se a título de exemplo casos particulares das três funções.

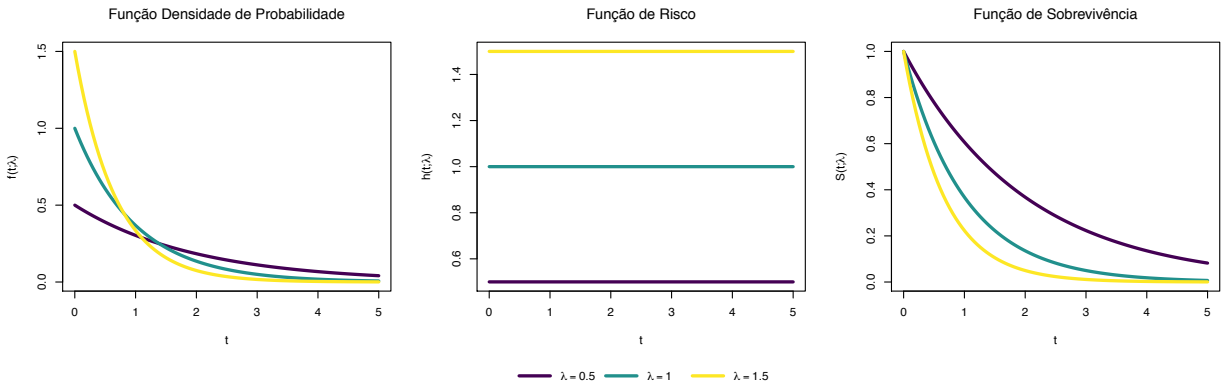


Figura 1.1: Representação de casos particulares da função densidade de probabilidade, função de risco e função de sobrevivência da distribuição exponencial.

Esta distribuição é a única a apresentar uma função de risco constante, i.e. o risco de se observar o evento de interesse é o mesmo em qualquer instante independente do

tempo decorrido desde o instante inicial. Esta propriedade é designada por ausência de memória.

Distribuição log-normal: Seja T uma v.a. que com distribuição log-normal, então $\log T$ segue uma distribuição Gaussiana com valor médio μ e variância σ^2. A função densidade de probabilidade, a função de sobrevivência, e a função de risco são dadas, respetivamente, por

$$f(t) = \frac{1}{t \cdot \sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(\log t - \mu)^2}{\sigma^2}\right),$$

$$S(t) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right),$$

$$h(t) = \frac{1}{t\sigma}\frac{\varphi\left(\frac{\log t - \mu}{\sigma}\right)}{\left[1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right)\right]},$$

onde $\varphi(\cdot)$ e $\Phi(\cdot)$ são, respetivamente, a função densidade de probabilidade e a função distribuição acumulada da distribuição Gaussiana $N(0,1)$. Na Figura 1.2 apresenta-se a título de exemplo casos particulares destas três funções.

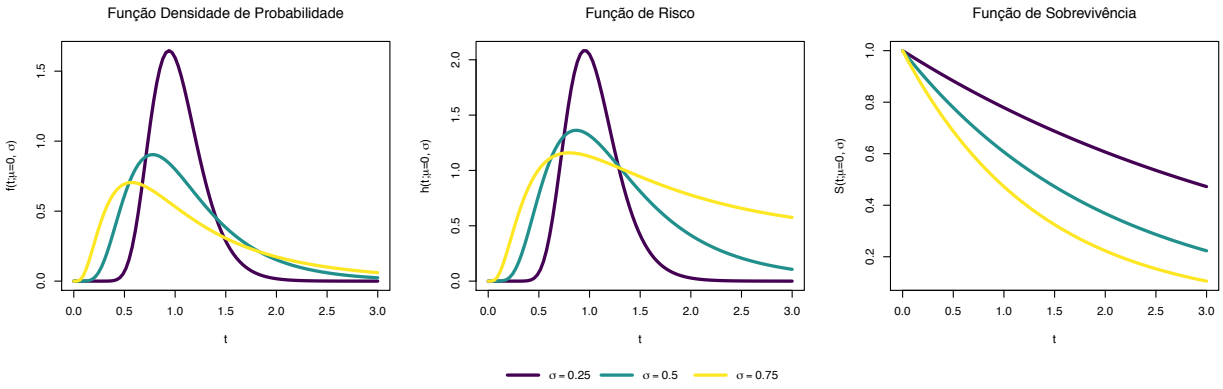


Figura 1.2: Representação de casos particulares da função densidade de probabilidade, função de risco e função de sobrevivência da distribuição log-normal.

A função de risco da distribuição log-normal é unimodal, tal que $h(0) = 0$, cresce até um valor máximo, e depois decresce com $\lim_{t\to\infty} h(t) = 0$.

Para uma análise mais pormenorizada das propriedades das distribuições apresentadas sugere-se a consulta de, por exemplo, Johson, Kotz & Balakrishnan (1995) [13].

Contudo, é frequente que a população não seja homogénea pela existência de fatores de risco que afetem os tempo de vida individuais. Exemplos de fatores de risco podem ser caraterísticas observáveis do indivíduo (e.g., o género, a idade), do desenho do estudo (e.g., tipo de tratamento), ou condições ambientais (e.g., poluição atmosférica). É, então, comum admitir-se homogeneidade entre indivíduos que apresentem os mesmos valores das covariáveis observadas. A associação entre o tempo de vida e estes fatores pode ser modelado através de um modelo de regressão, onde o tempo-até-evento é a variável resposta e as caraterísticas dos indivíduos as covariáveis. As covariáveis podem ser constantes ou variar ao longo do período de observação. As covariáveis dependentes do tempo podem ainda ser classificadas como covariáveis internas (endógenas) ou variáveis externas (exógenas). As covariáveis internas, por oposição às covariáveis externas, estão diretamente relacionadas com o mecanismo que regula o evento de interesse. As covariáveis internas requerem, por isso, a sobrevivência do paciente para a sua existência. A título de exemplo, num estudo em que se pretende estudar a recorrência de ataques de asma num grupo de pacientes, a saturação de oxigénio no sangue constitui um variável interna; e o nível de poluição atmosférica uma covariável externa.

A distribuição do tempo de vida T dado o vetor de covariáveis que descrevem um individuo pode ser definido através de uma família paramétrica de distribuições ou através de uma abordagem semi-paramétrica. A maioria dos modelos de regressão utilizados na análise de dados de tempo-até-evento podem ser enquadrados numa das seguintes categorias:

Modelos de riscos proporcionais: A função de risco de T, dado o conjunto de covariáveis $\boldsymbol{w}$, é escrito na forma

$$h(t;\boldsymbol{w}) = h_0(t) \cdot \psi(\boldsymbol{w};\boldsymbol{\alpha}),$$

em que $h_o(t)$ representa a função de risco para um indivíduo a que está associado o vetor de covariáveis $\boldsymbol{w} = \mathbf{0}$, porque $\psi(\mathbf{0}) = 1$. As covariáveis têm um efeito multiplicativo na função de risco de acordo com o fator $\psi(\boldsymbol{w})$, designado por risco relativo. Nestes modelos de regressão existe proporcionalidade entre as funções de risco correspondentes

a indivíduos com diferentes valores de covariáveis, i.e.,

$$\frac{h(t;\boldsymbol{w}_1)}{h(t;\boldsymbol{w}_2)} = \frac{h_0(t)\cdot\psi(\boldsymbol{w}_1;\boldsymbol{\alpha})}{h_0(t)\cdot\psi(\boldsymbol{w}_2;\boldsymbol{\alpha})} = \frac{\psi(\boldsymbol{w}_1;\boldsymbol{\alpha})}{\psi(\boldsymbol{w}_2;\boldsymbol{\alpha})}$$

A função de sobrevivência de T dado $\boldsymbol{w}$ é dada por

$$S(t;\boldsymbol{w}) = [S_0(t)]^{\psi(\boldsymbol{w};\boldsymbol{\alpha})} = \left[\exp\left(-\int_0^t h_0(u)\,du\right)\right]^{\psi(\boldsymbol{w};\boldsymbol{\alpha})}.$$

O modelo de regressão de Cox [14] é um modelo de riscos proporcionais extensivamente usado em análise de sobrevivência. É um modelo de regressão semi-paramétrico porque, enquanto o efeito das covariáveis é modelado parametricamente, $\psi(\boldsymbol{w};\boldsymbol{\alpha}) = \exp\left(\boldsymbol{\alpha}^\top\boldsymbol{w}\right)$, a função de risco subjacente $h_0(t)$ não é especificada. É esta última caraterística que torna o modelo muito flexível, adequado a um grande número de situações e, por isso, muito popular. No entanto, considerando formas particulares para $h_0(t)$ é possível obter modelos de regressão paramétricos de riscos proporcionais, e.g., modelo exponencial ou o modelo Weibull.

Modelos de tempo de vida acelerado: A função de sobrevivência de T, dado o conjunto de covariáveis $\boldsymbol{w}$, é

$$S(t;\boldsymbol{w}) = S_0\left(t\cdot\psi(\boldsymbol{w};\boldsymbol{\alpha})\right),$$

em que $S_0(t)$ representa a função de risco de um indivíduo que a que está associado o vetor de covariáveis $\boldsymbol{w} = \mathbf{0}$, com $\psi(\mathbf{0};\boldsymbol{\alpha}) = 1$. As covariáveis têm um efeito multiplicativo em t através do fator de aceleração $\psi(\boldsymbol{w};\boldsymbol{\alpha})$. Se $\psi(\boldsymbol{w};\boldsymbol{\alpha}) > 1$ as covariáveis aceleram o tempo até à ocorrência do evento de interesse. Se, por outro lado, $\psi(\boldsymbol{w};\boldsymbol{\alpha}) < 1$ o tempo até à ocorrência do evento de interesse é travado pelo efeito das covariáveis. A mediana do tempo de sobrevivência de um indivíduo com o vetor de covariáveis $\boldsymbol{w}_1$ é igual à mediana do tempo de sobrevivência do indivíduo de referência, i.e., com $\boldsymbol{w} = \mathbf{0}$, multiplicada pelo inverso do fator de aceleração $\psi(\boldsymbol{w}_1;\boldsymbol{\alpha})$.

A função de risco de T, dado o conjunto de covariáveis $\boldsymbol{w}$, é

$$h(t;\boldsymbol{w}) = h_0\left(t \cdot \psi(\boldsymbol{w};\boldsymbol{\alpha})\right) \cdot \psi(\boldsymbol{w};\boldsymbol{\alpha}),$$

em que $h_0(t)$ representa a função de risco de um indivíduo com $\boldsymbol{w} = \mathbf{0}$. Estes modelos são também conhecidos por modelos log-lineares ou modelos de localização-escala. Tal como nos modelos de riscos proporcionais, também neste tipo de modelos é possível considerar uma abordagem paramétrica ou semi-paramétrica, dependendo se a função $h_0(t)$ é ou não especificada.

Pela sua pertinência para o trabalho desenvolvido, dá-se destaque ao modelo de tempo de vida acelerado log-normal [15]. O modelo para a v.a. T_i, que descreve o tempo-até evento do indivíduo i, em que $\log T_i$ segue uma distribuição Gaussiana, pode ser escrito como

$$\log T_i = \boldsymbol{w}_i^\top \boldsymbol{\alpha} + \epsilon_i,$$

onde $\boldsymbol{\alpha}$ é o vetor de parâmetros de regressão e ϵ_i é uma v.a. que representa o erro de medição com distribuição Gaussiana, e cuja distribuição não depende de $\log T_i$. Considerando a v.a. $T_0 = \exp\epsilon$, que descreve o tempo de sobrevivência de um indivíduo padrão; o tempo de sobrevivência de um indivíduo descrito pelo conjunto de covariáveis $\boldsymbol{w}_i$ é descrito pela v.a. $T = T_0 \cdot \exp(\boldsymbol{w}_i^\top \boldsymbol{\alpha})$. Outro exemplo de modelo de tempo de vida acelerado é o modelo log-logístico na sua representação log-linear.

Modelos de chances proporcionais: A chance do evento de interesse não ser observado, ou chance de sobrevivência, para além do instante t define-se por,

$$\frac{S(t)}{1 - S(t)}.$$

A chance de sobrevivência de T é dada por

$$\frac{S(t;\boldsymbol{w})}{1 - S(t;\boldsymbol{w})} = \exp\left(\psi(\boldsymbol{w};\boldsymbol{\alpha})\right) \cdot \frac{S_0(t)}{1 - S_0(t)},$$

onde $S_0(t)$ é a função de sobrevivência subjacente. As covariáveis têm um efeito multiplicativo na chance de sobrevivência pela função $\psi(\boldsymbol{w};\boldsymbol{\alpha})$. $\psi(\boldsymbol{w};\boldsymbol{\alpha})$ pode ainda ser visto como o logaritmo da razão de chances de um indivíduo com vetor de covariáveis $\boldsymbol{w}$ e do indivíduo de referência.

A função de risco para T representa-se por

$$h(t;\boldsymbol{w}) = \frac{h_0(t)}{1 + (\exp(\psi(\boldsymbol{w};\boldsymbol{\alpha})) - 1) S_0(t)},$$

onde $h_0(t)$ é a função de sobrevivência subjacente. Neste tipo de modelos a função de risco subjacente pode ser estimada parametricamente ou não parametricamente. Este modelo apresenta uma propriedade particular, denominada por funções de risco convergentes. No instante inicial a razão das funções de risco é dada por $\frac{h(0)}{h_0(0)} = \exp(-\psi(\boldsymbol{w};\boldsymbol{\alpha}))$, no entanto com o avanço do tempo o efeito das covariáveis vai-se dissipando $\lim_{t\to\infty} \frac{h(t;\boldsymbol{w})}{h_0(t)} = 1$.

É importante notar que em casos particulares o mesmo modelo pode pertencer simultaneamente a duas classes de modelos de regressão. É o caso do modelo de regressão baseado na distribuição log-logística, que pode ser representado como modelo de chances proporcionais ou como modelo de tempo de vida acelerado. Outro caso particular é o do modelo de regressão de Weibull que pode ser expresso como modelo de riscos proporcionais ou como modelo de tempo de vida acelerado.

Como referido anteriormente, em estudos de análise de sobrevivência os dados recolhidos são frequentemente censurados, na medida em que o tempo decorrido até ao evento de interesse não é conhecido com exatidão. O tempo de vida pode exceder o tempo observado (censura à direita), pode ser inferior ao tempo observado (censura à esquerda), ou pode apenas delimitar-se um intervalo de tempo em que o evento ocorreu (censura intervalar). A Figura 1.3 ilustra os diferentes tipos de censura descritos. O evento de interesse foi observado nos indivíduos 1 e 2 durante o estudo, apesar de o indivíduo 2 ter entrado um pouco mais tarde do que os outros participantes. Assim, os dados destes dois indivíduos não apresentam qualquer tipo de censura. Por outro lado, os indivíduos 2 e 3 apresentam tempo-até-eventos censurados pela direita. O evento do indivíduo 2 ocorreu depois do período de observação, e

o tempo-de-vida do indivíduo 3 é desconhecido porque abandonou o estudo antes do seu fim. O instante inicial da sua exposição ao fator de risco em estudo é desconhecido para o indivíduo 5 e, por isso, contribui com uma observação censurada à esquerda. O tempo de vida do indivíduo 6 é inferior à duração do estudo, no entanto é desconhecido o momento exato em que o evento ocorreu; e, por isso, contribui com uma observação com censura intervalar.

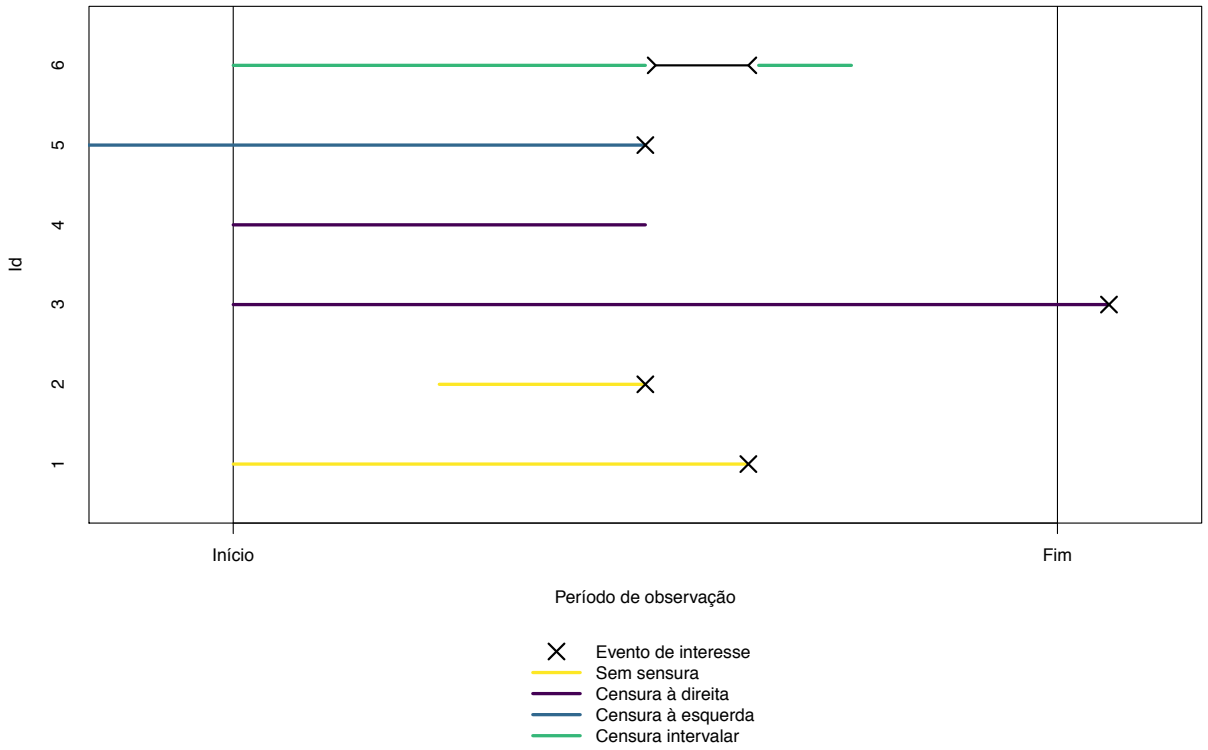


Figura 1.3: Representação gráfica dos diferentes tipos de censura que podem ser observados em conjuntos de dados tempo-até-evento.

Nesta introdução daremos à censura à direita pela sua relevância para o trabalho desenvolvido. Censura à direita verifica-se quando o evento de interesse não ocorre durante o período de observação, sabendo-se apenas que o tempo-até-evento excede o período de observação. Por exemplo, num estudo clínico em que se pretende estudar o tempo-até-morte; os pacientes que abandonam o estudo ou que sobrevivem até ao final do período de observação originam dados censurados à direita. Seja T_i a v.a. que descreve o tempo-até-evento do i-ésimo indivíduo. A cada indivíduo i corresponde um tempo de observação c_i, designado

habitualmente por tempo de censura potencial. Os dados que descrevem cada indivíduo são da forma (t_i, δ_i), onde $t_i = \min(T_i, c_i)$ e o indicador de censura δ_i toma a forma

$$\delta_i = \begin{cases} 0, \text{ se } T_i > c_i \\ 1, \text{ se } T_i \leq c_i \end{cases} .$$

A censura à direita é o tipo de censura mais comum e pode ser classificada em três categorias: de tipo I, de tipo II, ou aleatória. Diz-se que censura à direita é do tipo I, quando os períodos de observação c_i são fixados pelo desenho do estudo; o número de eventos observados é portanto aleatório. Estamos perante censura à direita do tipo II quando o desenho do estudo pré-determina que o período de observação termina quando for observado o n-ésimo evento; a duração do estudo é uma variável aleatória. A censura aleatória surge quando os indivíduos entram no estudo de forma aleatória. Neste último caso, o tempo potencial de censura de cada indivíduo c_i passa a ser aleatório, i.e., uma realização da v.a. C_i. A presença frequente de dados censurados faz com que os métodos de inferência estatística sejam, de um modo geral, baseados na teoria assintótica da máxima verosimilhança uma vez que não é possível obter distribuições de amostragem exatas [11].

O objetivo desta secção foi apresentar uma breve introdução das caraterísticas dos dados tempo-até-evento e dos modelos para a sua análise, para facilitar a compreensão do trabalho descrito no restante documento. Contudo, como a própria demonstra, esta introdução está longe de ser exaustiva. Há certamente técnicas para análise de dados tempo-até-evento que aqui não foram descritas, apesar do inquestionável mérito por elas já demonstrado na solução de inúmeros problemas nesta área do saber. Para uma introdução mais abrangente e detalhada dirige-se o leitor para os seguintes textos: Cox & Oakes (1984)[16], Klein & Moeschberger (2006)[12], e Kleinbaum & Klein (2011) [15].

Capítulo 2

Dados omissos em estudos longitudinais

Neste capítulo descreve-se a classificação de dados omissos em estudos longitudinais. É descrito o desenvolvimento da função `trim()` para o software R. Esta função permite gerar bases de dados a partir de diferentes mecanismos de dados faltantes por abandono a partir de uma base de dados completa. É ainda apresentado um estudo de simulação que faz uso da função desenvolvida.

2.1 Classificação de dados omissos

Num estudo longitudinal balanceado, espera-se que as unidades amostrados (e.g., indivíduos, animais) sejam observadas num conjunto de momentos pré-especificados. No entanto, em situações reais, é comum a presença de observações faltantes, vulgo dados omissos [2]. A omissão pode ocorrer por inúmeras razões. Por exemplo, (i) um paciente pode mudar de local de residência ou (ii) morrer da doença em estudo; nas duas situações o indivíduo deixa de estar disponível para continuar a participar no estudo e não é mais observado. Por outro lado, (iii) devido a um agravamento súbito do seu estado de saúde o paciente pode faltar a uma das visitas de monitorização, ou (iv) durante a visita um problema técnico no equipamento impede a recolha da observação; em ambos os casos o indivíduo falha uma observação mas retoma o acompanhamento na visita seguinte. Este exemplos mostram que os dados omissos num conjunto de dados longitudinais podem dizer respeito a omissões pontuais de observações (iii, iv) ou a sequências de observações (i, ii). Mostram ainda que a omissão pode, por um

Tabela 2.1: Hipotético estudo longitudinal com planeamento balanceado no qual se observam diferentes padrões de omissão. *NA* denota uma observação longitudinal omissa.

id	Tempo de Observação				
	0	2	4	6	8
1	36	32	30	27	28
2	50	44	*NA*	48	46
3	*NA*	48	33	35	39
4	75	65	70	*NA*	*NA*
5	61	*NA*	55	59	*NA*

lado, resultar de externalidades que nada têm a ver com o objecto em estudo (i, iv) ou, por outro lado, estar intimamente relacionada com o que se pretende medir (ii, iii). Estas caraterísticas permitem classificar os dados omissos em termos do seu *padrão* e *mecanismo de omissão* [17].

Em termos do *padrão* de omissão, identifica-se uma *omissão intermitente* quando um dado omisso é sucedido por pelo menos uma observação numa ocasião posterior no mesmo indivíduo. Apesar da omissão, o indivíduo permanece no estudo. Noutros casos, o indivíduo abandona o estudo antes da sua conclusão e não volta a ser observado; esse padrão é designado por *perda de acompanhamento* (*loss to follow-up*) ou *abandono* (*dropout*), consoante o mecanismo de omissão como será descrito adiante. Os dois padrões descritos anteriormente podem coexistir no mesmo indivíduo.

A Tabela 2.1 apresenta alguns padrões de omissão que podem ser encontrados na análise de dados longitudinais. Todas as observações planeadas para o indivíduo 1 foram recolhidas. Os indivíduos 2 e 3 apresentam observações omissas intermitentes nos tempos 4 e 0, respetivamente. Depois da terceira visita, o indivíduo 4 abandonou o estudo e, por isso, não foram recolhidas mais observações. O indivíduo 5 combina as duas tipologias descritas, a observação planeada para o tempo 2 não foi recolhida e deixou de ser acompanhado a partir do tempo 6.

É ainda importante notar que num estudo longitudinal com planeamento não balanceado existem inevitavelmente omissões intermitentes.

A presença de dados omissos influencia os métodos estatísticos de análise, uma vez que nem todos os métodos são capazes de trabalhar com dados incompletos. A exclusão da análise dos indivíduos com dados omissos, pode originar uma amostra não representativa da população de interesse. Por outro lado, a perda de informação associada à presença de dados omissos leva a uma redução na precisão do processo de estimação. Esta redução da precisão pode, por isso, também condicionar o modo como a análise dos dados é conduzida [7].

Outra caraterística importante a ter em conta aquando da análise de dados longitudinais com dados omissos é a razão que conduziu à sua omissão, geralmente referida como mecanismo de omissão de dados. Considerando apenas dados omissos por abandono, e seguindo a notação de Diggle et al. (2002)[2], $\boldsymbol{Y}$ denota a variável aleatória que descreve o conjunto completo de observações pretendidas num estudo longitudinal. Este vetor pode ser constituído por medições observadas ($\boldsymbol{Y}^o$) ou omissas (*missing*) ($\boldsymbol{Y}^m$), i.e.,

$$\boldsymbol{Y} = (\boldsymbol{Y}^o, \boldsymbol{Y}^m)^\top .$$

T é a variável aleatória que descreve o tempo que um indivíduo permanece dentro do estudo, i.e., tempo-até-abandono. Com isto em mente, pode-se pensar num mecanismo de omissão como a distribuição de T condicional em $\boldsymbol{Y}$,

$$[T \mid \boldsymbol{Y}] = \left[T \mid (\boldsymbol{Y}^o, \boldsymbol{Y}^m)^\top\right].$$

Rubin (1976)[18] e Rubin & Little (2019) [17] propuseram a seguinte taxonomia para os mecanismos de omissão de dados:

MCAR: Omissão completamente aleatória (*Missing completely at random*) se T é independente de ambos $\boldsymbol{Y}^m$ e $\boldsymbol{Y}^o$. A distribuição condicional de T em $\boldsymbol{Y}$ toma a forma

$$[T \mid \boldsymbol{Y}^o, \boldsymbol{Y}^m] = [T]\,; \tag{2.1}$$

MAR: Omissão aleatória (*Missing at random*) se T é dependente de $\boldsymbol{Y}^o$ mas independente de $\boldsymbol{Y}^m$, o que implica que

$$[T \mid \boldsymbol{Y}^o, \boldsymbol{Y}^m] = [T \mid \boldsymbol{Y}^o]\,; \tag{2.2}$$

MNAR: Omissão não aleatória (*Missing not at random*) se T depende de $\boldsymbol{Y}^m$ e pode ou não ser independente de $\boldsymbol{Y}^o$, i.e.,

$$[T \mid \boldsymbol{Y}^o, \boldsymbol{Y}^m] \text{ ou } [T \mid \boldsymbol{Y}^o, \boldsymbol{Y}^m] = [T \mid \boldsymbol{Y}^m]\,. \tag{2.3}$$

Acima apresentam-se os mecanismos de omissão no contexto do padrão de abandono, pela sua relevância para o trabalho desenvolvido. No entanto, os mesmos aplicam-se para observações omissas intermitentes.

Estamos na presença de um mecanismo de omissão MCAR quando a probabilidade das respostas serem omissas não depende do conjunto de respostas observadas ou a observar. Por exemplo, o paciente muda de residência para outro país e não está mais disponível para participar no estudo; o paciente morre por um motivo que não está relacionado com a doença em estudo (e.g., acidente de trânsito). Na presença de um mecanismo MCAR os dados observados podem ser considerados uma amostra aleatória dos dados completos. Os momentos - e.g. valor médio, variância, e a distribuição dos dados observados não diferem dos momentos e da distribuição dos dados completos. Por isso, em geral, os métodos de análise válidos para os dados completos continuam a produzir inferência válidas quando o mecanismo de omissão pode ser considerado MCAR [19].

Diz-se que se tem um mecanismo de omissão MAR quando a probabilidade de omissão das respostas está relacionada com o conjunto de respostas observadas. Por exemplo, o protocolo requer que pacientes cujo valor do biomarcador em análise exceda um limite por não responderem ao novo tratamento sejam removidos do ensaio clínico. Quando o mecanismo é MAR os dados observados não podem ser considerados uma amostra aleatória da população de interesse. Por isso, nem todos os métodos de análise de dados longitudinais produzem estimativas válidas na presença de um mecanismo MAR, como é o caso de métodos baseados nos momentos, e.g, o método de equações de estimação generalizadas [7]. Por outro

lado, métodos baseados na máxima verosimilhança, desde que especifiquem corretamente a distribuição de $\boldsymbol{Y}_i$, continuam conduzem a estimadores e inferências válidas [19].

Na presença de um mecanismo MNAR a probabilidade das respostas serem omissas depende do conjunto de respostas a observar, e possivelmente também das respostas observadas. Por exemplo, o paciente não completa o ensaio clínico porque morre da doença em estudo. Uma vez que a perda de informação não ocorre de forma completamente aleatória, tal como no MAR, os dados observados não representam uma amostra aleatória da resposta longitudinal de interesse. Como o mecanismo de omissão não pode ser ignorado, somente métodos de análise que modelem explicitamente a distribuição conjunta $(\boldsymbol{Y}^o, \boldsymbol{Y}^m, T)^\top$ conduzem a inferências válidas [19]. Pelo que, análises válidas quando o mecanismo de omissão é MAR deixam de ser válidas quando se tem MNAR. A partir da análise exclusiva dos dados não é possível distinguir se o mecanismo de omissão é MAR ou MNAR [20]. Para tal é necessário informação complementar por parte de quem recolheu os dados.

Neste documento, por uma questão de clareza, para distinguir padrões de omissão de abandono em função do mecanismo de omissão a ele associados, *dropout* é usado quando o indivíduo i deixa de participar no estudo por causas relacionadas com $\boldsymbol{Y}_i$ (MNAR ou MAR), e *perda de acompanhamento* caso contrário.

2.2 Desenvolvimento da função `trim()` em R

Neste trabalho desenvolveu-se em ambiente R a função `trim()` que visa simular dados omissos por abandono em conjuntos de dados longitudinais completos de acordo com diferentes mecanismos de omissão e controlando a proporção de participantes que abandonam o estudo. A designação advém do verbo em língua inglesa *to trim*, que significa aparar, cortar. Na Tabela 2.2 descreve-se a função `trim()` na perspetiva do utilizador.

Tabela 2.2: Descrição da função `trim()` em R.

Aplicação	`trim(data.arg, varying.arg, times.arg,`
	`  droptime.arg = times.arg[1], perc.drop,`
	`  mechanism.arg, summary.arg, lambdaf.arg,`
	`  nsim.arg = 1, seed.arg = NULL,`
	`  savedir.arg = NULL, savename.arg = NULL)`

Argumentos	Tipologia	Descrição
`data.arg`	`dataframe`	Conjunto de dados completo em formato *wide*.
`varying.arg`	`numeric vector`	Colunas relativas à resposta longitudinal.
`times.arg`	`numeric vector`	Tempos associados ao vetor `varying.arg`.
`droptime.arg`	`numeric`	Tempo a partir do qual pode ocorrer abandono.
`perc.drop`	`numeric`	Proporção de abandono desejada $[0, 1]$.
`mechanism.arg`	`character`	Mecanismo de omissão: `mcar`, `mar`, `mnar1`, ou `mnar2`.
`summary.arg`	`character`	Sumário da resposta longitudinal: `mean`, `median`, `range`, `gaus.wt`, ou `one`.
`lambdaf.arg`	`list`	Taxa da distribuição exponencial função de y: `list(y = c(···), rate = c(···))`.
`nsim.arg`	`numeric`	Número de simulações.
`seed.arg`	`numeric`	Semente aleatória.
`savedir.arg`	`character`	Diretório de exportação (.txt).
`savename.arg`	`character`	Nome do ficheiro a exportar em `savedir.arg`.

Para facilitar a explicação do funcionamento da função `trim()` considerar-se-á em alguns momentos o hipotético conjunto de dados apresentado na Figura 2.1 (esquerda). O conjunto de dados descreve um estudo longitudinal balanceado em que dois indivíduos foram observados em três ocasiões distintas. Admite-se que a partir deste conjunto de dados o utilizador pretender gerar um conjunto de dados incompleto, i.e., com dados omissos por abandono dos participantes. Na função `trim()` qualquer conjunto de dados incompleto observável é descrito por um vetor de dimensão m (número de indivíduos), onde o elemento na posição

Dados completos				Dados incompletos				
id	Y.t1	Y.t2	Y.t3	id	Y.t1	Y.t2	Y.t3	
1	100	80	90	1	100	80	90	$\longrightarrow \begin{bmatrix} NA \\ 2 \end{bmatrix} = \begin{pmatrix} NA & 2 \end{pmatrix}^\top$
2	110	100	80	2	110	NA	NA	

Figura 2.1: Esquerda: Hipotético conjunto de dados longitudinais completo. Direita: Hipotético conjunto de dados longitudinais com dados omissos por abandono de um dos participantes.

1	$(NA$	$NA)^\top$	⑤	$(1$	$NA)^\top$	⑨	$(2$	$NA)^\top$	⑬	$(3$	$NA)^\top$
②	$(NA$	$1)^\top$	6	$(1$	$1)^\top$	10	$(2$	$1)^\top$	14	$(3$	$1)^\top$
③	$(NA$	$1)^\top$	7	$(1$	$2)^\top$	11	$(2$	$2)^\top$	15	$(3$	$2)^\top$
④	$(NA$	$3)^\top$	8	$(1$	$3)^\top$	12	$(2$	$3)^\top$	16	$(3$	$3)^\top$

Figura 2.2: Representação de todos os conjuntos de dados incompletos observáveis a partir do conjunto de dados completos apresentados na Figura 2.1 (esquerda) por abandono do estudo dos participantes.

i indica o tempo t em que o abandono do indivíduo i foi registado, ou NA no caso de total conformidade. Este vetor é doravante denominado padrão global de abandono. Por exemplo, o conjunto de dados incompleto apresentado na Figura 2.1 (direita) é descrito pelo vetor $(NA \quad 2)^\top$. O indivíduo 1 não abandonou o estudo, pelo que o primeiro elemento do vetor é NA; o segundo elemento 2 informa que o abandono do indivíduo 2 foi detetado no tempo 2. Esta notação revelar-se-á importante na descrição do funcionamento da função `trim()` adiante. Admitindo apenas dados omissos por abandono e que é possível observar-se abandono em qualquer uma das três ocasiões em que os participantes são observados, a partir do conjunto de dados completo da Figura 2.1 (esquerda) podem ser observados 16 (4^2) conjuntos de dados incompletos (Figura 2.2).

A função `trim()` permite através do argumento `perc.drop` controlar a proporção do número total de participantes que abandonam o estudo e, assim, garantir que a proporção é observada nos conjuntos de dados gerados. Admitindo que o utilizador pretende gerar um conjunto de dados incompletos em que se observa 50% de abandono dos participantes,

`trim(···, perc.drop=0.5)`, só seria observável um dos 6 cenários assinalados com um círculo na Figura 2.2, ao invés dos 16 enumerados anteriormente. Se o utilizador pretendesse gerar dois conjuntos de dados que satisfaçam a mesma condição poderia fazê-lo fazendo `trim(···, n.sim=2)`.

A função `trim()` a partir da percentagem total de abandono especificada, do número de indivíduos no estudo, e do número de ocasiões em que é possível observar abandono, determina o número total de cenários possíveis fazendo

$$\begin{aligned} & \sum_{\forall r_1,r_2,\ldots,r_{n_D}\in\{0,1,\ldots,n_D\}:\sum_{j=1}^{n_D} r_j=n_D} \frac{n_D!}{r_1!\cdot r_2!\cdot\cdots\cdot r_{n_D}!}\cdot\binom{m}{m(1-p_D)} \\ = & \sum_{\forall r_1,r_2,\ldots,r_{n_D}\in\{0,1,\ldots,n_D\}:\sum_{j=1}^{n_D} r_j=n_D} \frac{n_D!}{r_1!\cdot r_2!\cdot\cdots\cdot r_{n_D}!}\cdot\frac{m!}{\big(m(1-p_D)\big)!\big(m+p_D\big)!}, \end{aligned}$$

onde m é o número de participantes no estudo; p_D é a proporção do número de indivíduos que abandonam o estudo (especificado pelo parâmetro `perc.drop`); n_D é o número de ocasiões em que pode ocorrer abandono (determinado a partir dos parâmetros `times.arg` e `droptime.arg`); e r_j é o número deM vezes que indivíduos abandonaram o estudo na ocasião j. A cada utilização da função é selecionado aleatoriamente um dos cenários possíveis. Os elementos do vetor que descrevem o cenário, i.e., os tempos de abandono que compõem o cenário são depois reorganizados de forma ponderada. Os factores de ponderação são calculados a partir de uma matriz de pesos que reflete a probabilidade de cada indivíduo observar abandono no tempo em análise, de acordo com o mecanismo de omissão especificado pelo utilizador pelos parâmetros da função `trim()`. No diagrama da Figura 2.3 descreve-se o modo como o cálculo dos pesos $\boldsymbol{W}$ é realizado na função. No caso do mecanismo `mcar` a reorganização do vetor é feita de forma totalmente aleatória. Esta abordagem permite, por um lado, uma computação célere. Por outro, como a seleção é feita em função do padrão global de abandono, torna possível a comparação do efeito dos diferentes mecanismos de omissão sobre os mesmos padrões globais de abandono observados, que de outro modo não seria possível.[1] Na tabela 2.3 apresenta-se uma descrição do algoritmo do funcionamento da

[1]Uma abordagem alternativa poderia passar por calcular a probabilidade de se observar cada um dos

Tabela 2.3: Algoritmo da função `trim()`:

Passo 1.	Cálculo do número total de padrões globais de abandono que satisfazem as condições impostas pelo utilizador.
Passo 2.	Seleção aleatória de `nsim` cenários.
Se `mechanism='mcar'`:	
Passo 3a.	Alocação aleatória dos registos de abandono pelos participantes no estudo.
Se `mechanism='mar'`, `'mnar1'`, ou `'mnar2'`:	
Passo 3b.	Cálculo matriz de pesos em função do mecanismo de omissão.
Passo 4b.	Alocação ponderada dos registos de abandono pelos participantes no estudo.

função `trim()`.

O utilizador especifica o modo como pretende que abandono dos indivíduos seja determinado através dos parâmetros `mechanism.arg`, `summary.arg` e `lambdaf.arg` (Tabela 2.2). As opções disponíveis para `mechanism.arg` tentam representar a taxonomia proposta por Rubin (1976)[18] e Rubin & Little (2019) [17] para os mecanismos de omissão de dados, descrita anteriormente em (2.1)-(2.3) (p. 21). Nos casos em que o mecanismo de omissão não é completamente aleatório, i.e, `mechanism.arg = mar`, `mnar1`, ou `mnar2`; admite-se que o tempo D_i que cada indivíduo permanece no estudo, i.e. tempo até abandono informativo, segue uma distribuição exponencial

$$D_{ij} \sim Exponencial\left(\lambda_{ij}\right),$$

onde o parâmetro da distribuição (taxa) λ é função de pelo menos uma das observações

cenários possíveis e a cada simulação selecionar por amostragem ponderada um cenário em função da sua probabilidade. Contudo, esta abordagem é apenas viável para conjuntos de dados completos de reduzida dimensão; tornando-se computacionalmente muito exigente e moroso para dados de média dimensão. A título de exemplo, considere-se um conjunto de dados longitudinais completos relativos a 30 indivíduos os quais podem abandonar o estudo em uma de 5 ocasiões distintas. Admitindo que se pretende simular cenários em que observe uma proporção de abandono de 0.3, podem-se observar 27943650000000 cenários.

$$\underset{m\times n}{\boldsymbol{Y}}$$

$$\begin{bmatrix} y_{11} & \cdots & y_{1n} \\ & \ddots & \\ y_{m1} & \cdots & y_{mn} \end{bmatrix}$$

$$\Big\downarrow \; y_{ij}^{s} = \mathcal{F}_1(\boldsymbol{y}_i)$$

$$\underset{m\times n}{\boldsymbol{Y}^{s}}$$

$$\begin{bmatrix} y_{11}^{s} & \cdots & y_{1n}^{s} \\ & \ddots & \\ y_{m1}^{s} & \cdots & y_{mn}^{s} \end{bmatrix}$$

$$\Big\downarrow \; \lambda_{ij} = \mathcal{F}_2(y_{ij}^{s})$$

$$\underset{m\times n}{\boldsymbol{\Lambda}}$$

$$\begin{bmatrix} \lambda_{11} & \cdots & \lambda_{1n} \\ & \ddots & \\ \lambda_{m1} & \cdots & \lambda_{mn} \end{bmatrix}$$

$$\Big\downarrow \; D_{ij} \sim Exp\,(\lambda_{ij})$$

$$\underset{m\times (n+1)}{\boldsymbol{W}}$$

$$\begin{array}{c} \\ 1 \\ \vdots \\ m \end{array} \begin{bmatrix} \overset{t_1}{} & \overset{t_2}{} & & \overset{t_n}{} & \overset{NA}{} \\ F(t_1;\lambda_{11}) & F(t_2-t_1;\lambda_{11}) & \cdots & F(t_n-t_{n-1};\lambda_{1n}) & 1-F(t_n-t_{n-1};\lambda_{1n}) \\ & \ddots & & & \vdots \\ F(t_1;\lambda_{m1}) & F(t_2-t_1;\lambda_{m1}) & \cdots & F(t_n-t_{n-1};\lambda_{mn}) & 1-F(t_n-t_{n-1};\lambda_{mn}) \end{bmatrix}$$

Figura 2.3: Diagrama do modo como a matriz de pesos para a alocação aos participantes dos tempos de abandono do cenário selecionado é realizado na função `trim()`.

longitudinais do indivíduo i na forma

$$\lambda_{ij} = \mathcal{F}_1\left(\mathcal{F}_2\left(\boldsymbol{y}_{ij}\right)\right). \tag{2.4}$$

A função $\mathcal{F}_2\left(\boldsymbol{y}_i\right)$ é especificada pelo utilizador através dos parâmetros `mechanism.arg` e `summary.arg`. Na Tabela 2.4 apresenta-se em detalhe as 13 combinações disponibilizadas ao utilizador, que lhe conferem uma grande flexibilidade de simulação. O parâmetro `mechanism.arg` determina quais as observações longitudinais do indivíduo i que devem ser consideradas no tempo j, correspondentes a $\boldsymbol{y}_{ij}$ em (2.4). De forma sucinta: a opção `mcar`, tratando-se de uma mecanismo completamente aleatório, não é função das respostas longitudinais do indivíduo; a opção `mar` é função das suas respostas longitudinais passadas; a opção `mnar1` é função de toda a sequência longitudinal recolhida do indivíduo; e a opção `mnar2` é função das suas respostas longitudinais presentes e/ou futuras. O parâmetro `summary.arg` descreve a função $\mathcal{F}_2\left(\cdot\right)$ com argumento $\boldsymbol{y}_{ij}$ em (2.4). As opções `mean` (média), `median` (mediana), `range` (variação) são autoexplicativas pela sua designação. A opção `gaus.wt` descreve uma média aritmética ponderada pela função densidade de probabilidade (f.d.p.) de uma distribuição Gaussiana padronizada, $N(0,1)$, no qual a contribuição (peso) de cada observação é função da sua distância temporal. As observações mais próximas no tempo recebem um peso superior, e vice-versa, como ilustrado na Figura 2.4. Na opção `one` a função considera apenas a observação observada ou a observar consoante o mecanismo de omissão é MAR ou MNAR, respetivamente. A função $\mathcal{F}_1\left(\cdot\right)$ em (2.4) é especificada pelo utilizador pelo parâmetro `lambdaf.arg`. Não se restringe o utilizador a um conjunto de funções pre-definidas, conferindo-lhe liberdade total para especificar a sua própria função através de uma variável do tipo `list` na forma `list(y = c(···), rate = c(···))`. Algumas possíveis funções são apresentadas na Tabela 2.5.

A função `trim()` devolve uma estrutura de dados `data.frame` que reúne o(s) conjunto(s) de dados com observações omissas e outras informações relativas à simulação (Tabela 2.6). Ao utilizador é ainda dada a possibilidade de exportar os resultados em formato *.txt* para análise posterior, definindo a diretoria e o nome do ficheiro através dos parâmetros `savedir.arg` e `savename.arg`, respetivamente. O código R integral da função `trim()` é disponibilizado no anexo A.1.

Tabela 2.4: Parametrização imposta na função `trim()` para $\mathcal{F}_2\left(\boldsymbol{y}_{ij}\right)$ em (2.4) pelos parâmetros `summary.arg` e `mechanism.arg`.

`mechanism.arg`	`summary.arg`				
	`mean`	`median`	`range`	`gaus.wt`	`one`
`mcar`	—	—	—	—	—
`mar`	$\dfrac{\sum_{k=1}^{j} y_{ik}}{j}$	$Q_{.5}\left\{y_{i1}, \cdots, y_{i(j-1)}\right\}$	$\max\left\{y_{i1}, \cdots, y_{i(j-1)}\right\} - \min\left\{y_{i1}, \cdots, y_{i(j-1)}\right\}$	$\sum_{k=1}^{j-1} w'_{ik} \cdot y_{ik} = \sum_{k=1}^{j-1} \left(\dfrac{\varphi\left(\dfrac{t_k - \mu_{mar}}{\sigma_{mar}}\right)}{\sum_{l=1}^{j-1} \varphi\left(\dfrac{t_l - \mu_{mar}}{\sigma_{mar}}\right)} \right) \cdot y_{ik}$	$y_{i(j-1)}$
`mnar1`	$\dfrac{\sum_{k=1}^{n_i} y_{ik}}{n_i}$	$Q_{.5}\left\{y_{i1}, \cdots, y_{in_i}\right\}$	$\max\left\{y_{i1}, \cdots, y_{in_i}\right\} - \min\left\{y_{i1}, \cdots, y_{in_i}\right\}$	—	—
`mnar2`	$\dfrac{\sum_{k=j}^{n_i} y_{ik}}{n_i - j}$	$Q_{.5}\left\{y_{ij}, \cdots, y_{in_i}\right\}$	$\max\left\{y_{ij}, \cdots, y_{in_i}\right\} - \min\left\{y_{ij}, \cdots, y_{in_i}\right\}$	$\sum_{k=1}^{j} w'_{ik} \cdot y_{ik} = \sum_{k=j}^{n_i} \left(\dfrac{\varphi\left(\dfrac{t_k - \mu_{mnar}}{\sigma_{mnar}}\right)}{\sum_{l=1}^{j} \varphi\left(\dfrac{t_l - \mu_{mnar}}{\sigma_{mnar}}\right)} \right) \cdot y_{ik}$	y_{ij}

Onde φ é a função densidade de probabilidade da distribuição Gaussiana $N\left(0,1\right)$, $\mu_{mar} = t_{i(j-1)}$, $\sigma_{mar} = \dfrac{t_{i(j-1)} - t_{i1}}{3}$, $\mu_{mnar} = t_{ij}$, e $\sigma_{mnar} = \dfrac{t_{in_i} - t_{ij}}{3}$.

Tabela 2.5: Exemplos de possíveis funções para $\mathcal{F}_1(\cdot)$ em (2.4) que podem ser implementadas pelo utilizador através do parâmetro `lambdaf.arg` na função `trim()`.

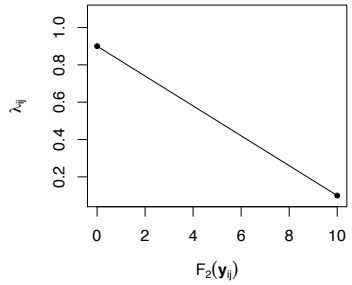
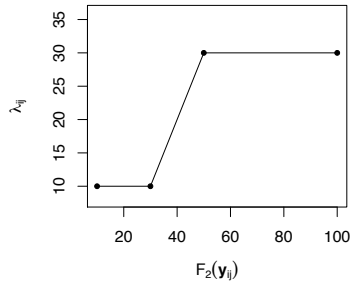
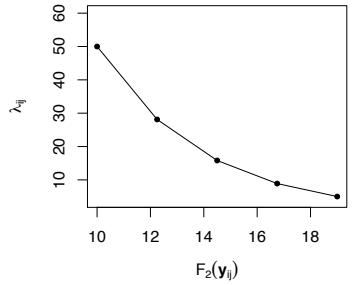
`lambdaf.arg`	Representação gráfica
`list(y = c(0, 10), rate = c(0.9, 0.1))`	
`list(y = c(10, 30, 50, 100), rate = c(10, 10, 30, 30))`	
`list(y = c(19, 16.75, 14.50, 12.25, 10), rate = c(5, 8.89, 15.81, 28.12, 50))`	

Tabela 2.6: Descrição do conjunto de dados longitudinais com dados omissos por abandono gerados pela função `trim()` a partir de um conjunto de dados completo, em formato `data.frame`.

Variáveis	Tipologia	Descrição
...	...	Variáveis presentes no conjunto de dados original.
`sim`	`integer`	Identificador da simulação.
`perc.drop`	`numeric`	Proporção do número de indivíduos que abandonaram o estudo $[0, 1]$.
`perc.miss`	`numeric`	Proporção do número de observações omissas $[0, 1]$.
`dur1`	`numeric`	Duração do processo de inicialização (s). (Comum a todos os conjuntos de dados do mesmo grupo de simulação.)*
`dur2`	`numeric`	Duração do processo de criação de cada conjunto de dados incompleto (s).*
`dtime`	`numeric`	Ocasião de registo de abandono do indivíduo, ou `NA` se o indivíduo não abandona o estudo.

∗A duração total da simulação determina-se por `>data$dur1[1] + sum(data$dur2)`

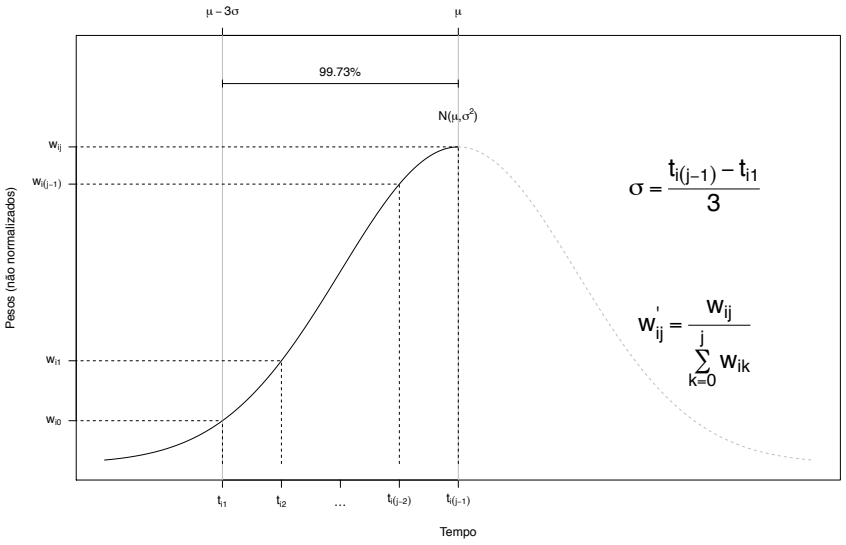


Figura 2.4: Representação do modo como a contribuição (peso) das observações é calculado recorrendo à f.d.p. da distribuição $N(0,1)$, no cálculo da média aritmética ponderada nas condições `trim(···, summary.arg = gaus.wt, mechanism.arg = mar)`.

2.3 Estudo de simulação

A presença de dados omissos num conjunto de dados longitudinais leva a uma perda de informação que por sua reduz a precisão com que a evolução da resposta média longitudinal é estimada. Para avaliar de que forma a presença de valores omissos por abandono de participantes pode afetar a precisão da estimação da progressão média da resposta longitudinal conduziu-se um estudo de simulação. A utilização de conjuntos de dados completos simulados e a função `trim()` permite-nos avaliar de forma controlada o impacto de diferentes caraterísticas, quer do conjunto de dados completo e quer dos dados omissos, no processo de estimação. O mesmo não seria possível por meio de uma base de dados real.

Este estudo visou, assim, o duplo objetivo de exemplificar uma aplicação da função `trim()`, e ilustrar de que forma as seguintes caraterísticas do conjunto de dados longitudinais e dos dados omissos,

- o número de indivíduos,
- o número de observações por indivíduo,

- a proporção de abandono,
- e o mecanismo de omissão

podem influenciar as inferências baseadas exclusivamente nos dados observados. Para cumprir este objectivo foram simulados conjuntos dados longitudinais completos com diferentes caraterísticas, aos quais foram posteriormente omitidos dados por abandono de acordo com diferentes condições.

2.3.1 Materiais e métodos

Foram geradas conjuntos de dados completos a partir de um modelo linear misto, descrito por Laird & Ware (1982) [8], da forma

$$Y_{ij} = \beta_0 + \beta_1 \cdot tempo_{ij} + U_i + Z_{ij}, \tag{2.5}$$

com $i = 1, \ldots, m$, e $j = 1, \ldots, n$. Y_{ij} representa o valor da resposta do indivíduo i na tempo j. $\beta_0 + \beta_1 \cdot tempo_{ij}$ é a componente fixa do modelo, onde $tempo_{ij}$ representa a covariável tempo, i.e., a ocasião em que a observação foi registada. $U_i + Z_{ij}$ descreve a componente dos efeitos aleatórios, onde U_i é um efeito aleatório na interseção, e Z_{ij} é o erro de medição ou ruído. As condições subjacentes ao modelo são:

$$\begin{cases} U_i \sim N(0, \nu^2) \\ Z_{ij} \sim N(0, \tau^2) \\ U_i \perp\!\!\!\perp Z_{ij} \end{cases} \tag{2.6}$$

Este modelo é comumente referido como modelo com efeito aleatório na interseção e erros heterogéneos. Com base no modelo (2.5) e nas condições subjacentes (2.6), conclui-se que a distribuição de $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in})^\top$ é Gaussiana multivariada

$$\underset{n\times 1}{\boldsymbol{Y}_i} \sim MVN\left(\underset{n\times 1}{\boldsymbol{X}_i} \cdot \boldsymbol{\beta}, \tau^2 \underset{n\times n}{\boldsymbol{I}_n} + \nu^2 \boldsymbol{J}_n\right), \tag{2.7}$$

onde $\boldsymbol{I}$ é uma matriz identidade, e $\boldsymbol{J}$ uma matriz com o número 1 em todas as posições. $\boldsymbol{X}_i$ é a matriz de covariáveis dos efeitos fixos (matriz desenho) e $\boldsymbol{\beta}$ é o vetor dos efeitos fixos (parâmetros desconhecidos), e escreve-se

$$\underset{n\times 1}{\boldsymbol{X}_i \cdot \boldsymbol{\beta}} = \underset{n\times 2}{\begin{bmatrix} 1 & t_1 \\ \cdots & \cdots \\ 1 & t_n \end{bmatrix}} \cdot \underset{2\times 1}{\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}}. \tag{2.8}$$

Combinando todos os modelos especificados por (2.7) para cada indivíduo (grupo) obtém-se o modelo

$$\underset{(m\cdot n)\times 1}{\boldsymbol{Y}} = \underset{(m\cdot n)\times 2}{\boldsymbol{X}} \cdot \underset{2\times 1}{\boldsymbol{\beta}} + \underset{(m\cdot n)\times 1}{\boldsymbol{U}} + \underset{(m\cdot n)\times 1}{\boldsymbol{Z}}, \tag{2.9}$$

Denotando o número total de observações por N, $N = m \cdot n$, a distribuição da variável aleatória $\boldsymbol{Y}$ é Gaussiana multivariada dada por,

$$\underset{N\times 1}{\boldsymbol{Y}} \sim MVN\left(\underset{N\times 1}{\boldsymbol{X}_i\boldsymbol{\beta}},\ diag_N\left(\underset{N\times N}{\tau^2\boldsymbol{I}_n + \nu^2\boldsymbol{J}_n, \ldots, \tau^2\boldsymbol{I}_n + \nu^2\boldsymbol{J}_n}\right)\right), \tag{2.10}$$

onde $diag_N(\cdot)$ é uma matriz diagonal $N \times N$ de blocos com dimensão $n \times n$.

Na Tabela 2.7 é apresentada a parametrização utilizada para gerar os diferentes conjuntos de dados longitudinais completos. As funções de densidade de probabilidade das distribuições Gaussianas de U_i e Z_{ij} são representadas na Figura 2.5. A sua representação gráfica permite uma melhor percepção da distribuição da variabilidade presente nos dados entre a variabilidade entre indivíduos e o ruído.

Os conjuntos de dados completos foram gerados a partir de um modelo com a mesma estrutura, mas diferem entre si pelo número m de indivíduos e pelo número n de observações por indivíduo, com $m \in \{10, 30, 70, 120\}$ e $n \in \{5, 15, 20, 50\}$; perfazendo um total de 16 parametrizações possíveis. Para cada um dos 16 modelos geraram-se 500 conjuntos de dados completos, perfazendo um total de 8000 conjuntos de dados. Na Figura 2.6 está representada a evolução de $\boldsymbol{Y}_i$ de cada indivíduo ao longo do tempo para um conjunto de dados selecionado aleatoriamente para cada um dos diferentes cenários. O código R utilizado pode ser consultado no anexo A.2 (p. 95), e os dados gerados em `https://bit.ly/2Oe17ga`.

Tabela 2.7: Parametrização aplicada no modelo (2.5).

Parâmetro	Valor
β_0	1000
β_1	$-3 \cdot \frac{n^*_{min}-1}{n-1}$
ν^2	100
τ^2	1

$*$ $n_{min} = \min\{5, 15, 20, 50\}$

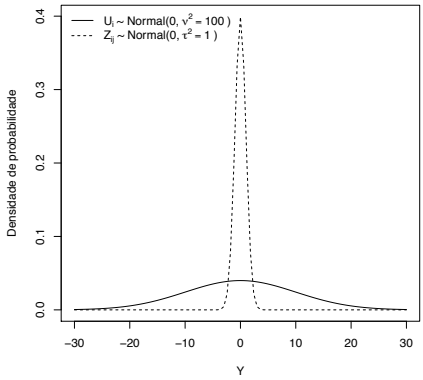


Figura 2.5: Densidade de probabilidade das distribuições de U_i e Z_{ij} presentes no modelo (2.5).

Cada conjunto de dados é identificado por $m\boxed{\mu}n\boxed{\nu}.txt$, onde μ descreve o número m de indivíduos e ν descreve o número n de observações por indivíduo. As 500 réplicas de cada cenário encontram-se no mesmo ficheiro *.txt* identificadas pela variável numérica `sim`, com `sim` $\in \{1, 2, \ldots, 500\}$. Os conjuntos de dados completos foram gerados fixando uma semente aleatória com o objetivo de permitir a sua reprodutibilidade pelo leitor.

A cada conjunto de dados completos foram removidas observações pelo abandono de alguns dos indivíduos, recorrendo à função `trim()`, respeitando diferentes condições. Cada um dos 8000 conjuntos de dados completo foi submetido a três diferentes mecanismos de omissão e dentro de cada mecanismo garantiu-se que eram observados as proporções de abandono 0.1, 0.4, e 0.6. Na Figura 2.8, a título de exemplo, apresenta-se o processo aplicado

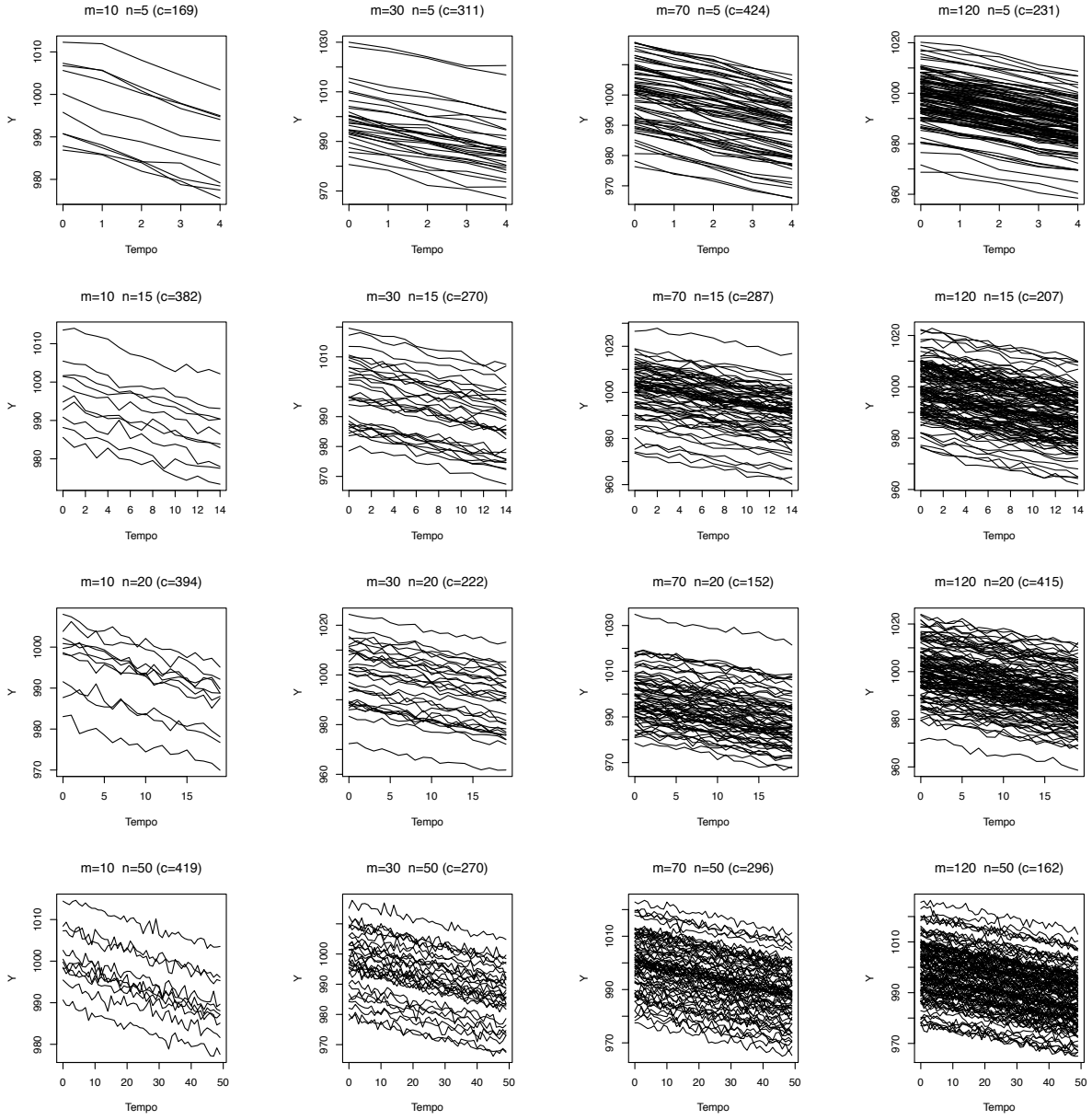


Figura 2.6: Gráfico dos perfis individuais para um conjunto de dados completo selecionado aleatoriamente entre os 500 gerados para cada uma das 16 parametrizações aplicadas. O conjunto de dados selecionado aleatoriamente é identificado pela letra *c*.

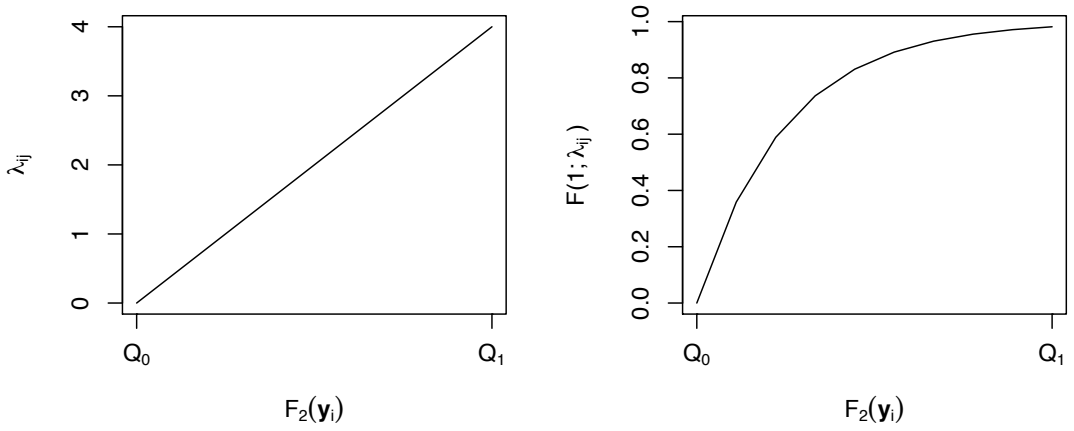


Figura 2.7: Representação gráfica da função $\mathcal{F}_1(\cdot)$ aplicada em (2.4).

ao conjunto de dados completo *m10n5.txt*. Deste modo, cada um dos 8000 conjuntos de dados completos originou 9 conjuntos de dados incompletos, originando um total de 72000 conjuntos de dados. Ao nível dos mecanismos de omissão MAR e MNAR, para $\mathcal{F}_2(\boldsymbol{y}_i)$ admitiu-se uma única observação, `trim(···, summary.arg = "one")`, e uma função linear crescente para $\mathcal{F}_1(\cdot)$, como apresentado na Figura 2.7. Os conjuntos de dados gerados são disponibilizados em `https://bit.ly/20e17ga`. Cada conjunto de dados é identificado por $m\boxed{\mu}n\boxed{\nu}\boxed{\pi}d\boxed{\delta}.txt$, onde μ descreve o número m de indivíduos, com $\mu \in \{10, 30, 70, 120\}$; ν o número n de observações planeadas por indivíduo, com $\nu \in \{5, 15, 20, 50\}$; π o mecanismo de omissão, com $\pi \in \{mcar, mar, mnar2\}$; e δ a proporção de indivíduos que abandonaram o estudo com $\delta \in \{0.1, 0.4, 0.6\}$. As réplicas dentro de cada cenário encontram-se dentro mesmo ficheiro *.txt* identificadas pela variável numérica `sim`, com `sim` $\in \{1, 2, \ldots, 500\}$. Mais uma vez, pela utilização de uma semente de aleatoriedade, o leitor conseguirá reproduzir os resultados apresentados a partir do código R disponibilizado no anexo A.2 (p. 95).

O modo como a função `trim()` foi implementada permite tempos de computação céleres, com $Q_{0.25} = 0.006s$ e $Q_{0.75} = 0.765s$ por conjunto de dados para o estudo conduzido. Na Figura 2.9 (topo esquerda) apresenta-se o histograma dos tempos de simulação para cada conjunto de dados incompleto gerado. Na mesma Figura (topo direita e base) apresentam-se os diagramas de caixa dos tempos de simulação para cada conjunto de dados incompleto em

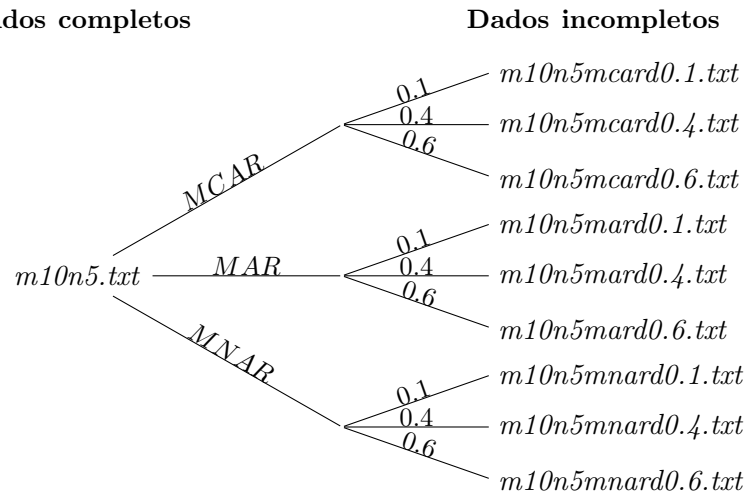


Figura 2.8: Diagrama das condições impostas a um dos conjuntos de dados completo para gerar conjuntos de dados com observações omissas.

função de diferentes caraterísticas dos dados completos ou dos dados omissos [2].

A utilização de uma semente de aleatoriedade na função `trim()` através do parâmetro `seed` permitiu garantir que os dados incompletos gerados a partir de diferentes mecanismos de omissão apresentem a mesma progressão no tempo de dados omissos e de abandono de indivíduos, removendo essa fonte de variabilidade da análise. No anexo B.1 (p. 122) podem ser consultados os gráficos da proporção de observações omissas em cada um dos cenários (Figuras B.1-B.3), e ainda a evolução da proporção de dados omissos e da proporção de indivíduos que abandonam o estudo em função do tempo (Figuras B.4-B.6) para os diferentes mecanismos.

A cada um dos 72000 conjuntos de dados com observações omissas foi ajustado um modelo linear misto com a estrutura de correlação correta, ie., a estrutura de correlação do modelo a partir do qual se geraram os dados completos. Os modelos foram ajustados fazendo uso da função `lme` do *package* `nlme` do software R, desenvolvido por Pinheiro & Bates (2006) [10]. O código R utilizado é disponibilizado no anexo A.2 (p. 95)

[2] Caraterísticas computador e software: Sistema operativo: macOS Mojave Versão 10.14.4; Processador: 2,5 GHz Intel Core i5; Memória: 8 GB 1600 MHZz DDR3; R versão 3.6.1 (2019-07-05).

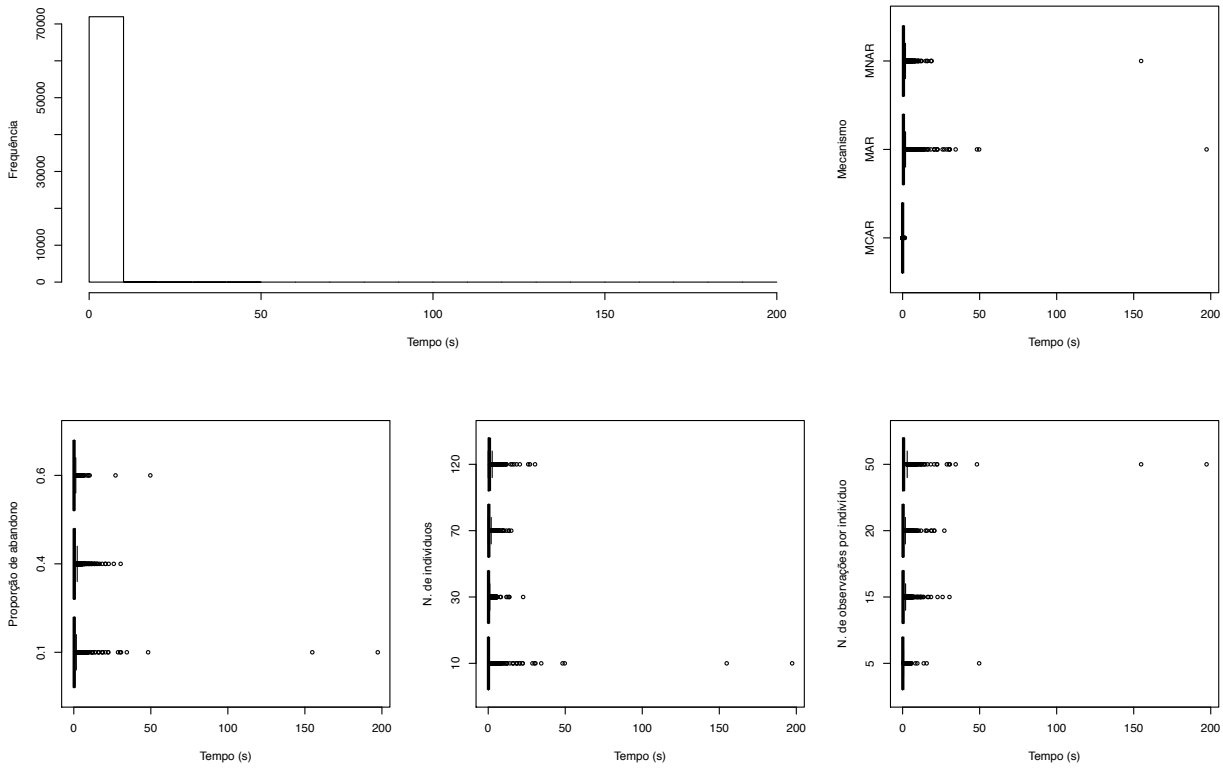


Figura 2.9: Topo esquerda: Histograma de frequência do tempo de simulação (s) para cada conjunto de dados incompleto. Topo direira e base: Diagramas de caixa do tempo de simulação (s) para cada conjunto de dados incompleto em função do mecanismo de omissão (topo esquerda), proporção de abandono (topo direita), número m de indivíduos (base esquerda), e número n de observações por indivíduo (base direita).

2.3.2 Resultados

Nas Figuras 2.10, B.7 (p. 129), e B.8 (p. 130) apresentam-se os diagramas de caixa da diferença da média de Y_j, no tempo j, entre os dados completos e os dados observados nas 500 simulações, nas últimas 5 ocasiões de observação para cada um dos 16 cenários, e em função das três proporções de abandono consideradas. A inspeção gráfica destes três conjuntos de gráficos sugere que o aumento do número m de participantes no estudo, fixando o número n de observações e a proporção de abandono, diminui a variabilidade da resposta média observada mitigando o efeito da presença de dados omissos.

O aumento da proporção de abandono, fixando o número m de indivíduos e o número de observações n por indivíduo, traduz-se num aumento da variabilidade da resposta média observada. Este efeito pode ser mais facilmente constatado na Figura 2.11.

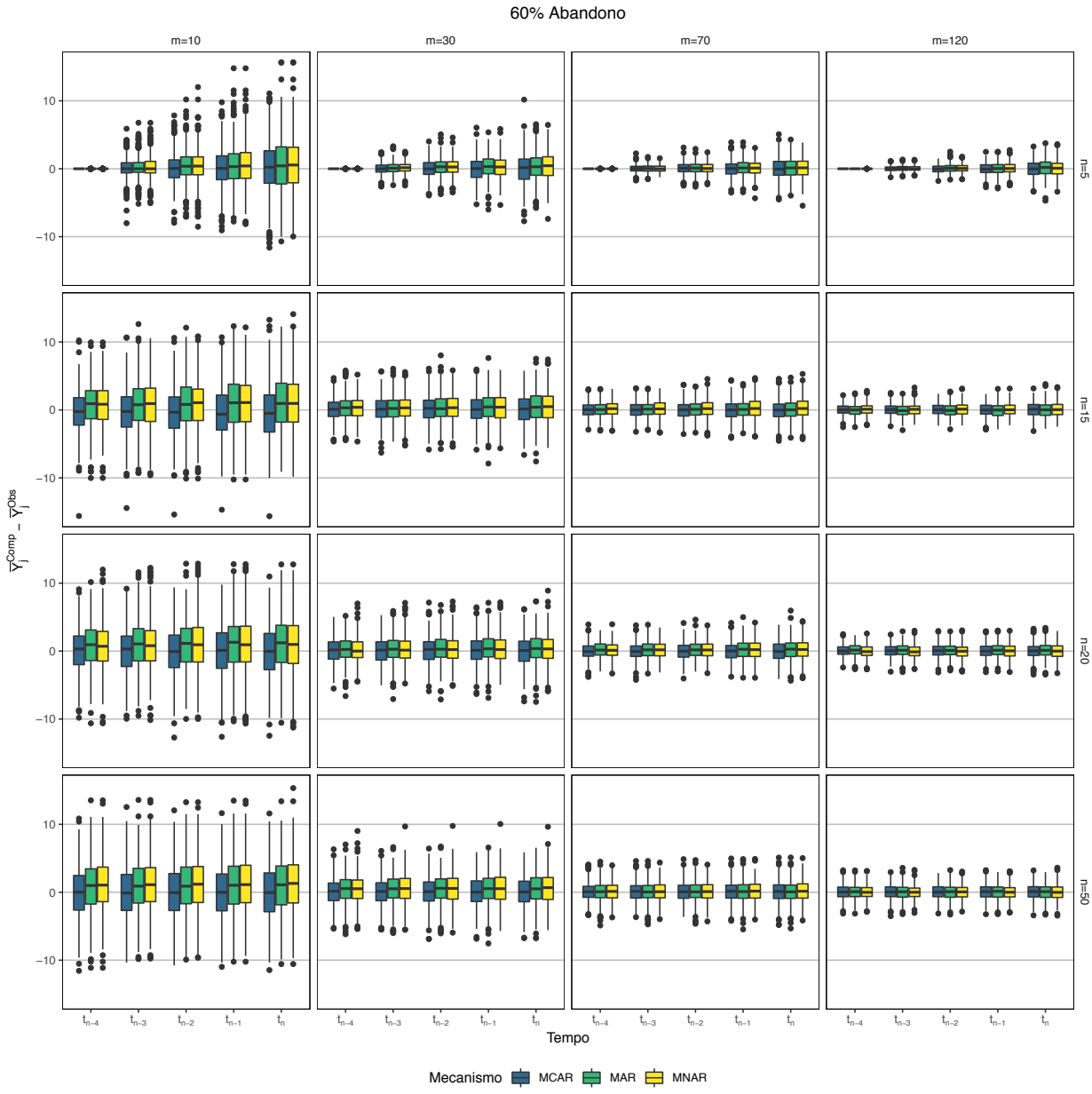


Figura 2.10: Diagramas de caixa da diferença da média de $\boldsymbol{Y}_j$, no tempo j, entre os dados completos e os dados observados nas 500 simulações, nas últimas 5 ocasiões de observação, e na presença de 60% de abandono dos indivíduos.

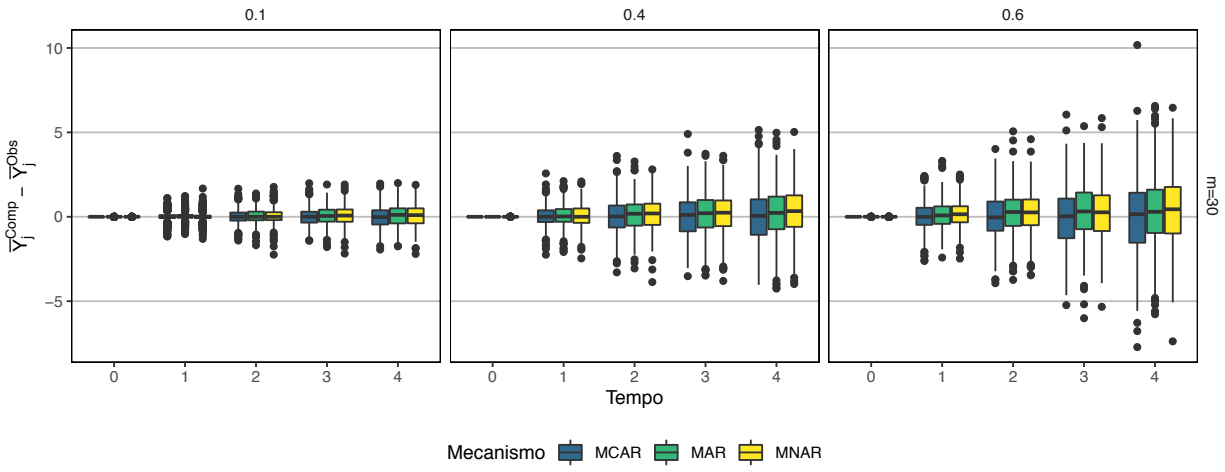


Figura 2.11: Diagramas de caixa da diferença da média de $\boldsymbol{Y}_j$, no tempo j, entre os dados completos e os dados observados nas 500 simulações, considerando os conjuntos de dados com 30 participantes e 5 observações por indivíduo.

Ao nível dos mecanismos de omissão, é possível constatar que a mediana da diferença entre os valores médios de $\boldsymbol{Y}_j$, nos tempos de observação j, entre os dados completos e os dados observados é aproximadamente nula no mecanismo MCAR. Verificando-se desvios positivos para os mecanismos MAR e MNAR. Estes desvios refletem o mecanismo de omissão imposto na função `trim()` e apresentado na Figura 2.7. Indivíduos que apresentam medições mais elevadas são mais susceptíveis de sofrer abandono, permanecendo no estudo os indivíduos que apresentam medições mais reduzidas. Assim, é expetável que a resposta nos dados incompletos pelos mecanismos MAR e MNAR seja inferior à resposta média nos dados completos, e por isso $\bar{\boldsymbol{Y}}_j^{Comp} - \bar{\boldsymbol{Y}}_j^{Obs} > 0$. Este efeito é mais facilmente observável na Figura 2.12. Estes resultados suportam a ideia reportada na literatura, e já descrita na secção 2.1, de que na presença de um mecanismo MCAR os dados observados podem ser considerados uma amostra aleatória dos dados completos.

Nas Figuras 2.13, B.9 (p. 131), e B.10 (p. 132) apresentam-se os diagramas de caixa da diferença da média de $\boldsymbol{Y}_k$ nos últimos 4 momentos até evento k entre o coorte de indivíduos que completaram o estudo e o coorte dos que o abandonaram nas 500 simulações, para cada uma das três proporções de abandono consideradas. No coorte dos indivíduos que completaram o estudo o evento é o fim do estudo, i.e., o registo da última observação. No coorte

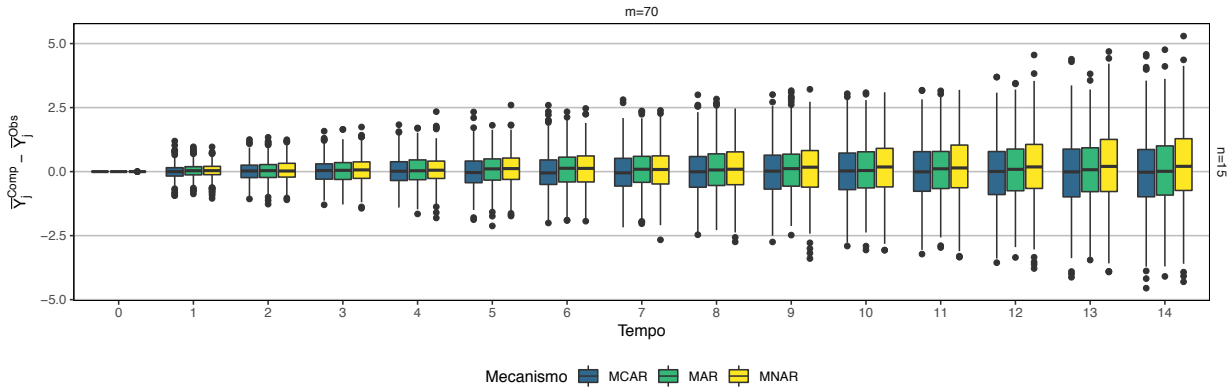


Figura 2.12: Diagramas de caixa da diferença da média de $\boldsymbol{Y}_j$, no tempo j, entre os dados completos e os dados observados nas 500 simulações, considerando os conjuntos de dados com 70 participantes, 15 observações por indivíduo e 60% de abandono.

dos indivíduos que sofreram abandono o evento de interesse é o seu abandono. Chama-se a atenção do leitor para o facto de apesar destes três gráficos serem muito semelhantes aos apresentados nas Figuras 2.10, B.7 (p. 129), e B.8 (p. 130), a informação por eles vinculada é diferente. Verifica-se que a diferença entre o valor médio é mais próxima de zero quando a razão para abandono não depende da resposta longitudinal - mecanismo MCAR. Nos restantes mecanismos, a diferença é agravada pela dependência que existe entre a resposta longitudinal e tempo-até-*dropout* ilustrada na Figura 2.7. Indivíduos com respostas $\boldsymbol{Y}_i$ mais elevadas tendem a permanecer menos tempo no estudo, i.e., são mais susceptíveis de sofrer abandono. Por outro lado, indivíduos que apresentem uma resposta longitudinal mais baixa são mais propícios a completarem o estudo. Na Figura 2.14 apresenta-se a progressão completa de um dos conjuntos de dados analisados para melhor ilustrar este fenómeno.

Nas Figuras 2.15, 2.16, 2.17, e 2.18 apresentam-se os mapas de calor[3] relativos ao erro percentual médio (EPM) dos parâmetros β_0, β_1, τ^2, e η^2, respetivamente, estimados considerando a estrutura de correlação correta especificada em (2.5) com (2.6). O EPM em cada quadrícula é calculado a partir do erro percentual observado em cada uma das 500 simulações dentro de cada cenário fazendo

[3]Os mapas, a par de todos gráficos de cor apresentados neste documento, fazem uso do esquema de cores *viridis*, desenvolvido pelos designers Stéfan van der Walt e Nathaniel Smith, para que sejam também compreendidos pelos leitores com a forma mais comum de daltonismo [21].

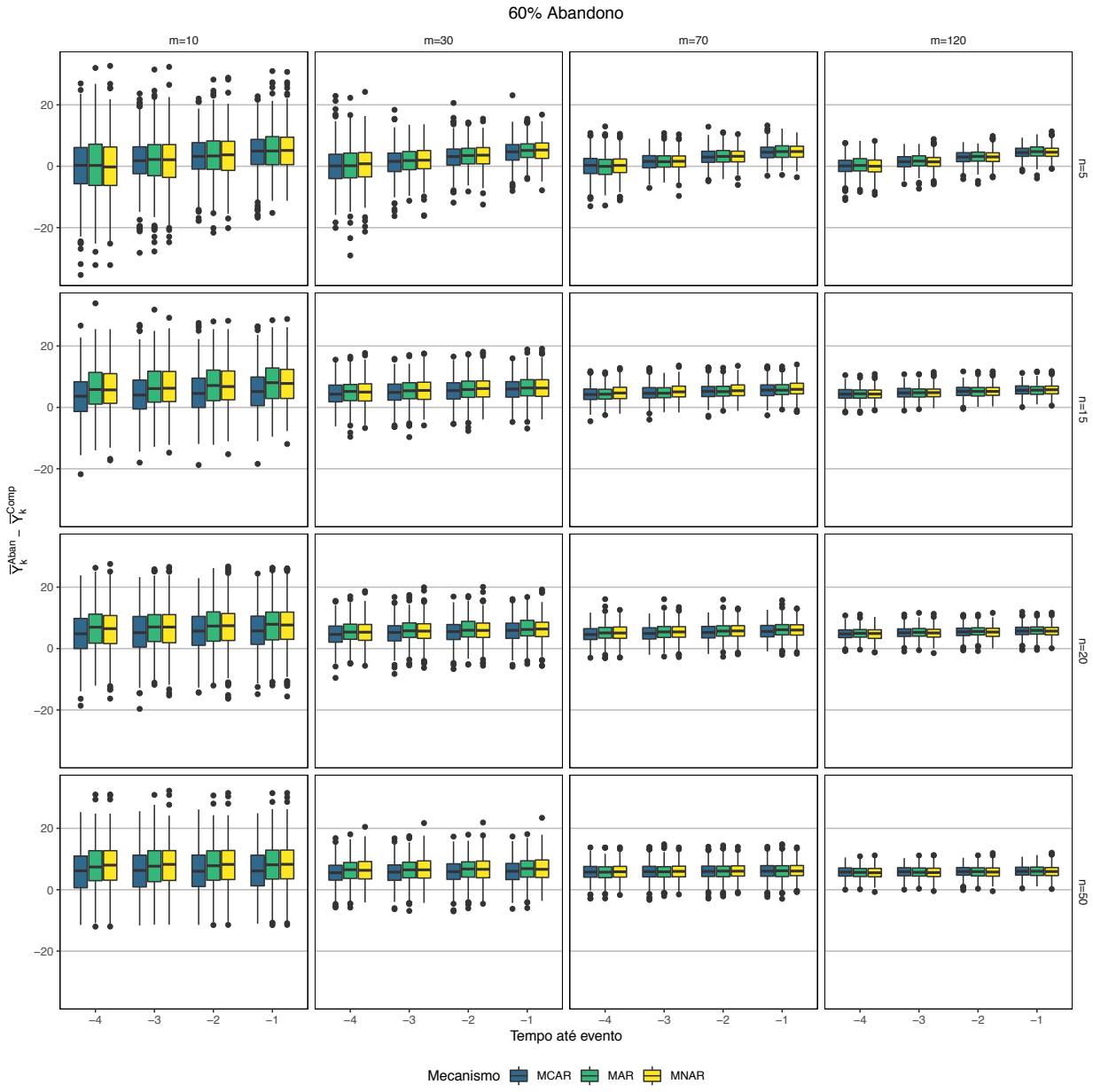


Figura 2.13: Diagramas de caixa da diferença da média de $\boldsymbol{Y}_k$, no tempo até evento k, entre os indivíduos que completaram e os que abandonaram o estudo nas 500 simulações, nos últimos 4 instantes, e na presença de 60% de abandono dos indivíduos.

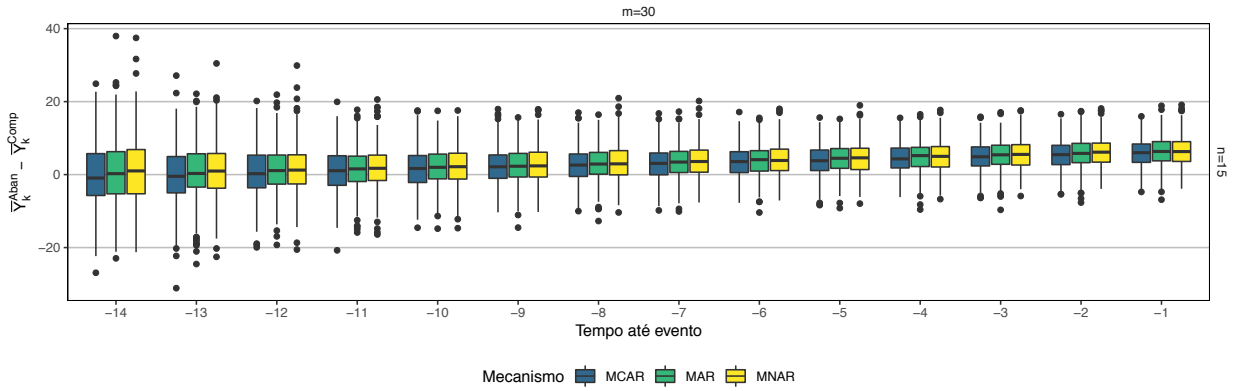


Figura 2.14: Diagramas de caixa da diferença da média de $\boldsymbol{Y}_k$, no tempo até evento k, entre os indivíduos que completaram e os que abandonaram o estudo nas 500 simulações, considerando os conjuntos de dados com 30 participantes, 15 observações por indivíduo e 60% de abandono.

$$EPM(\theta) = \frac{100}{n_{sim}} \sum_{i=1}^{n_{sim}} \left| \frac{\hat{\theta}_i - \theta}{\theta} \right|.$$

A inspeção gráfica sugere que, de um modo geral, quer o aumento do número m de indivíduos que participam no estudo quer o aumento do número n de ocasiões em que são observados reduz o EPM dos parâmetros estimados. É importante notar que o benefício do aumento de m não parece ocorrer de forma linear, diminuindo com o aumento de m. Por outro lado, o aumento do número n de observações parece não surtir efeito na estimação dos parâmetros β_0 e ν^2, tal sucede porque estes parâmetros dizem respeito ao momento inicial de cada indivíduo e no desenho do estudo não foi permitido que os indivíduos pudessem abandonar o estudo no momento inicial. Esta restrição foi imposta para garantir que se observavam os mesmos padrões globais de omissão entre os diferentes mecanismos de omissão. O mecanismo MAR depende das respostas longitudinais observadas, e como no instante inicial ainda não existem observações não é possível informar este mecanismo. Esta caraterística do desenho do estudo explica também o porquê das diferentes proporções de abandono e dos mecanismos de omissão não influenciarem a estimação destes dois parâmetros associados à ordenada na origem, β_0 e ν^2. Ao nível dos parâmetros β_1 e τ^2, um aumento da proporção da

proporção de abandono agrava o erro de estimação. Os EPM obtidos para estes dois últimos parâmetros entre os diferentes mecanismos de omissão são muito semelhantes, apresentando diferenças inexpressivas. É importante notar que no desenho do estudo se impôs a observação dos mesmos padrões globais de abandono entre os diferentes mecanismos e ainda que os conjuntos de dados fossem ajustados com a estrutura de correlação correta. Por outro lado, estes resultados podem advir do facto de as progressões individuais geradas pelo modelo considerado serem relativamente monótonas. Estas caraterísticas podem explicar estes resultados, no entanto estas hipóteses deverão ser exploradas em estudos de simulação posteriores.

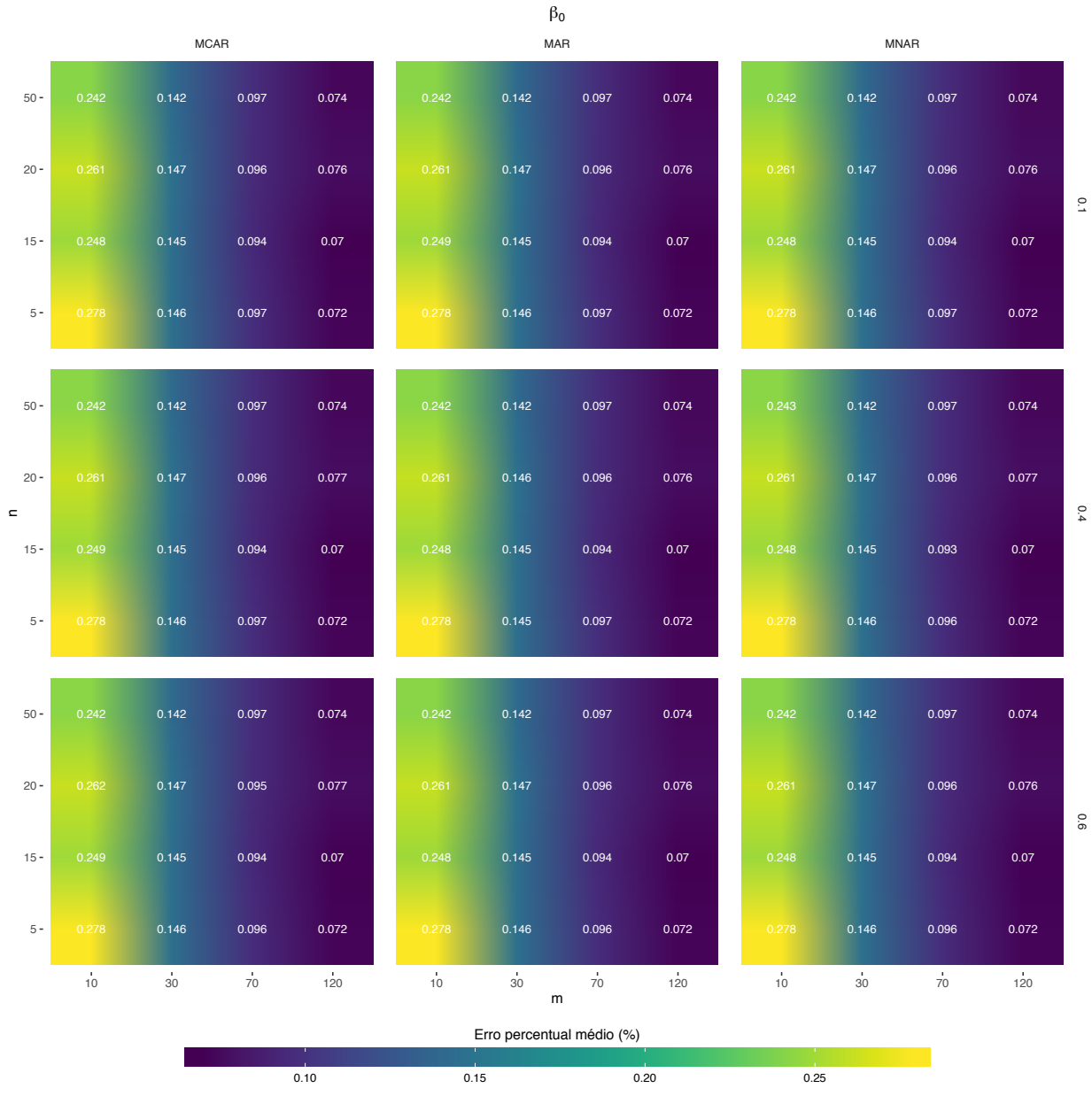


Figura 2.15: Mapa de calor do EPM do parâmetro β_0 estimado considerando a estrutura de correlação correta nas 500 simulações.

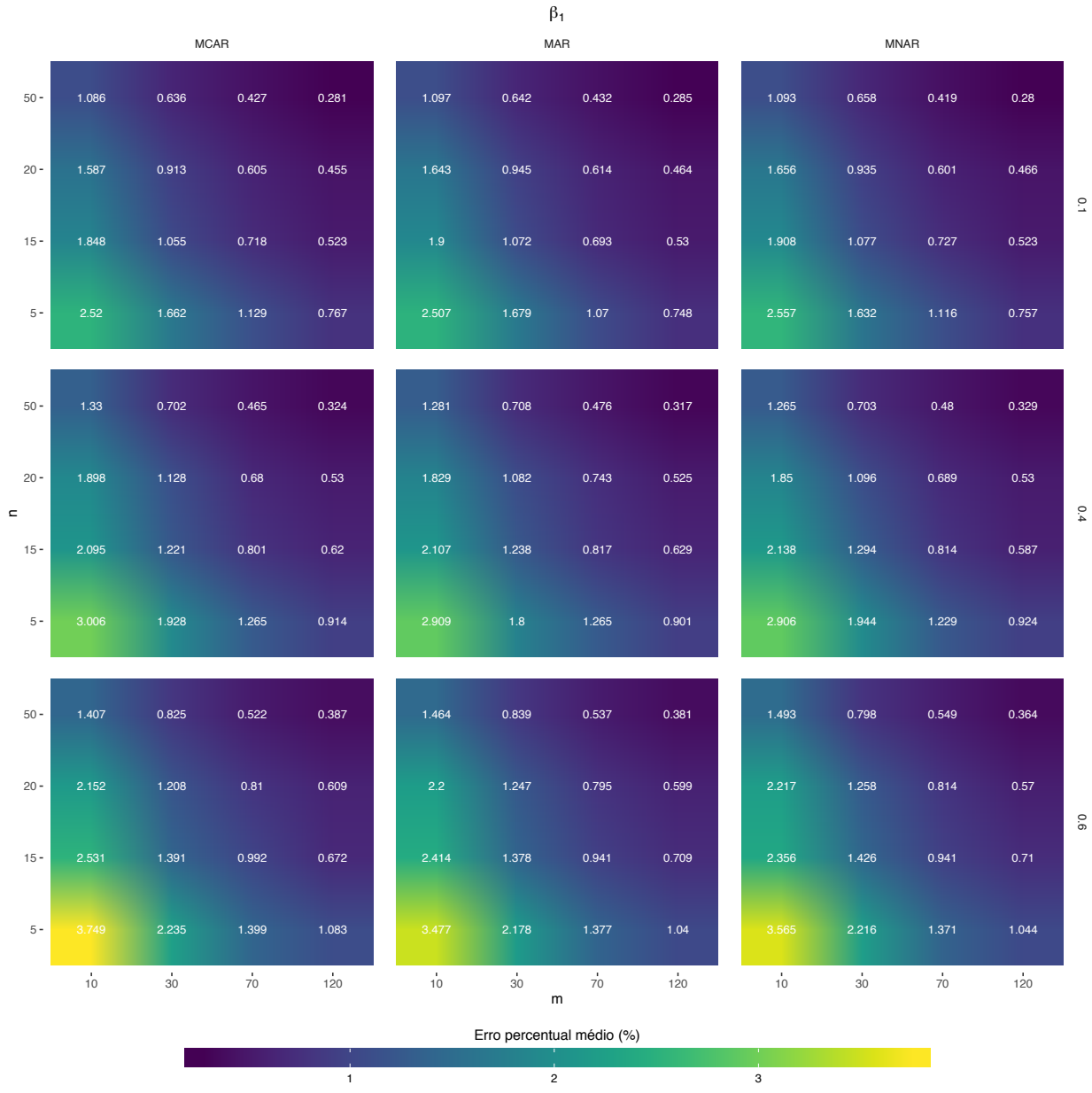


Figura 2.16: Mapa de calor do EPM do parâmetro β_1 estimado considerando a estrutura de correlação correta nas 500 simulações.

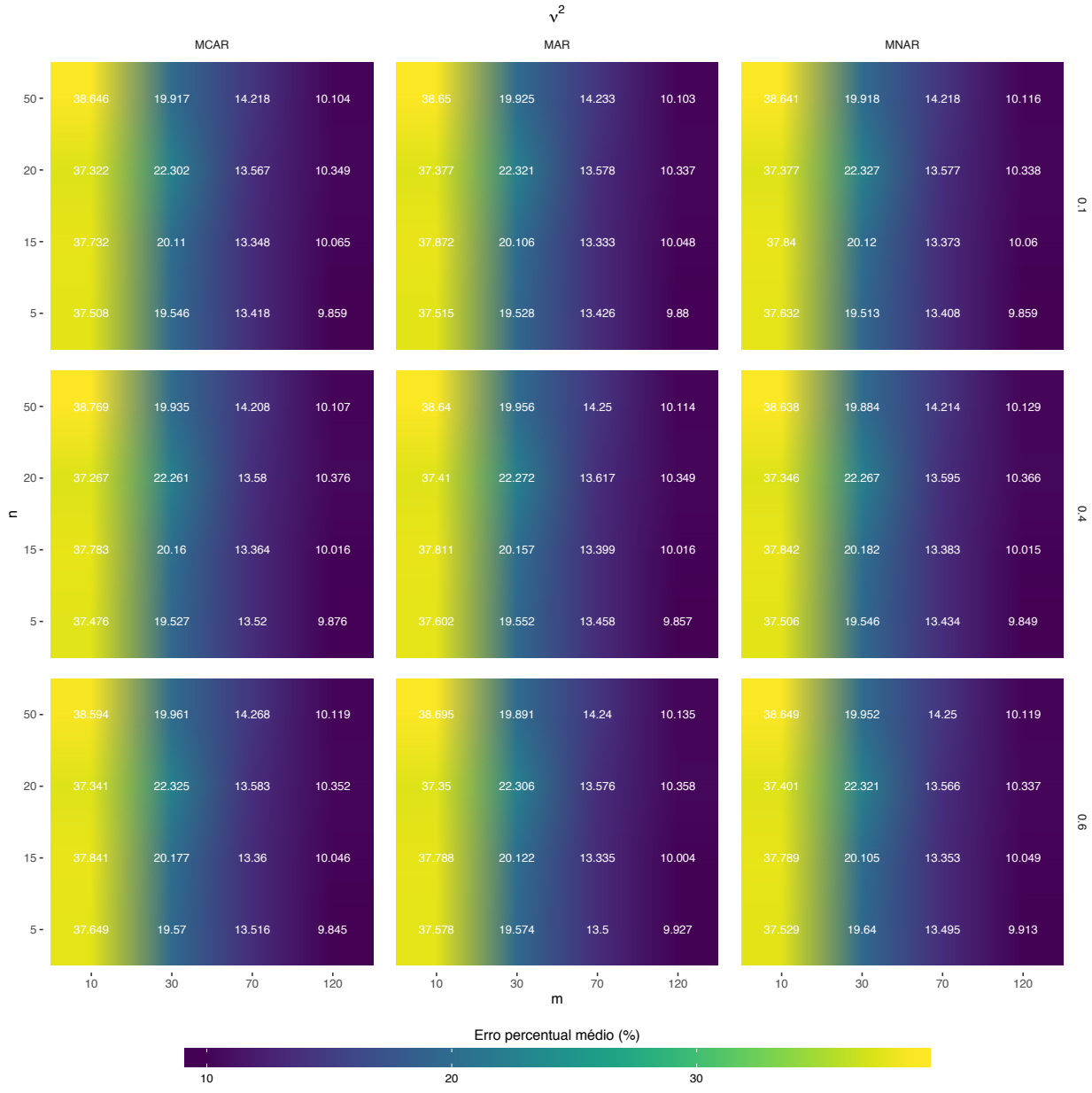


Figura 2.17: Mapa de calor do EPM do parâmetro ν^2 estimado considerando a estrutura de correlação correta nas 500 simulações.

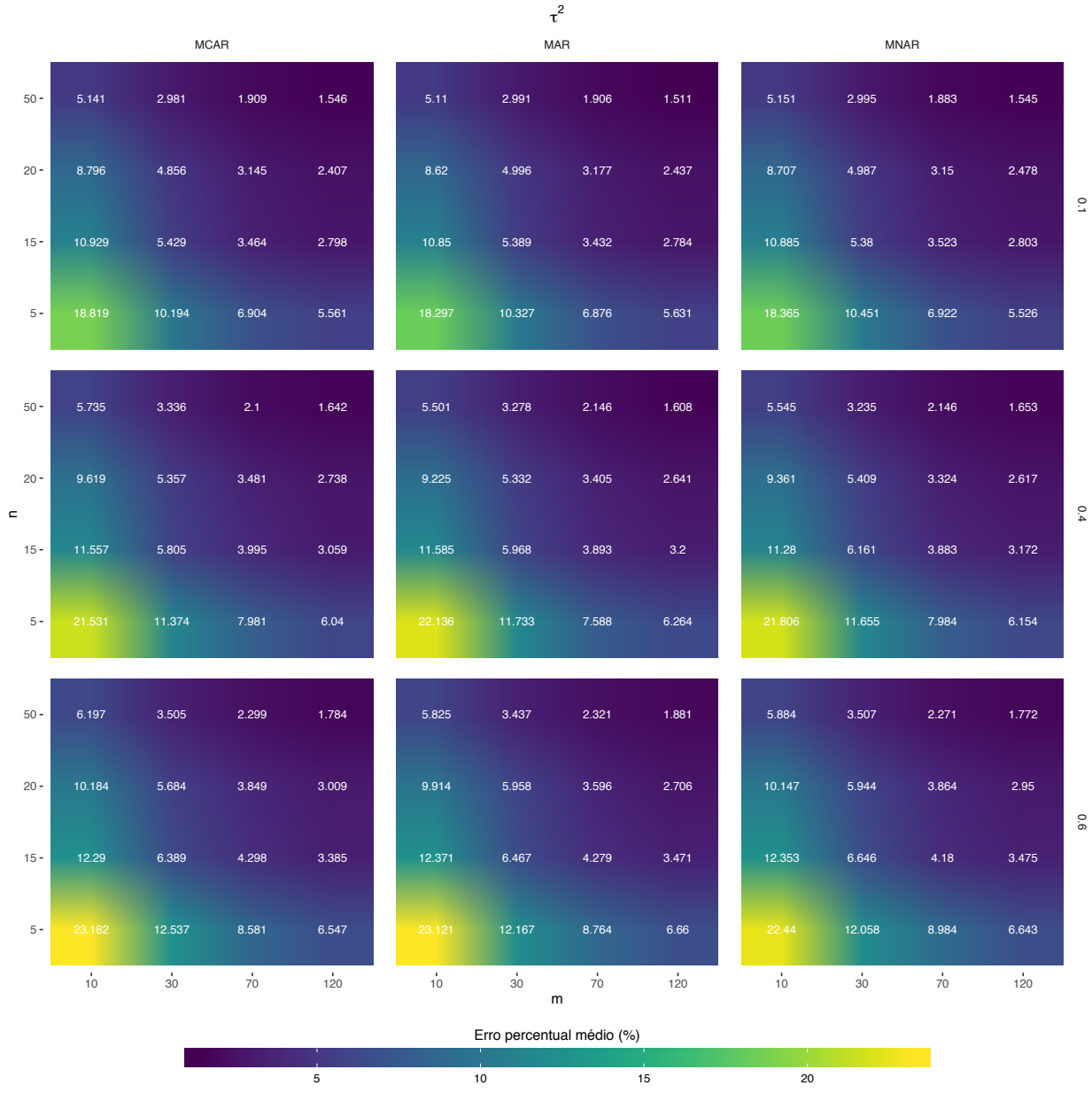


Figura 2.18: Mapa de calor do EPM do parâmetro τ^2 estimado considerando a estrutura de correlação correta nas 500 simulações.

Capítulo 3

Modelos conjuntos para dados longitudinais e de tempo-até-evento

Neste capítulo introduzimos os modelos conjuntos para dados longitudinais e de tempo-até-evento e as condições que motivam a sua aplicação. É apresentada uma revisão das principais famílias de modelos propostos na literatura.

Quando o mecanismo de omissão em dados longitudinais é informativo, é possível obter inferências válidas de análises que sejam baseadas na distribuição conjunta do processo longitudinal e do processo de omissão $[\boldsymbol{Y}, D]$ [19]. Estes modelos designam-se por modelos conjuntos para dados longitudinais e tempo-até-evento. Estes modelos surgem como alternativa a abordagens simplistas como o método última observação efetuada (*last observation carried forward*), que consiste em extrapolar as observações em falta para os indivíduos que abandonaram o estudo a partir da última observações registada [2]. No entanto, a aplicação de modelos conjuntos não se esgota na modelação de dados longitudinais em que se pretende corrigir abandono informativo. Estes modelos são também úteis quando estamos interessados na resposta tempo-até-evento mas queremos ter conta o efeito de uma covariável interna dependente do tempo medida com erro [19].

Diferentes modelos têm sido apresentados na literatura baseados em diferentes fatorizações da distribuição conjunta $[\boldsymbol{Y}, T]$, onde $\boldsymbol{Y}$ e T são as v.a. que descrevem a resposta

longitudinal e a resposta tempo-até-evento. As três principais famílias de modelos propostas para esta distribuição conjunta são [22]: (i) Modelos de seleção (*selection models*), (ii) Modelos de mistura de padrões (*pattern-mixture models*), e (iii) Modelos de efeitos aleatórios ou de parâmetros partilhados (*random effects models* ou *shared parameter models*).

3.1 Modelos de seleção

A fatorização dos modelos de seleção é baseada na relação

$$[\boldsymbol{Y}, T] = [\boldsymbol{Y}][T|\boldsymbol{Y}],$$

onde o primeiro fator é a distribuição marginal das medições longitudinais, e o segundo é a densidade do processo de omissão condicional nas respostas longitudinais. Este último fator pode ser visto como o mecanismo probabilístico que descreve a auto-seleção de cada indivíduo abandonar ou continuar no estudo. Diggle & Kenward (1994) [23] foram pioneiros na utilização de modelos de seleção para correção dados omissos por abandono não aleatório. A abordagem mais comum para a distribuição de medições repetidas é um modelo linear de efeitos mistos. A escolha para a distribuição condicional geralmente recai num modelo de regressão linear logístico, ou *probit*, onde as probabilidades são modeladas em função de algumas medições longitudinais. Contudo, alguns trabalhos propõem a utilização do modelo de riscos proporcionais de Cox ou um modelo de tempo de vida acelerado.

Esta classe de modelos podem ainda incorporar efeitos aleatórios na distribuição marginal do processo longitudinal, na forma

$$[\boldsymbol{Y}, T, \boldsymbol{U}] = [\boldsymbol{U}][\boldsymbol{Y}|\boldsymbol{U}][T|\boldsymbol{Y}],$$

passando a designar-se por modelos de seleção aleatórios (*random selection models*).

3.2 Modelos de mistura de padrões

Nos modelos de mistura de padrões a distribuição conjunta é fatorizada na forma

$$[\boldsymbol{Y}, T] = [T][\boldsymbol{Y}|T],$$

onde o primeiro fator é a densidade marginal do tempo-até-evento, e o segundo é a densidade do modelo do processo longitudinal condicional no processo de omissão. Esta fatorização atribui um modelo longitudinal distinto a cada padrão de omissão observado. A mistura de padrões é pesada pela probabilidade de se observar cada um desses padrões de omissão. Esta classe de modelos para tratamento de dados omissos de forma não aleatória foram introduzidos por Little (1993, 1994) [24, 25]. A distribuição marginal de T pode ser descrita por uma distribuição multinomial ou através da modelação da função de risco com um modelo de Cox, ou um modelo de tempo de vida acelerado. A distribuição condicional $\boldsymbol{Y}|T$ nem sempre é especificada, uma vez que o espaço amostral dos padrões de tempo-até-evento pode ser integrado a partir da distribuição condicional e, assim, as inferências são feitas diretamente na distribuição marginal de $\boldsymbol{Y}$. O principal objetivo dos modelos de mistura de padrões é ajustar a inferência sobre $\boldsymbol{Y}$ para os efeitos do evento de interesse, com a conveniência de não ser necessário especificar a distribuição marginal do tempo-até-evento.

Tal como na fatorização anterior, este modelo pode ser estendido para incluir efeitos aleatórios não observados, passando a designar-se por modelos de mistura de padrões aleatórios (*random pattern-mixture models*). A fatorização da distribuição conjunta das três v.a. toma então a forma,

$$[\boldsymbol{Y}, T, \boldsymbol{U}] = [\boldsymbol{U}][T|\boldsymbol{U}][\boldsymbol{Y}|T],$$

onde os efeitos aleatórios são incluídos na distribuição marginal do tempo-até-evento.

3.3 Modelos de efeitos aleatórios

Nos modelos de efeitos aleatórios, também designados por modelos de parâmetros partilhados, são introduzidos efeitos aleatórios que capturam a associação entre o processo longitudinal e o processo de omissão. A fatorização toma a forma

$$[\boldsymbol{Y}, T, U_1, U_2] = [U_1, U_2][T|U_1][\boldsymbol{Y}|U_2],$$

onde $\boldsymbol{U} = (U_1, U_2)^\top$ representa os efeitos aleatórios. Este tipo de modelos assenta na ideia de que existe um processo latente subjacente, descrito pelos efeitos aleatórios, que afeta os dois processos observados. Quando condicionados aos efeitos aleatórios os dois processos tornam-se independentes. Alguns dos primeiros modelos incluídos nesta classe podem ser encontrados nos trabalhos de Wu & Carrol (1988) [26], Follmann & Wu (1995) [27], e Hogan & Laird (1997)[28].

Apesar de os modelos descritos anteriormente serem matematicamente equivalentes na medida em que descrevem a mesma distribuição conjunta de variáveis aleatórias, eles têm interpretações distintas. Os parâmetros envolvidos em cada um dos componentes dos modelos têm diferentes interpretações; enquanto que num modelo podem ser os parâmetros da distribuição condicional, no outro são os parâmetros da distribuição marginal. A distribuição conjunta pode ser fatorizada de diferentes modos, e cada um desses modos traduz-se numa estratégia de modelação distinta que condiciona as interpretações possíveis e, por isso, deve ser adequada ao problema em análise [22]. Por exemplo, os modelos de seleção são usados principalmente quando o interesse inferencial está nos parâmetros do modelo de sobrevivência, melhorando a inferência permitindo a correlação entre as medições longitudinais. Por oposição, quando o interesse principal recai na trajetória longitudinal, que pode estar associada a um padrão de evento, os modelos de mistura de padrões são mais comumente usados. Como Sousa (2011) [22] sugere, a seleção da fatorização deve ser guiada, por um lado, pela questão científica que precisa ser respondidas e, por outro, pela natureza da associação entre processos.

A maioria da literatura de modelos conjuntos considera o modelo de riscos proporcionais de Cox [14] para modelar a v.a. tempo até evento. Como descrito na secção 1.3, este modelo prescinde da especificação da função de risco subjacente conferindo-lhe uma grande flexibilidade. No entanto, o modelo pressupõe que o rácio das funções de risco é constante ao longo do tempo (riscos proporcionais). Este pressuposto de proporcionalidade nem sempre é satisfeito e, portanto, há necessidade de utilizar outros modelos que prescindam dessa condição. Uma alternativa passa pela utilização de modelos paramétricos que assumem que o tempo de sobrevivência é descrito por um distribuição específica. Algumas das distribuições utilizadas são: Weibull, log-logística, log-normal, ou Gompertz. Estes modelos permitem a inferência direta do efeito que as variáveis têm no tempo de sobrevivência, a estimação por máxima verosimilhança, e incorporar funções de risco com diferentes formas.

A inferência em modelos paramétricos baseia-se essencialmente na função de verosimilhança. A verosimilhança conjunta é de fácil escrever, contudo os integrais que a compõem podem não ter soluções analíticas e requerem, por isso, a utilização de métodos numéricos de integração (e.g., aproximação Gauss-Hermite, aproximação Laplace) [29]. As técnicas de estimação podem incluir métodos Bayesianos baseados na cadeia de Markov Monte Carlo, ou métodos frequentistas como, e.g., a máxima verosimilhança, a máxima verosimilhança com estimação por *bootstrap* para os erros padrão [29].

A construção de modelos conjuntos é disponibilizada nos softwares de análise estatística padrão, como o R, SAS [30], e Stata [31]. No software R existem diferentes *packages* que permitem ajustar este tipo de modelos, nomeadamente: o `JoineR`, desenvolvido por Philipson et al. [32]; o `JM` e `JMbayes`, desenvolvido por Rizopoulos et al. [33, 34], e o `lcmm`, desenvolvido por Proust-Lima et al. [35].

Uma revisão mais detalhada da modelação conjunta de dados longitudinais e de tempo-até-evento pode ser encontrada nos trabalhos de Tsiatis & Davidian (2004) [36], Sousa (2011) [22], e Papageorgiou et al. (2019) [37].

Capítulo 4

Modelo Gaussiano transformado

Neste capítulo descreve-se o modelo Gaussiano transformado proposto por Diggle et al. (2008) [1]. São apresentadas novas estruturas de correlação para este modelo conjunto, com uma interpretação mais intuitiva. Para uma estrutura particular, apresentamos as equações dos estimadores dos parâmetros desconhecidos do modelo obtidos (i) pela derivação da função de log-verosimilhança, e (ii) pela aplicação do algoritmo EM.

Diggle et al. (2008) [1] propuseram um modelo simples totalmente paramétrico, o Modelo Gaussiano Transformado (TGM, *Transformed Gaussian Model*), para descrever a distribuição conjunta da resposta longitudinal $\boldsymbol{Y}_i$ e a transformação log do tempo-até-abandono informativo D_i, $\log D_i$. O TGM assume que o vetor de respostas $(\boldsymbol{Y}_i, \log D_i)^\top$ do i-ésimo indivíduo é uma realização de uma variável aleatória Gaussiana multivariada.

Considerando um conjunto de indivíduos independentes i, $i = 1, \ldots, m$, que se pretendem observar num conjunto de tempos pré-definidos t_j, $j = 1, \ldots, n$. Cada indivíduo é, portanto, descrito pelo conjunto de medidas longitudinais $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})^\top$, uma realização da variável aleatória $\boldsymbol{Y}_i$. Se um indivíduo abandona o estudo antes do fim do tempo de observação t_n, a observação da variável aleatória $\boldsymbol{Y}_i$ é interrompida e são recolhidas menos realizações ($n_i < n$). Como descrito em 2.1, esta interrupção pode ocorrer por causas relacionadas com $\boldsymbol{Y}_i$ (*dropout*), ou não (*perda de acompanhamento*).

A variável aleatória D_i descreve o tempo até *dropout*. Se uma sequência de medidas

longitudinais $\boldsymbol{y}_i$ é interrompido por um tempo de *dropout* d_i, isso é dizer que as medidas longitudinais y_{ij} são omissas para tempo t_j superiores ou iguais a d_i. Como descrito na secção 2.1, cada vetor $\boldsymbol{Y}_i$ pode então ser particionado no sub-vetor dos dados observados $\boldsymbol{Y}_i^o$ e no sub-vetor dos dados omissos (*missing*) $\boldsymbol{Y}_i^m$: $\boldsymbol{Y}_i = (\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m)^\top$. Se um indivíduo providenciar todas as observações pretendidas, o seu tempo tempo-até-*dropout* é censurado à direita.[1] É ainda importante notar que a presença de observações omissas não implica necessariamente a observação de um evento de *dropout*; esta pode resultar de *perda de acompanhamento* (abandono não informativo), o que leva simultaneamente a um conjunto incompleto de medições longitudinais e a um tempo-até-*dropout* censurado à direita. Portanto, o tempo de conformidade registado t_i pode ser um tempo-até-*dropout* d_i ou um tempo de censura à direita c_i. Assim sendo, $t_i = \min\{d_i, c_i\}$ e δ_i indica se a resposta longitudinal do indivíduo i terminou por *dropout* ($\delta_i = 1$) ou por perda de acompanhamento (censura) ($\delta_i = 0$). Na Tabela 4.1 apresenta-se um exemplo hipotético. Tendo em conta o acima exposto, os dados relativos ao indivíduo i podem ser convenientemente representados pelo vetor $(\boldsymbol{y}_i^o, t_i, \delta_i)^\top$, que deve ser visto como uma realização do vetor aleatório $(\boldsymbol{Y}_i, T_i, \Delta_i)^\top$ onde

$$T_i = \min\{D_i, C_i\}. \tag{4.1}$$

O TGM assume que os m vetores $(\boldsymbol{y}_i, \log d_i)^\top$ são realizações da variável aleatória multivariada Gaussiana $(\boldsymbol{Y}_i, \log D_i)^\top$, da forma

$$\underset{(n_i+1)\times 1}{\begin{bmatrix} \underset{n_i\times 1}{\boldsymbol{Y}_i} \\ \underset{1\times 1}{\log D_i} \end{bmatrix}} \sim MVN \left(\underset{(n_i+1)\times 1}{\begin{bmatrix} \underset{n_i\times p}{\boldsymbol{X}_i} \cdot \underset{p\times 1}{\boldsymbol{\beta}} \\ \underset{1\times q}{\boldsymbol{W}_i^\top} \cdot \underset{q\times 1}{\boldsymbol{\alpha}} \end{bmatrix}}, \underset{(n_i+1)\times(n_i+1)}{\begin{bmatrix} \underset{n_i\times n_i}{\boldsymbol{V}_i(\boldsymbol{\psi})} & \underset{n_i\times 1}{\boldsymbol{g}_i(\boldsymbol{\gamma})} \\ \underset{1\times n_i}{\boldsymbol{g}_i^\top(\boldsymbol{\gamma})} & \underset{1\times 1}{\eta^2} \end{bmatrix}} \right). \tag{4.2}$$

[1] Numa primeira leitura isto pode parecer confuso, dado que o *dropout* é uma causa comum de censura à direita em estudos de análise de sobrevivência. Por exemplo, se um indivíduo abandona um estudo durante o tempo de observação o evento de interesse *morte* não é observado, e o seu tempo de sobrevivência é considerado censurado à direita. No entanto, no contexto desta exposição, o evento de interesse é o *dropout* do estudo em si mesmo. Este tempo-até-*dropout* pode não ser observado nos casos de total conformidade com o protocolo e por isso gerar dados censurados à direita da variável aleatória tempo-até-evento D_i.

Tabela 4.1: Hipotético estudo longitudinal com indivíduos com conformidade total (a), *dropout* (b), e perda de acompanhamento (c).

id	Tempo de Observação					d_i	c_i	t_i	δ_i
	0	2	4	6	8				
1	y_{10}	y_{12}	y_{14}	y_{16}	y_{18}	12*	8	8	0
2	y_{20}	y_{22}	y_{24}	NA	NA	6	10*	6	1
3	y_{30}	y_{32}	NA	NA	NA	6*	10*	4	0

∗ Não observável.

As componentes do vetor de valores médios da distribuição multivariada Gaussiana em (4.2) são:

- $\boldsymbol{X}_i\boldsymbol{\beta}$ é o vetor valor médio da variável longitudinal, que é função da matriz desenho $\boldsymbol{X}_i$ e do vetor desconhecido dos efeitos fixos $\boldsymbol{\beta}$;
- $\boldsymbol{W}_i^\top\boldsymbol{\alpha}$ é a média da variável tempo-até-*dropout* transformado, que é especificado pelo vetor das variáveis independentes $\boldsymbol{W}_i$ e pelo vetor dos parâmetros desconhecidos $\boldsymbol{\alpha}$.

A sua matriz de (co)variâncias é especificada por:

- $\boldsymbol{V}_i(\boldsymbol{\psi})$ é a matriz de (co)variâncias da resposta longitudinal $\boldsymbol{Y}_i$;
- η^2 é a variância do tempo-até-*dropout* transformado $\log D_i$;
- $\boldsymbol{g}(\boldsymbol{\gamma})$ é o vetor de covariância cruzada entre cada medição longitudinal Y_{ij} e $\log D_i$.

4.1 Estruturas de correlação

O modelo TGM destaca-se de outras abordagens de modelação conjunta descritas na literatura pela sua simplicidade e rapidez de computação. Contudo, a interpretação puramente empírica do vetor de correlação cruzada $\boldsymbol{g}(\boldsymbol{\gamma})$ entre as medições longitudinais e o tempo até abandono pode ser vista como uma fraqueza. Neste trabalho, propomos novas estruturas de

correlação com uma interpretação mais intuitiva pela partilha de efeitos aleatórios e inclusão de um parâmetro de associação γ entre as duas respostas. A inclusão de efeitos aleatórios partilhados permite a fatorização da distribuição conjunta como um modelo de efeitos aleatórios, como descrito no capítulo 3.

Consideramos o modelo para o vetor das variáveis reposta $\boldsymbol{Y}_i$ e transformação do tempo-até-*dropout* $\log D_i$

$$\begin{cases} \boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{M}_i\boldsymbol{U}_i + \boldsymbol{Z}_i \\ \log D_i = \boldsymbol{W}_i^\top\boldsymbol{\alpha} + \boldsymbol{V}_i^\top\boldsymbol{\gamma} + \epsilon_i \end{cases}, \tag{4.3}$$

onde

$$\begin{cases} \boldsymbol{U}_i \sim MVN(\boldsymbol{0}, \boldsymbol{\Upsilon}) \\ \boldsymbol{Z}_i \sim MVN(\boldsymbol{0}, \tau^2\boldsymbol{I}) \\ \epsilon_i \sim N(0, \eta^2) \\ U_i \perp\!\!\!\perp Z_{ij} \perp\!\!\!\perp \epsilon_i \end{cases}.$$

O vetor $\boldsymbol{Y}_i$ é descrito por um modelo linear misto para um nível de agrupamento, como descrito por Laird & Ware (1982) [8]. A matriz $\boldsymbol{X}_i$ é a matriz de covariáveis dos efeitos fixos e $\boldsymbol{\beta}$ o vetor dos efeitos fixos. $\boldsymbol{M}_i$ é a matriz de covariáveis dos efeitos aleatórios e $\boldsymbol{U}_i$ o vetor dos efeitos aleatórios. O vetor $\boldsymbol{Z}_i$ é o vetor dos erros aleatórios dentro do grupo.

O modelo log-normal proposto para o tempo-até-*dropout* do indivíduo i, D_i, é um modelo de tempo de vida acelerado, introduzido na secção 1.3. $\boldsymbol{\alpha}$ é o vetor dos parâmetros de regressão e $\boldsymbol{W}_i$ o vetor das covariáveis. ϵ_i é o erro aleatório. $\boldsymbol{V}_i$ é o vetor de efeitos aleatórios partilhados pela resposta longitudinal $\boldsymbol{Y}_i$ e $\boldsymbol{\lambda}$ o vetor de escala associado.

Possíveis estruturas para $\boldsymbol{M}_i\boldsymbol{U}_i$ e $\boldsymbol{V}_i^\top\boldsymbol{\gamma}$ são apresentadas na Tabela 4.2. O(s) efeito(s) aleatórios pretendem explicar caraterísticas não observáveis que afetam a progressão individual da resposta longitudinal e o tempo-até-*dropout*; e, por isso, explicam a associação entre os dois processos. É pouco realista admitir que as caraterísticas observadas esvaziam a heterogeneidade populacional, i.e., que indivíduos com os mesmos valores nas covariáveis observadas são homogéneos. Tal deve-se ao facto de não ser possível incluir no nosso modelo

todos as caraterísticas/fatores de risco relevantes como covariáveis, ora pela impossibilidade prática da sua observação, ora pela sua existência ser desconhecida. Esta heterogeneidade individual não observada é extremamente importante e não pode ser descurada. Exemplos destas caraterísticas não observáveis dos indivíduos podem ser fatores genéticos, ou condições ambientais não quantificáveis (e.g., poluição doméstica, stress). Assim, as duas respostas em (4.3) partilham pelo menos um mesmo efeito aleatório, que descreve caraterísticas ambientais e/ou genéticas latentes.

Tabela 4.2: Possíveis estruturas para $\boldsymbol{M}_i\boldsymbol{U}_i$ e $\boldsymbol{V}_i^\top\boldsymbol{\gamma}$ no modelo (4.3).

Modelo	$\boldsymbol{V}_i^\top\boldsymbol{\gamma}$	$\boldsymbol{M}_i\boldsymbol{U}_i$
1	γU_i^1	U_i^1 / $U_i^1 + t_{ij}U_j^2$
2	γU_i^2	$U_i^1 + t_{ij}U_j^2$
3	$\gamma_1 U_i^1 + \gamma_2 U_i^2$	$U_i^1 + t_{ij}U_j^2$

No modelo

$$\begin{aligned} \log D_i &= \boldsymbol{w}_i^\top\boldsymbol{\alpha} + \boldsymbol{V}_i^\top\boldsymbol{\gamma} + \epsilon_i \\ \Leftrightarrow D_i &= \exp\{\boldsymbol{w}_i^\top\boldsymbol{\alpha} + \boldsymbol{V}_i^\top\boldsymbol{\gamma} + \epsilon_i\} \\ \Leftrightarrow D_i &= \exp\{\boldsymbol{w}_i^\top\boldsymbol{\alpha} + \boldsymbol{V}_i^\top\boldsymbol{\gamma}_i\}\exp\{\epsilon_i\} \end{aligned} ,$$

onde $\epsilon_i \sim N(0, \eta^2)$, e $\boldsymbol{w}_i^\top\boldsymbol{\alpha} + \boldsymbol{V}_i^\top\boldsymbol{\gamma}$ descreve caraterísticas do indivíduo i, que podem ser ou não observáveis. Notar que como

$$\begin{aligned} \log D_0 &= \epsilon_i \\ \Leftrightarrow D_0 &= \exp\{\epsilon_i\} \end{aligned} ,$$

D_i pode ser escrito na forma

$$\begin{aligned} D_i &= \exp\{\boldsymbol{w}_i^\top\boldsymbol{\alpha} + \boldsymbol{V}_i^\top\boldsymbol{\gamma}\}\exp\{\epsilon_i\} \\ &= \exp\{\boldsymbol{w}_i^\top\boldsymbol{\alpha} + \boldsymbol{V}_i^\top\boldsymbol{\gamma}\}D_0 \end{aligned} ,$$

onde D_0 pode ser vista como a distribuição para um indivíduo de referência com $\boldsymbol{w}_i^\top \boldsymbol{\alpha} + \boldsymbol{V}_i^\top \boldsymbol{\gamma} = 0$. A função de sobrevivência para qualquer indivíduo i no tempo t, $S_i(d)$, em termos da função de sobrevivência subjacente $S_0(d)$ é

$$\begin{aligned} S_i(d) &= P\left(D_i > d\right) \\ &= P\left(D_0 \exp\{\log D_i - \log D_0\} > d\right) \\ &= P\left(D_0 > d \cdot \exp\{-\left(\log D_i - \log D_0\right)\}\right) \\ &= S_0(d \cdot \exp\{-\left(\log D_i - \log D_0\right)\}) \\ &= S_0(d \cdot \exp\{-\left(\boldsymbol{w}_i^\top \boldsymbol{\alpha} + \boldsymbol{V}_i^\top \boldsymbol{\gamma} + \epsilon_i - \epsilon_i\right)\}) \\ &= S_0(d \cdot \exp\{-(\boldsymbol{w}_i^\top \boldsymbol{\alpha} + \boldsymbol{V}_i^\top \boldsymbol{\gamma})\}) \end{aligned} \tag{4.4}$$

A probabilidade de um indivíduo com covariáveis $\boldsymbol{w}_i$ e/ou efeitos aleatórios $\boldsymbol{V}_i$, permanecer no estudo longitudinal no tempo d é a mesma do indivíduo de referência no tempo $d \cdot \exp\{-(\boldsymbol{w}_i^\top \boldsymbol{\alpha} + \boldsymbol{V}_i^\top \boldsymbol{\gamma})\}$. Isto pode facilmente ser interpretado como o encolhimento ou alongamento do eixo do tempo, isto é, como o tempo passando mais rápido ou mais devagar por um fator $\exp\{-(\boldsymbol{w}_i^\top \boldsymbol{\alpha} + \boldsymbol{V}_i^\top \boldsymbol{\gamma})\}$. Característica que justifica a sua designação como modelo de tempo de vida acelerado. Se

$$\begin{aligned} &\exp\{-(\boldsymbol{w}_i^\top \boldsymbol{\alpha} + \boldsymbol{V}_i^\top \boldsymbol{\gamma})\} > 1 \\ \Leftrightarrow &\boldsymbol{w}_i^\top \boldsymbol{\alpha} + \boldsymbol{V}_i^\top \boldsymbol{\gamma} < 0, \end{aligned}$$

o tempo até *dropout* é reduzido pelo efeito de $\boldsymbol{w}_i^\top \boldsymbol{\alpha} + \boldsymbol{V}_i^\top \boldsymbol{\gamma}$. Se, pelo contrário,

$$\begin{aligned} &\exp\{-(\boldsymbol{w}_i^\top \boldsymbol{\alpha} + \boldsymbol{V}_i^\top \boldsymbol{\gamma})\} < 1 \\ \Leftrightarrow &\boldsymbol{w}_i^\top \boldsymbol{\alpha} + \boldsymbol{V}_i^\top \boldsymbol{\gamma} > 0, \end{aligned}$$

o tempo de permanência no estudo é alongado pelo efeito de $\boldsymbol{w}_i^\top \boldsymbol{\alpha} + \boldsymbol{V}_i^\top \boldsymbol{\gamma}$.

Atendendo que a função densidade de referência é

$$f_0(d) = \frac{1}{d\sqrt{2\pi\eta^2}} \exp\left\{-\frac{(\log d)^2}{2\eta^2}\right\},$$

a relação entre a função densidade para qualquer sujeito em termos da função densidade subjacente obtêm-se resolvendo em ordem a p_1 a equação

$$f_i(d) = f_0\left(d \cdot \exp\left\{-\left(\log D_i - \log D_0\right)\right\}\right) \cdot p_1.$$

Obtém-se a relação

$$\begin{aligned}
f_i(d) &= f_0\left(d \cdot \exp\left\{-\left(\log D_i - \log D_0\right)\right\}\right) \times \exp\left\{-\left(\log D_i - \log D_0\right)\right\} \\
&\quad \times \exp\left\{\frac{1}{2}\left[\frac{\left(\log d - \left(\log D_i - \log D_0\right)\right)^2}{Var[\log D_0]} - \frac{\left(\log d - E[\log D_i]\right)^2}{Var[\log D_i]}\right]\right\} \times \sqrt{\frac{Var[\log D_0]}{Var[\log D_i]}} \\
&= f_0\left(d \cdot \exp\left\{-(\boldsymbol{w}_i^\top \boldsymbol{\alpha} + \boldsymbol{V}_i^\top \boldsymbol{\gamma})\right\}\right) \times \exp\left\{-(\boldsymbol{w}_i^\top \boldsymbol{\alpha} + \boldsymbol{V}_i^\top \boldsymbol{\gamma})\right\} \\
&\quad \times \exp\left\{\frac{1}{2}\left[\frac{\left(\log d - (\boldsymbol{w}_i^\top \boldsymbol{\alpha} + \boldsymbol{V}_i^\top \boldsymbol{\gamma})\right)^2}{Var[\log D_0]} - \frac{\left(\log d - E[\log D_i]\right)^2}{Var[\log D_i]}\right]\right\} \times \sqrt{\frac{Var[\log D_0]}{Var[\log D_i]}}
\end{aligned} \tag{4.5}$$

Usando o facto que $h(d) = f(d)/S(d)$ a função de risco pode ser escrita como

$$\begin{aligned}
h_i(d) &= \frac{f_i(d)}{S_i(d)} \\
&= \frac{f_0\left(d \cdot \exp\left\{-(\boldsymbol{w}_i^\top \boldsymbol{\alpha} + \boldsymbol{V}_i^\top \boldsymbol{\gamma})\right\}\right)}{S_0(d \cdot \exp\{-(\boldsymbol{w}_i^\top \boldsymbol{\alpha} + \boldsymbol{V}_i^\top \boldsymbol{\gamma})\})} \times \exp\left\{-(\boldsymbol{w}_i^\top \boldsymbol{\alpha} + \boldsymbol{V}_i^\top \boldsymbol{\gamma})\right\} \\
&\quad \times \exp\left\{\frac{1}{2}\left[\frac{\left(\log d - (\boldsymbol{w}_i^\top \boldsymbol{\alpha} + \boldsymbol{V}_i^\top \boldsymbol{\gamma})\right)^2}{Var[\log D_0]} - \frac{\left(\log d - E[\log D_i]\right)^2}{Var[\log D_i]}\right]\right\} \times \sqrt{\frac{Var[\log D_0]}{Var[\log D_i]}}, \\
&= h_0\left(d \cdot \exp\left\{-(\boldsymbol{w}_i^\top \boldsymbol{\alpha} + \boldsymbol{V}_i^\top \boldsymbol{\gamma})\right\}\right) \times \exp\left\{-(\boldsymbol{w}_i^\top \boldsymbol{\alpha} + \boldsymbol{V}_i^\top \boldsymbol{\gamma})\right\} \\
&\quad \times \exp\left\{\frac{1}{2}\left[\frac{\left(\log d - (\boldsymbol{w}_i^\top \boldsymbol{\alpha} + \boldsymbol{V}_i^\top \boldsymbol{\gamma})\right)^2}{Var[\log D_0]} - \frac{\left(\log d - E[\log D_i]\right)^2}{Var[\log D_i]}\right]\right\} \times \sqrt{\frac{Var[\log D_0]}{Var[\log D_i]}}
\end{aligned} \tag{4.6}$$

onde $h_0(d)$ é a função de risco subjacente.

Ao contrário da função de sobrevivência (4.4), a relação entre funções de densidade (4.5) e de risco (4.6) e as respetivas funções subjacentes não é tão simples.

4.2 Inferência

Das estruturas de correlação propostas na Tabela 4.2, consideraremos o caso particular

$$\begin{cases} \boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + U_i + \boldsymbol{Z}_i \\ \log D_i = \boldsymbol{w}_i^\top \boldsymbol{\alpha} + \gamma\, U_i + \epsilon_i \end{cases} \tag{4.7}$$

onde

$$(Z_{ij}, U_i, \epsilon_i) \sim MVN\left(\mathbf{0}, \begin{bmatrix} \tau^2 & 0 & 0 \\ & \nu^2 & 0 \\ & & \eta^2 \end{bmatrix}\right). \tag{4.8}$$

Nestas condições, o modelo (4.2) toma a forma

$$\begin{bmatrix} \boldsymbol{Y}_i \\ \log D_i \end{bmatrix} \sim MVN\left(\begin{bmatrix} \boldsymbol{X}_i\boldsymbol{\beta} \\ \boldsymbol{w}_i^\top\boldsymbol{\alpha} \end{bmatrix}, \begin{bmatrix} \boldsymbol{V}_i(\nu^2,\tau^2) & \boldsymbol{g}_i(\nu^2) \\ \boldsymbol{g}_i^\top(\nu^2) & \gamma\,\nu^2 + \eta^2 \end{bmatrix}\right). \tag{4.9}$$

Atendendo que

$$\begin{aligned} Var(Y_{ij}) &= Var(\boldsymbol{X}_{ij}\boldsymbol{\beta} + U_i + Z_{ij}) \\ &= 0 + Var(U_i) + Var(Z_{ij}) \\ &= \nu^2 + \tau^2, \end{aligned}$$

e

$$\begin{aligned} Cov(Y_{ij}, Y_{ik}) &= Cov(\boldsymbol{X}_{ij}\boldsymbol{\beta} + U_i + Z_{ij}, \boldsymbol{X}_{ik}\boldsymbol{\beta} + U_i + Z_{ik}) \\ &= 0 + Cov(U_i, U_i) \\ &= Var(U_i) \\ &= \nu^2, \end{aligned}$$

a matriz de (co)variâncias da resposta longitudinal $\boldsymbol{Y}_i$ é dada por

$$\boldsymbol{V}_i(\nu^2,\tau^2) = \begin{bmatrix} (\nu^2+\tau^2) & \nu^2 & \cdots & \nu^2 \\ & (\nu^2+\tau^2) & \cdots & \nu^2 \\ & & \ddots & \vdots \\ & & & (\nu^2+\tau^2) \end{bmatrix} = \nu^2\,\boldsymbol{J} + \tau^2\,\boldsymbol{I}.$$

A variância do tempo-até-*dropout* transformado $\log D_i$ obtém-se fazendo

$$\begin{aligned} Var(\log D_i) &= Var(\boldsymbol{w}_i^\top\boldsymbol{\alpha} + \gamma\,U_i + \epsilon_i) \\ &= 0 + Var(\gamma\,U_i) + Var(\epsilon_i) \\ &= \gamma^2\,Var(U_i) + Var(\epsilon_i) \\ &= \gamma^2\,\nu^2 + \eta^2 \end{aligned}.$$

Como

$$\begin{aligned} Cov(Y_{ij}, \log D_i) &= Cov(\boldsymbol{X}_{ij}\boldsymbol{\beta} + U_i + Z_{ij}, \boldsymbol{w}_i^\top\boldsymbol{\alpha} + \gamma\,U_i + \epsilon_i) \\ &= 0 + Cov(U_i, \gamma\,U_i) \\ &= \gamma\,Var(U_i) \\ &= \gamma\,\nu^2, \end{aligned}$$

o vetor de covariância cruzada entre a resposta longitudinal $\boldsymbol{Y}_i$ e $\log D_i$ é

$$\boldsymbol{g}_i(\nu^2) = \left(\gamma\,\nu^2, \ldots, \gamma\,\nu^2\right)^\top = \gamma\,\nu^2\mathbf{1}.$$

Pelos resultados anteriores, a matriz de (co)variâncias em (4.9) tem a forma

$$\begin{bmatrix} \boldsymbol{V}_i(\nu^2,\tau^2) & \boldsymbol{g}_i(\nu^2) \\ \boldsymbol{g}_i^\top(\nu^2) & \gamma\,\nu^2+\eta^2 \end{bmatrix} = \begin{bmatrix} \nu^2\,\boldsymbol{J} + \tau^2\,\boldsymbol{I} & \gamma\,\nu^2\mathbf{1} \\ \gamma\,\nu^2\mathbf{1}^\top & \gamma\,\nu^2+\eta^2 \end{bmatrix}.$$

4.2.1 Estimação por máxima verosimilhança

A função verosimilhança dos parâmetros $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \gamma, \nu^2, \tau^2, \eta^2)$, dado os dados observados $(\boldsymbol{y}^o, \log \boldsymbol{t}, \boldsymbol{\delta})$ do modelo (4.7), é apresentada abaixo. É importante notar no desenvolvimento que t_i é a versão observável de d_i ou c_i como descrito em (4.1), e discriminado pelo indicador δ_i.

$$
\begin{aligned}
L(\boldsymbol{\theta}; \boldsymbol{y}^o, \log \boldsymbol{t}, \boldsymbol{\delta}) &= \prod_{i=1}^{m} \int [(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m), \log t_i, \delta_i]_{\boldsymbol{Y}_i, \log D_i, \Delta_i} \, d\boldsymbol{y}_i^m \\
&= \prod_{i=1}^{m} \int [\boldsymbol{y}_i^o, \log t_i, \delta_i \mid \boldsymbol{y}_i^m]_{\boldsymbol{Y}_i, \log D_i, \Delta_i \mid \boldsymbol{Y}_i^m} \times [\boldsymbol{y}_i^m]_{\boldsymbol{Y}_i} \, d\boldsymbol{y}_i^m \\
&= \prod_{i=1}^{m} [\boldsymbol{y}_i^o, \log t_i, \delta_i]_{\boldsymbol{Y}_i, \log D_i, \Delta_i} \\
&= \prod_{i=1}^{m} [\boldsymbol{y}_i^o]_{\boldsymbol{Y}_i} \times [\log t_i, \delta_i \mid \boldsymbol{y}_i^o]_{\log D_i, \Delta_i \mid \boldsymbol{Y}_i} \\
&= \prod_{i=1}^{m} f_{\boldsymbol{Y}_i}(\boldsymbol{y}_i^o) \Big\{ f_{\log D_i \mid \boldsymbol{Y}_i}\big(\log t_i\big) S_{C_i \mid \boldsymbol{Y}_i}\big(\log t_i\big) \Big\}^{\delta_i} \Big\{ f_{C_i \mid \boldsymbol{Y}_i}\big(\log t_i\big) S_{\log D_i \mid \boldsymbol{Y}_i}\big(\log t_i\big) \Big\}^{1-\delta_i} \\
&= \prod_{i=1}^{m} f_{\boldsymbol{Y}_i}(\boldsymbol{y}_i^o) \Big\{ f_{\log D_i \mid \boldsymbol{Y}_i}\big(\log t_i\big) \Big\}^{\delta_i} \Big\{ S_{\log D_i \mid \boldsymbol{Y}_i}\big(\log t_i\big) \Big\}^{1-\delta_i} \\
&\quad \times \prod_{i=1}^{m} \Big\{ S_{C_i \mid \boldsymbol{Y}_i}\big(\log t_i\big) \Big\}^{\delta_i} \Big\{ f_{C_i \mid \boldsymbol{Y}_i}\big(\log t_i\big) \Big\}^{1-\delta_i}
\end{aligned}
\tag{4.10}
$$

onde δ_i é o indicador de *dropout*

$$
\delta_i = \begin{cases} 1 & \text{se } \textit{dropout} \text{ foi observado} \\ 0 & \text{se } \textit{dropout} \text{ foi censurado} \end{cases}. \tag{4.11}
$$

Se assumirmos que o processo de censura é independente quer do processo de *dropout* quer do processo longitudinal (censura não informativa), a distribuição do tempo de censura C_i não depende dos parâmetros de interesse $\boldsymbol{\theta}$. Assim, o último produtório na equação (4.10) é constante, e toda a inferência sobre $\boldsymbol{\theta}$ pode ser baseada na função de verosimilhança dada por

$$L(\boldsymbol{\theta}; \boldsymbol{y}^o, \log \boldsymbol{t}, \boldsymbol{\delta}) \propto \prod_{i=1}^{m} f_{\boldsymbol{Y}_i}(\boldsymbol{y}_i^o) \times \left\{ f_{\log D_i | \boldsymbol{Y}_i}\big(\log t_i\big) \right\}^{\delta_i} \left\{ S_{\log D_i | \boldsymbol{Y}_i}\big(\log t_i\big) \right\}^{1-\delta_i} \tag{4.12}$$

A função de verosimilhança (4.12) envolve a distribuição marginal de $\boldsymbol{Y}_i$ e a distribuição condicional de $\log D_i \mid \boldsymbol{Y}_i$. A distinção entre a contribuição de um indivíduo que sofreu ou não *dropout* faz-se através da distribuição condicional $\log D_i \mid \boldsymbol{Y}_i$. Quando o tempo de *dropout* é observado, o indivíduo contribui com densidade de probabilidade condicional; quando é censurado, i.e. o tempo de *dropout* não é observado, ele contribui com a probabilidade condicional acumulada.

Por (4.2) e (4.7) conseguimos ver que

$$\boldsymbol{Y}_i \sim MVN\left(\boldsymbol{X}_i\boldsymbol{\beta},\ \boldsymbol{V}_i(\nu^2, \tau^2)\right),$$

e

$$\log D_i \mid \boldsymbol{Y}_i = \boldsymbol{y}_i \sim N\left(\mu_{\log D_i | \boldsymbol{y}_i}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \nu^2, \tau^2),\ \sigma^2_{\log D_i | \boldsymbol{y}_i}(\nu^2, \tau^2, \eta^2)\right), \tag{4.13}$$

onde

$$\begin{aligned} \mu_{\log D_i | \boldsymbol{y}_i}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \nu^2, \tau^2) &= \boldsymbol{w}_i^\top \boldsymbol{\alpha} + \boldsymbol{g}_i^\top(\nu^2)\boldsymbol{V}_i^{-1}(\nu^2, \tau^2)(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \\ &= \boldsymbol{w}_i^\top \boldsymbol{\alpha} + \gamma\,\nu^2 \mathbf{1}^\top \left(\nu^2\,\boldsymbol{J} + \tau^2\,\boldsymbol{I}\right)^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}), \end{aligned} \tag{4.14}$$

e, pelas propriedades da distribuição multivariada Gaussiana condicional descritas no anexo C.1 (p. 133), que

$$\begin{aligned} \sigma^2_{\log D_i | \boldsymbol{y}_i}(\nu^2, \tau^2, \eta^2) &= \left(\gamma\,\nu^2 + \eta^2\right) - \boldsymbol{g}_i^\top(\nu^2)\boldsymbol{V}_i^{-1}(\nu^2, \tau^2)\boldsymbol{g}_i(\nu^2). \\ &= \left(\gamma\,\nu^2 + \eta^2\right) - \left(\gamma\,\nu^2\right)^2 \mathbf{1}^\top \left(\nu^2\,\boldsymbol{J} + \tau^2\,\boldsymbol{I}\right)^{-1}\mathbf{1} \end{aligned} \tag{4.15}$$

Substituindo as respetivas funções de densidade de probabilidade e de distribuição acumulada em (4.12), a função de verosimilhança é escrita como

$$L(\boldsymbol{\theta};\boldsymbol{y}^o,\log\boldsymbol{t},\boldsymbol{\delta}) \propto \prod_{i=1}^{m} \frac{\exp\left\{-\frac{1}{2}(\boldsymbol{y}_i^o-\boldsymbol{X}_i\boldsymbol{\beta})^\top \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i^o-\boldsymbol{X}_i\boldsymbol{\beta})\right\}}{(2\pi)^{n_i/2} \mid \boldsymbol{V}_i \mid^{1/2}} \times \left\{\frac{\exp\left\{-\frac{1}{2}u_i(\log t_i)^2\right\}}{(2\pi\sigma^2_{\log D_i|\boldsymbol{y}_i})^{1/2}}\right\}^{\delta_i} \left\{1-\Phi(u_i(\log t_i))\right\}^{1-\delta_i} \tag{4.16}$$

onde

$$u_i(\log t_i) = \frac{\log t_i - \mu_{\log D_i|\boldsymbol{y}_i}}{\sigma_{\log D_i|\boldsymbol{y}_i}},$$

e $\Phi(\cdot)$ a função de distribuição acumulada da distribuição Gaussiana padronizada, $N(0,1)$. O logaritmo da função de verosimilhança é então

$$\begin{aligned} l(\boldsymbol{\theta};\boldsymbol{y}^o,\log\boldsymbol{c},\boldsymbol{\delta}) \propto & \sum_{i=1}^{m}\Bigg\{ \log\left(\frac{\exp\left\{-\frac{1}{2}(\boldsymbol{y}_i^o-\boldsymbol{X}_i\boldsymbol{\beta})^\top \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i^o-\boldsymbol{X}_i\boldsymbol{\beta})\right\}}{(2\pi)^{n_i/2} \mid \boldsymbol{V}_i \mid^{1/2}}\right) \\ & + \delta_i \log\left(\frac{\exp\left\{-\frac{1}{2}u_i(\log t_i)^2\right\}}{(2\pi\sigma^2_{\log D_i|\boldsymbol{y}_i})^{1/2}}\right) \\ & + (1-\delta_i)\log\Big(1-\Phi(u_i(\log t_i))\Big)\Bigg\} \\ = & \sum_{i=1}^{m}\Bigg\{ -\frac{n_i}{2}\log(2\pi) - \frac{1}{2}\log(\mid \boldsymbol{V}_i \mid) \\ & - \frac{1}{2}(\boldsymbol{y}_i^o-\boldsymbol{X}_i\boldsymbol{\beta})^\top \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i^o-\boldsymbol{X}_i\boldsymbol{\beta}) \\ & + \delta_i\Big(-\frac{1}{2}\log(2\pi) - \log(\sigma_{\log D_i|\boldsymbol{y}_i}) - \frac{1}{2}u_i(\log t_i)^2\Big) \\ & + (1-\delta_i)\log\Big(1-\Phi(u_i(\log t_i)\Big)\Bigg\} \\ \propto & \sum_{i=1}^{m}\Bigg\{ -\frac{1}{2}\log(\mid \boldsymbol{V}_i \mid) - \frac{1}{2}(\boldsymbol{y}_i^o-\boldsymbol{X}_i\boldsymbol{\beta})^\top \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i^o-\boldsymbol{X}_i\boldsymbol{\beta}) \\ & + \delta_i\Big(-\frac{1}{2}\log(\sigma^2_{\log D_i|\boldsymbol{y}_i}) - \frac{1}{2}u_i(\log t_i)^2\Big) \\ & + (1-\delta_i)\log\Big(1-\Phi(u_i(\log t_i)\Big)\Bigg\} \end{aligned} \tag{4.17}$$

Analisando a contribuição do indivíduo i em (4.17), o primeiro termo refere-se à contribuição da distribuição Gaussiana multivariada dos dados longitudinais. O segundo e terceiro termo correspondem à contribuição do tempo-até-*dropout* ou tempo de censura, respetivamente. Um atributo desejável para um modelo conjunto é que, na ausência de associação entre os dados longitudinais e os tempos dos eventos, o modelo deve gerar os mesmos resultados obtidos pelas análises independentes de cada uma das componentes. Se a associação entre as medidas longitudinais e tempo até *dropout* não for estatisticamente significativo, i.e. $\gamma = 0$, os parâmetros da distribuição condicional $D_i \mid \boldsymbol{Y}_i$, (4.14) e (4.15) tornam-se

$$\mu_{\log D_i|\boldsymbol{y}_i} = \boldsymbol{w}_i^\top \boldsymbol{\alpha} = \mu_{\log D_i},$$

e

$$\sigma^2_{\log D_i|\boldsymbol{y}_i} = \eta^2 = \sigma^2_{\log D_i}.$$

A log-verosimilhança (4.17) reduz-se à soma das funções de log-verosimilhança que seriam obtidas pelas análises individuais dos dados longitudinais e tempo-até-evento. Ou seja, nessas circunstâncias, maximizar a função de log-verosimilhança é equivalente a maximizar as duas funções que a compõem separadamente.

As estimativas dos parâmetros de interesse $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \tau^2, \gamma, \eta^2, \nu^2)$ podem ser obtidas através da diferenciação da função log-verosimilhança (4.17). Apresentam-se abaixo os cálculos da sua derivada com respeito a cada um dos parâmetros Os cálculos fazem uso de um conjunto de propriedades de operações sobre matrizes e da fórmula de Leibniz descritas nos anexos C.2 e C.3 (p. 135 e 136), respetivamente.

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}} &= 0 + \frac{\partial}{\partial \boldsymbol{\alpha}} \sum_{i=1}^{m} \left\{ \delta_i \left(-\frac{1}{2} u_i (\log t_i)^2 \right) + (1-\delta_i) \log \left(1 - \Phi(u_i(\log t_i) \right) \right\} \\
&= \sum_{i=1}^{m} \left\{ \delta_i \left(\frac{u_i(\log t_i)}{\sigma_{\log D_i|\boldsymbol{y}_i}} \boldsymbol{w}_i \right) + (1-\delta_i) \left(\frac{\varphi(u_i(\log t_i))}{1 - \Phi(u_i(\log t_i))} \frac{\boldsymbol{w}_i}{\sigma_{\log D_i|\boldsymbol{y}_i}} \right) \right\} \\
&= \sum_{i=1}^{m} \boldsymbol{w}_i \left\{ \delta_i \left(\frac{u_i(\log t_i)}{\sigma_{\log D_i|\boldsymbol{y}_i}} \right) + (1-\delta_i) \left(\frac{\varphi(u_i(\log t_i))}{1 - \Phi(u_i(\log t_i))} \frac{1}{\sigma_{\log D_i|\boldsymbol{y}_i}} \right) \right\},
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} &= 0 + \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{i=1}^{m} \Bigg\{ -\frac{1}{2} (\boldsymbol{y}_i^o - \boldsymbol{X}_i \boldsymbol{\beta})^\top \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i^o - \boldsymbol{X}_i \boldsymbol{\beta}) \\
&\quad + \delta_i \bigg(-\frac{1}{2} u_i (\log t_i)^2 \bigg) + (1 - \delta_i) \log \bigg(1 - \Phi(u_i(\log t_i) \bigg) \Bigg\} \\
&= \sum_{i=1}^{m} \Bigg\{ \boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i^o - \boldsymbol{X}_i \boldsymbol{\beta}) \\
&\quad + \delta_i \bigg(-\frac{u_i(\log t_i)}{\sigma_{\log D_i | \boldsymbol{y}_i}} \boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1} \boldsymbol{g}_i \bigg) \\
&\quad + (1 - \delta_i) \bigg(-\frac{\varphi(u_i(\log t_i))}{1 - \Phi(u_i(\log t_i))} \frac{1}{\sigma_{\log D_i | \boldsymbol{y}_i}} \boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1} \boldsymbol{g}_i \bigg) \Bigg\} \\
&= \sum_{i=1}^{m} \Bigg\{ \boldsymbol{X}_i^\top (\nu^2 \boldsymbol{J} + \tau^2 \boldsymbol{I})^{-1} (\boldsymbol{y}_i^o - \boldsymbol{X}_i \boldsymbol{\beta}) \\
&\quad + \delta_i \bigg(-\frac{u_i(\log t_i)}{\sigma_{\log D_i | \boldsymbol{y}_i}} \boldsymbol{X}_i^\top (\nu^2 \boldsymbol{J} + \tau^2 \boldsymbol{I})^{-1} \gamma \, \nu^2 \mathbf{1} \bigg) \\
&\quad + (1 - \delta_i) \bigg(-\frac{\varphi(u_i(\log t_i))}{1 - \Phi(u_i(\log t_i))} \frac{1}{\sigma_{\log D_i | \boldsymbol{y}_i}} \boldsymbol{X}_i^\top (\nu^2 \boldsymbol{J} + \tau^2 \boldsymbol{I})^{-1} \gamma \, \nu^2 \mathbf{1} \bigg) \Bigg\},
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial l(\boldsymbol{\theta})}{\partial \eta^2} &= 0 + \frac{\partial}{\partial \eta^2} \sum_{i=1}^{m} \Bigg\{ \delta_i \bigg(-\frac{1}{2} \log(\sigma^2_{\log D_i | \boldsymbol{y}_i}) - \frac{1}{2} u_i (\log t_i)^2 \bigg) \\
&\quad + (1 - \delta_i) \log \bigg(1 - \Phi(u_i(\log t_i) \bigg) \Bigg\} \\
&= \sum_{i=1}^{m} \frac{1}{2} \frac{1}{\sigma^2_{\log D_i | \boldsymbol{y}_i}} \Bigg\{ \delta_i \bigg(u_i (\log t_i)^2 - 1 \bigg) + (1 - \delta_i) u_i (\log t_i) \frac{\varphi(u_i(\log t_i))}{1 - \Phi(u_i(\log t_i))} \Bigg\},
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial l(\boldsymbol{\theta})}{\partial \gamma} &= 0 + \frac{\partial}{\partial \gamma} \sum_{i=1}^{m} \Bigg\{ \delta_i \Bigg(-\frac{1}{2} \log(\sigma^2_{\log D_i|\boldsymbol{y}_i}) - \frac{1}{2} u_i (\log t_i)^2 \Bigg) \\
&\quad + (1-\delta_i) \log \Bigg(1 - \Phi(u_i(\log t_i)) \Bigg) \Bigg\} \\
&= \sum_{i=1}^{m} \frac{\nu^2}{\sigma^2_{\log D_i|\boldsymbol{y}_i}} \Bigg\{ \delta_i \Bigg(u_i(\log t_i) \sigma_{\log D_i|\boldsymbol{y}_i} \mathbf{1}^\top \boldsymbol{V}_i^{-1} \left(\boldsymbol{y}_i^o - \boldsymbol{X}_i \boldsymbol{\beta} \right) \\
&\quad + \gamma \, \left(u_i^2(\log t_i) - 1 \right) \left(1 - \nu^2 \mathbf{1}^\top \boldsymbol{V}_i^{-1} \mathbf{1} \right) \Bigg) \\
&\quad + (1-\delta_i) \frac{\varphi(u_i(\log t_i))}{1 - \Phi(u_i(\log t_i))} \Bigg(\sigma_{\log D_i|\boldsymbol{y}_i} \mathbf{1}^\top \boldsymbol{V}_i^{-1} \left(\boldsymbol{y}_i^o - \boldsymbol{X}_i \boldsymbol{\beta} \right) \\
&\quad + \gamma \, u_i(\log t_i) \left(1 - \nu^2 \mathbf{1}^\top \boldsymbol{V}_i^{-1} \mathbf{1} \right) \Bigg) \Bigg\} \\
&= \sum_{i=1}^{m} \frac{\nu^2}{\sigma^2_{\log D_i|\boldsymbol{y}_i}} \Bigg\{ \delta_i \Bigg(u_i(\log t_i) \sigma_{\log D_i|\boldsymbol{y}_i} \mathbf{1}^\top (\nu^2 \boldsymbol{J} + \tau^2 \boldsymbol{I})^{-1} \left(\boldsymbol{y}_i^o - \boldsymbol{X}_i \boldsymbol{\beta} \right) \\
&\quad + \gamma \, \left(u_i^2(\log t_i) - 1 \right) \left(1 - \nu^2 \mathbf{1}^\top (\nu^2 \boldsymbol{J} + \tau^2 \boldsymbol{I})^{-1} \mathbf{1} \right) \Bigg) \\
&\quad + \ (1-\delta_i) \frac{\varphi(u_i(\log t_i))}{1 - \Phi(u_i(\log t_i))} \Bigg(\sigma_{\log D_i|\boldsymbol{y}_i} \mathbf{1}^\top (\nu^2 \boldsymbol{J} + \tau^2 \boldsymbol{I})^{-1} \left(\boldsymbol{y}_i^o - \boldsymbol{X}_i \boldsymbol{\beta} \right) \\
&\quad + \ \gamma \, u_i(\log t_i) \left(1 - \nu^2 \mathbf{1}^\top (\nu^2 \boldsymbol{J} + \tau^2 \boldsymbol{I})^{-1} \mathbf{1} \right) \Bigg) \Bigg\}.
\end{aligned}
$$

Notando que

$$
\begin{aligned}
\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{V}_i} &= \frac{\partial}{\partial \boldsymbol{V}_i}\sum_{i=1}^{m}\Bigg\{ -\frac{1}{2}\log\left(\mid \boldsymbol{V}_i \mid\right) - \frac{1}{2}(\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta})^\top \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta}) \\
&\quad + \delta_i\bigg(-\frac{1}{2}\log(\sigma^2_{\log D_i|\boldsymbol{y}_i}) - \frac{1}{2}u_i(\log t_i)^2\bigg) + (1-\delta_i)\log\bigg(1-\Phi(u_i(\log t_i)\bigg)\Bigg\} \\
&= \sum_{i=1}^{m}\boldsymbol{V}_i^{-1}\Bigg\{\frac{1}{2}\left((\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta})(\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta})^\top\boldsymbol{V}_i^{-1} - \boldsymbol{I}\right) \\
&\quad - \boldsymbol{g}_i\Bigg[\delta_i\bigg((\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta})^\top\frac{u_i(\log t_i)}{\sigma_{\log D_i|\boldsymbol{y}_i}} - \frac{1}{2}\frac{(u_i^2(\log t_i)-1)}{\sigma^2_{\log D_i|\boldsymbol{y}_i}}\boldsymbol{g}_i^\top\bigg) \\
&\quad + (1-\delta_i)\frac{\varphi(u_i(\log t_i))}{1-\Phi(u_i(\log t_i))}\frac{1}{\sigma_{\log D_i|\boldsymbol{y}_i}}\bigg((\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta})^\top - \frac{1}{2}\frac{u_i(\log t_i)}{\sigma_{\log D_i|\boldsymbol{y}_i}}\boldsymbol{g}_i^\top\bigg)\Bigg]\boldsymbol{V}_i^{-1}\Bigg\} \\
&= \sum_{i=1}^{m}\left(\nu^2\boldsymbol{J} + \tau^2\boldsymbol{I}\right)^{-1}\Bigg\{\frac{1}{2}\left((\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta})(\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta})^\top\left(\nu^2\boldsymbol{J} + \tau^2\boldsymbol{I}\right)^{-1} - \boldsymbol{I}\right) \\
&\quad - \gamma\,\nu^2\mathbf{1}\Bigg[\delta_i\bigg((\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta})^\top\frac{u_i(\log t_i)}{\sigma_{\log D_i|\boldsymbol{y}_i}} - \frac{1}{2}\frac{(u_i^2(\log t_i)-1)}{\sigma^2_{\log D_i|\boldsymbol{y}_i}}\gamma\,\nu^2\mathbf{1}^\top\bigg) \\
&\quad + (1-\delta_i)\frac{\varphi(u_i(\log t_i))}{1-\Phi(u_i(\log t_i))}\frac{1}{\sigma_{\log D_i|\boldsymbol{y}_i}}\bigg((\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta})^\top - \frac{1}{2}\frac{u_i(\log t_i)}{\sigma_{\log D_i|\boldsymbol{y}_i}}\gamma\,\nu^2\mathbf{1}^\top\bigg)\Bigg] \\
&\quad \times\left(\nu^2\boldsymbol{J} + \tau^2\boldsymbol{I}\right)^{-1}\Bigg\}.
\end{aligned}
$$

a derivada da função log-verosimilhança em função dos parâmetros τ^2 e ν^2 é dada por

$$
\begin{aligned}
\frac{\partial l(\boldsymbol{\theta})}{\partial \tau^2} &= \frac{\partial}{\partial \tau^2}\sum_{i=1}^{m}\Bigg\{ -\frac{1}{2}\log(\mid \boldsymbol{V}_i \mid) - \frac{1}{2}(\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta})^\top \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta}) \\
&\quad + \delta_i\bigg(-\frac{1}{2}\log(\sigma^2_{\log D_i|\boldsymbol{y}_i}) - \frac{1}{2}u_i(\log t_i)^2\bigg) + (1-\delta_i)\log\bigg(1-\Phi(u_i(\log t_i)\bigg)\Bigg\} \\
&= \sum_{i=1}^{m}\mathrm{Tr}\Bigg[\boldsymbol{V}_i^{-1}\Bigg\{\frac{1}{2}\left((\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta})(\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta})^\top\boldsymbol{V}_i^{-1} - \boldsymbol{I}\right) \\
&\quad - \Bigg[\delta_i\bigg((\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta})\frac{u_i(\log t_i)}{\sigma_{\log D_i|\boldsymbol{y}_i}} - \frac{1}{2}\frac{(u_i^2(\log t_i)-1)}{\sigma^2_{\log D_i|\boldsymbol{y}_i}}\boldsymbol{g}_i\bigg) \\
&\quad + (1-\delta_i)\frac{\varphi(u_i(\log t_i))}{1-\Phi(u_i(\log t_i))}\frac{1}{\sigma_{\log D_i|\boldsymbol{y}_i}}\bigg((\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta}) - \frac{1}{2}\frac{u_i(\log t_i)}{\sigma_{\log D_i|\boldsymbol{y}_i}}\boldsymbol{g}_i\bigg)\Bigg]\boldsymbol{g}_i^\top\boldsymbol{V}_i^{-1}\Bigg\}\Bigg] \\
&= \sum_{i=1}^{m}\mathrm{Tr}\Bigg[\left(\nu^2\boldsymbol{J} + \tau^2\boldsymbol{I}\right)^{-1}\Bigg\{\frac{1}{2}\left((\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta})(\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta})^\top\left(\nu^2\boldsymbol{J} + \tau^2\boldsymbol{I}\right)^{-1} - \boldsymbol{I}\right) \\
&\quad - \Bigg[\delta_i\bigg((\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta})\frac{u_i(\log t_i)}{\sigma_{\log D_i|\boldsymbol{y}_i}} - \frac{1}{2}\frac{(u_i^2(\log t_i)-1)}{\sigma^2_{\log D_i|\boldsymbol{y}_i}}\gamma\,\nu^2\mathbf{1}\bigg) \\
&\quad + (1-\delta_i)\frac{\varphi(u_i(\log t_i))}{1-\Phi(u_i(\log t_i))}\frac{1}{\sigma_{\log D_i|\boldsymbol{y}_i}}\bigg((\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta}) - \frac{1}{2}\frac{u_i(\log t_i)}{\sigma_{\log D_i|\boldsymbol{y}_i}}\gamma\,\nu^2\mathbf{1}\bigg)\Bigg] \\
&\quad \times \gamma\,\nu^2\mathbf{1}^\top\left(\nu^2\boldsymbol{J} + \tau^2\boldsymbol{I}\right)^{-1}\Bigg\}\Bigg],
\end{aligned}
$$

e

$$
\begin{aligned}
\frac{\partial l(\boldsymbol{\theta})}{\partial \nu^2} &= \frac{\partial}{\partial \nu^2}\sum_{i=1}^{m}\Bigg\{ -\frac{1}{2}\log\left(\mid \boldsymbol{V}_i \mid\right) - \frac{1}{2}(\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta})^\top \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta}) \\
&\quad + \delta_i\Big(-\frac{1}{2}\log(\sigma^2_{\log D_i|\boldsymbol{y}_i}) - \frac{1}{2}u_i(\log t_i)^2\Big) + (1-\delta_i)\log\Big(1-\Phi(u_i(\log t_i)\Big)\Bigg\} \\
&= \sum_{i=1}^{m}\Bigg\{ -\frac{1}{2}\mathrm{Tr}[\boldsymbol{V}_i^{-1}\boldsymbol{J}] + \frac{1}{2}\mathrm{Tr}[\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta})(\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta})^\top \boldsymbol{V}_i^{-1}\boldsymbol{J}] \\
&\quad + \delta_i \frac{1}{2}\frac{\gamma}{\sigma^2_{\log D_i|\boldsymbol{y}_i}}\Bigg((\gamma-1)\Big\{\mathbf{1}^\top\boldsymbol{V}_i^{-1}\mathbf{1} - \nu^2\mathrm{Tr}[\boldsymbol{V}_i^{-1}\boldsymbol{J}\boldsymbol{V}_i^{-1}\boldsymbol{J}]\Big\} \\
&\quad + 2u_i(\log t_i)\Big\{\sigma^2_{\log D_i|\boldsymbol{y}_i}\left(\mathbf{1}^\top\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta}) - \nu^2\mathrm{Tr}\left[\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta})\mathbf{1}^\top\boldsymbol{V}_i^{-1}\boldsymbol{J}\right]\right) \\
&\quad + \frac{1}{2}u_i(\log t_i)(1-\gamma)\left(\mathbf{1}^\top\boldsymbol{V}_i^{-1}\mathbf{1} - \nu^2\mathrm{Tr}\left[\boldsymbol{V}_i^{-1}\boldsymbol{J}\boldsymbol{V}_i^{-1}\boldsymbol{J}\right]\right)\Big\}\Bigg) \\
&\quad - (1-\delta_i)\frac{\varphi(u_i(\log t_i))}{1-\Phi(u_i(\log t_i))}\frac{\gamma}{\sigma^2_{\log D_i|\boldsymbol{y}_i}}\Bigg(\sigma_{\log D_i|\boldsymbol{y}_i}\left(\mathbf{1}^\top\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta})\right. \\
&\quad \left. - \nu^2\mathrm{Tr}\left[\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta})\mathbf{1}^\top\boldsymbol{V}_i^{-1}\boldsymbol{J}\right]\right) \\
&\quad + \frac{1}{2}u_i(\log t_i)(\gamma-1)\left(\mathbf{1}^\top\boldsymbol{V}_i^{-1}\mathbf{1} - \nu^2\mathrm{Tr}\left[\boldsymbol{V}_i^{-1}\boldsymbol{J}\boldsymbol{V}_i^{-1}\boldsymbol{J}\right]\right)\Bigg)\Bigg\}
\end{aligned}
$$

(Continua na página seguinte.)

$$
\begin{aligned}
&= \sum_{i=1}^{m} \Bigg\{ -\frac{1}{2}\mathrm{Tr}[\left(\nu^2\,\boldsymbol{J}+\tau^2\,\boldsymbol{I}\right)^{-1}\boldsymbol{J}] \\
&+\frac{1}{2}\mathrm{Tr}[\left(\nu^2\,\boldsymbol{J}+\tau^2\,\boldsymbol{I}\right)^{-1}(\boldsymbol{y}_i^o-\boldsymbol{X}_i\boldsymbol{\beta})(\boldsymbol{y}_i^o-\boldsymbol{X}_i\boldsymbol{\beta})^\top\left(\nu^2\,\boldsymbol{J}+\tau^2\,\boldsymbol{I}\right)^{-1}\boldsymbol{J}] \\
&+\delta_i\frac{1}{2}\frac{\gamma}{\sigma^2_{\log D_i|\boldsymbol{y}_i}}\Bigg((\gamma-1)\Big\{\mathbf{1}^\top\left(\nu^2\,\boldsymbol{J}+\tau^2\,\boldsymbol{I}\right)^{-1}\mathbf{1} \\
&-\nu^2\mathrm{Tr}[\left(\nu^2\,\boldsymbol{J}+\tau^2\,\boldsymbol{I}\right)^{-1}\boldsymbol{J}\left(\nu^2\,\boldsymbol{J}+\tau^2\,\boldsymbol{I}\right)^{-1}\boldsymbol{J}]\Big\} \\
&+2u_i(\log t_i)\Big\{\sigma^2_{\log D_i|\boldsymbol{y}_i}\big(\mathbf{1}^\top\left(\nu^2\,\boldsymbol{J}+\tau^2\,\boldsymbol{I}\right)^{-1}(\boldsymbol{y}_i^o-\boldsymbol{X}_i\boldsymbol{\beta}) \\
&-\nu^2\mathrm{Tr}\left[\left(\nu^2\,\boldsymbol{J}+\tau^2\,\boldsymbol{I}\right)^{-1}(\boldsymbol{y}_i^o-\boldsymbol{X}_i\boldsymbol{\beta})\mathbf{1}^\top\left(\nu^2\,\boldsymbol{J}+\tau^2\,\boldsymbol{I}\right)^{-1}\boldsymbol{J}\right]\big) \\
&+\frac{1}{2}u_i(\log t_i)(1-\gamma)\big(\mathbf{1}^\top\left(\nu^2\,\boldsymbol{J}+\tau^2\,\boldsymbol{I}\right)^{-1}\mathbf{1} \\
&-\nu^2\mathrm{Tr}\left[\left(\nu^2\,\boldsymbol{J}+\tau^2\,\boldsymbol{I}\right)^{-1}\boldsymbol{J}\left(\nu^2\,\boldsymbol{J}+\tau^2\,\boldsymbol{I}\right)^{-1}\boldsymbol{J}\right]\big)\Big\}\Bigg) \\
&-(1-\delta_i)\frac{\varphi(u_i(\log t_i))}{1-\Phi(u_i(\log t_i))}\frac{\gamma}{\sigma^2_{\log D_i|\boldsymbol{y}_i}}\Bigg(\sigma_{\log D_i|\boldsymbol{y}_i}\big(\mathbf{1}^\top\left(\nu^2\,\boldsymbol{J}+\tau^2\,\boldsymbol{I}\right)^{-1}(\boldsymbol{y}_i^o-\boldsymbol{X}_i\boldsymbol{\beta}) \\
&-\nu^2\mathrm{Tr}\left[\left(\nu^2\,\boldsymbol{J}+\tau^2\,\boldsymbol{I}\right)^{-1}(\boldsymbol{y}_i^o-\boldsymbol{X}_i\boldsymbol{\beta})\mathbf{1}^\top\left(\nu^2\,\boldsymbol{J}+\tau^2\,\boldsymbol{I}\right)^{-1}\boldsymbol{J}\right]\big) \\
&+\frac{1}{2}u_i(\log t_i)(\gamma-1)\big(\mathbf{1}^\top\left(\nu^2\,\boldsymbol{J}+\tau^2\,\boldsymbol{I}\right)^{-1}\mathbf{1} \\
&-\nu^2\mathrm{Tr}\left[\left(\nu^2\,\boldsymbol{J}+\tau^2\,\boldsymbol{I}\right)^{-1}\boldsymbol{J}\left(\nu^2\,\boldsymbol{J}+\tau^2\,\boldsymbol{I}\right)^{-1}\boldsymbol{J}\right]\big)\Bigg)\Bigg\}.
\end{aligned}
$$

Como os resultados acima demonstram, na presença de tempo-até-*dropout* censurados, não é possível obter expressões de forma fechada para nenhum dos parâmetros desconhecidos do modelo.

4.2.2 Estimação por algoritmo EM

A inclusão de efeitos aleatórios partilhados no modelo (4.3) possibilita a aplicação do algoritmo Estimação-Maximização (EM, *Expectation Maximization*), proposto por Dempster et al. (1977) [38], tratando os efeitos aleatórios como dados em falta, para obter as estimativas por máxima verosimilhança do modelo (4.2) [39], como alternativa à diferenciação da log-verosimilhança descrita anteriormente.

O algoritmo EM é um método iterativo aplicado na estimação por máxima verosimilhança em problemas com dados incompletos. A ideia intuitiva deste algoritmo é a de que a log-verosimilhança relativa aos dados completos é tipicamente mais simples de maximizar, geralmente em forma fechada. A cada iteração o algoritmo percorre dois passos: o passo de Esperança (E) e o passo de Maximização (M). No passo E preenchemos os dados incompletos e trocamos a log-verosimilhança dos dados observados por uma função substituta. Esta função é então maximizada no passo M, que simula a estimação por máxima verosimilhança que seria possível na presença dos dados completos [40].

Uma das grandes vantagens do algoritmo é a sua estabilidade numérica. Dempster et. al (1997) [38] demonstrou que que a verosimilhança dos dados observados aumenta a cada iteração do algoritmo, i.e, $\log f_{\boldsymbol{Y}}(\boldsymbol{y}^o;\boldsymbol{\theta}^{it+1}) \geq \log f_{\boldsymbol{Y}}(\boldsymbol{y}^o;\boldsymbol{\theta}^{it+1})$. As iterações iniciais do método aproximam-se rapidamente do ótimo, contudo próximo do ponto máximo a convergência é lenta. A velocidade de convergência do algoritmo é geralmente linear, dependendo dos dados e da estrutura da matriz de (co)variâncias [7]. A restrição de não negatividade para as componentes da variância são automaticamente satisfeitas a cada iteração se os valores iniciais cumprirem essa condição [41].

A função verosimilhança dos parâmetros $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \gamma, \nu^2, \tau^2, \eta^2)$ do modelo (4.7), dado os dados observados $(\boldsymbol{y}^o, \log \boldsymbol{t}, \boldsymbol{\delta})$, é

$$
\begin{aligned}
L(\boldsymbol{\theta}; \boldsymbol{y}^o, \log \boldsymbol{t}, \boldsymbol{\delta}) &= \prod_{i=1}^{m} \int [(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m), \log t_i, \delta_i]_{\boldsymbol{Y}_i, \log T_i, \Delta_i} \, d\boldsymbol{y}_i^m \\
&= \prod_{i=1}^{m} \int [\boldsymbol{y}_i^o, \log t_i, \delta_i \mid \boldsymbol{y}_i^m]_{\boldsymbol{Y}_i, \log T_i, \Delta_i \mid \boldsymbol{Y}_i} \times [\boldsymbol{y}_i^m]_{\boldsymbol{Y}_i} \, d\boldsymbol{y}_i^m \\
&= \prod_{i=1}^{m} [\boldsymbol{y}_i^o, \log t_i, \delta_i]_{\boldsymbol{Y}_i, \log T_i, \Delta_i} \\
&= \prod_{i=1}^{m} \int [\boldsymbol{y}_i^o, \log t_i, \delta_i, \boldsymbol{u}_i]_{\boldsymbol{Y}_i, \log T_i, \Delta_i, \boldsymbol{U}_i} \, d\boldsymbol{u}_i \\
&= \prod_{i=1}^{m} \int [\boldsymbol{u}_i]_{\boldsymbol{U}_i} \times [\boldsymbol{y}_i^o, \log t_i, \delta_i \mid \boldsymbol{u}_i]_{\boldsymbol{Y}_i, \log T_i, \Delta_i \mid \boldsymbol{U}_i} \, d\boldsymbol{u}_i \\
&= \prod_{i=1}^{m} \int [\boldsymbol{u}_i]_{\boldsymbol{U}_i} \times [\boldsymbol{y}_i^o \mid \boldsymbol{u}_i]_{\boldsymbol{Y}_i \mid \boldsymbol{U}_i} \times [\log t_i, \delta_i \mid \boldsymbol{u}_i]_{\log T_i, \Delta_i \mid \boldsymbol{U}_i} \, d\boldsymbol{u}_i \\
&\quad \textit{iaomi}, \text{ dado que } [\log D_i, \Delta_i \mid \boldsymbol{U}_i] \perp\!\!\!\perp [\boldsymbol{Y}_i \mid \boldsymbol{U}_i] \\
&= \prod_{i=1}^{m} \int f_{\boldsymbol{U}_i}(\boldsymbol{u}_i) f_{\boldsymbol{Y}_i \mid \boldsymbol{U}_i}(\boldsymbol{y}_i^o) \Big\{ f_{\log D_i \mid \boldsymbol{U}_i}\big(\log t_i\big) S_{\log C_i \mid \boldsymbol{U}_i}\big(\log t_i\big) \Big\}^{\delta_i} \\
&\quad \times \Big\{ f_{\log C_i \mid \boldsymbol{U}_i}\big(\log t_i\big) S_{\log D_i \mid \boldsymbol{U}_i}\big(\log t_i\big) \Big\}^{1-\delta_i} \, d\boldsymbol{u}_i \\
&= \prod_{i=1}^{m} \int f_{\boldsymbol{U}_i}(\boldsymbol{u}_i) f_{\boldsymbol{Y}_i \mid \boldsymbol{U}_i}(\boldsymbol{y}_i^o) \Big\{ f_{\log D_i \mid \boldsymbol{U}_i}\big(\log t_i\big) \Big\}^{\delta_i} \Big\{ S_{\log D_i \mid \boldsymbol{U}_i}\big(\log t_i\big) \Big\}^{1-\delta_i} \\
&\quad \times \prod_{i=1}^{m} \Big\{ S_{\log C_i \mid \boldsymbol{U}_i}\big(\log t_i\big) \Big\}^{\delta_i} \Big\{ f_{\log C_i \mid \boldsymbol{U}_i}\big(\log t_i\big) \Big\}^{1-\delta_i} \, d\boldsymbol{u}_i
\end{aligned}
\tag{4.18}
$$

Onde δ_i é o indicador de *dropout* como definido em (4.11). Admitindo que o processo de censura é independente quer do processo de *dropout* quer do processo longitudinal (censura não-informativa), a distribuição do tempo de censura $\boldsymbol{C}_i$ não é função dos parâmetros de interesse $\boldsymbol{\theta}$. Assim, o último termo em (4.18) é constante, e a inferência sobre $\boldsymbol{\theta}$ pode então basear-se na função de verosimilhança

$$
L(\boldsymbol{\theta}; \boldsymbol{y}^o, \log \boldsymbol{t}, \boldsymbol{\delta}) \propto \prod_{i=1}^{m} \int f_{U_i}(u_i) f_{\boldsymbol{Y}_i \mid U_i}(\boldsymbol{y}_i^o) \Big\{ f_{\log D_i \mid U_i}\big(\log t_i\big) \Big\}^{\delta_i} \Big\{ S_{\log D_i \mid U_i}\big(\log t_i\big) \Big\}^{1-\delta_i} \, dU_i
\tag{4.19}
$$

A função de verosimilhança (4.19) envolve a distribuição marginal de U_i

$$U_i \sim N\left(0,\ \nu^2\right),$$

a distribuição condicional de $\log D_i \mid \boldsymbol{U}_i$

$$\log D_i \mid U_i \sim N(\boldsymbol{W}_i^\top \boldsymbol{\alpha} + \gamma U_i,\ \eta^2),$$

e a distribuição condicional de de $\boldsymbol{Y}_i \mid \boldsymbol{U}_i$

$$\boldsymbol{Y}_i \mid \boldsymbol{U}_i \sim MVN(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{U}_i,\ \tau^2\boldsymbol{I}).$$

Substituindo as respetivas funções densidade de probabilidade e de sobrevivência em (4.19), a verosimilhança dos parâmetros do modelo pode ser escrita como

$$\begin{aligned} L(\boldsymbol{\theta}; \boldsymbol{y}^o, \log \boldsymbol{t}, \boldsymbol{\delta}) \propto \prod_{i=1}^{m} \int & \frac{\exp\left\{-\frac{1}{2}\frac{U_i}{\nu^2}\right\}}{(2\pi\nu^2)^{1/2}} \\ & \times \frac{\exp\left\{-\frac{1}{2}\frac{1}{\tau^2}(\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{U}_i)^\top(\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{U}_i)\right\}}{(2\pi\tau^2)^{n_i/2}} \\ & \times \left\{\frac{\exp\left\{-\frac{1}{2}u_i(\log t_i)^2\right\}}{(2\pi\eta^2)^{1/2}}\right\}^{\delta_i} \times \left\{1 - \Phi\left(u_i(\log t_i)\right)\right\}^{1-\delta_i} dU_i \end{aligned} \tag{4.20}$$

onde

$$u_i(\log t_i) = \frac{\log t_i - \mu_{\log D_i|U_i}}{\sigma_{\log D_i|U_i}} = \frac{\log t_i - \boldsymbol{W}_i^\top\boldsymbol{\alpha} - \gamma U_i}{\sqrt{\eta^2}}. \tag{4.21}$$

Aplicando a transformação do logaritmo Neperiano a (4.20), obtém-se a função de log-verosimilhança

$$\begin{aligned} l(\boldsymbol{\theta}; \boldsymbol{y}^o, \log \boldsymbol{t}, \boldsymbol{\delta}) \propto \sum_{i=1}^{m} \log \Bigg(\int & \frac{\exp\left\{-\frac{1}{2}\frac{U_i}{\nu^2}\right\}}{(2\pi\nu^2)^{1/2}} \\ & \times \frac{\exp\left\{-\frac{1}{2}\frac{1}{\tau^2}(\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{U}_i)^\top(\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{U}_i)\right\}}{(2\pi\tau^2)^{n_i/2}} \\ & \times \left\{\frac{\exp\left\{-\frac{1}{2}u_i(\log t_i)^2\right\}}{(2\pi\eta^2)^{1/2}}\right\}^{\delta_i} \times \left\{1 - \Phi\left(u_i(\log t_i)\right)\right\}^{1-\delta_i} dU_i \Bigg). \end{aligned} \tag{4.22}$$

Passo E

Aplicando o algoritmo EM, tratando os efeitos aleatórios como dados em falta, nós pretendemos encontrar os parâmetros $\boldsymbol{\theta}$ do modelo de dados completo, mas usando apenas a informação observada. No passo E calculamos o valor esperado da log-verosimilhança dos dados completos, i.e.,

$$
\begin{aligned}
Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(it)}) =& E\left[l(\boldsymbol{\theta}; \boldsymbol{y}^o, \log \boldsymbol{t}, \boldsymbol{\delta}, U_i) \mid \boldsymbol{y}^o, \log \boldsymbol{t}, \boldsymbol{\delta}; \boldsymbol{\theta}^{(it)}\right] \\
=& \sum_i^m \int \Bigg\{ \log f_{U_i}(U_i; \nu^2) + \log f_{\boldsymbol{Y}_i|U_i}(\boldsymbol{y}^o; \boldsymbol{\beta}, \tau^2) \\
& + \log f_{\log T_i, \Delta_i|U_i}(\log t_i; \boldsymbol{\alpha}, \gamma, \eta^2) \Bigg\} \cdot f_{U_i|\boldsymbol{Y}_i, \log T_i, \Delta_i}\left(U_i; \boldsymbol{\theta}^{(it)}\right) dU_i \\
=& \sum_i^m \int \Bigg\{ \frac{1}{2}\left(- \log(2\pi) - \log(\nu^2) - \frac{U_i}{\nu^2} \right) \\
& + \frac{1}{2}\left(-n_i \log(2\pi) - n_i \log(\tau^2) - \frac{(\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{u}_i)^\top (\boldsymbol{y}_i^o - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{u}_i)}{\tau^2} \right) \\
& + \delta_i \frac{1}{2}\left(- \log(2\pi) - \log(\eta^2) - u_i (\log t_i)^2 \right) \\
& + (1 - \delta_i) \log\left(1 - \Phi\left(u_i(\log t_i)\right)\right) \Bigg\} \cdot f_{U_i|\boldsymbol{Y}_i, \log T_i, \Delta_i}\left(U_i; \boldsymbol{\theta}^{(it)}\right) dU_i.
\end{aligned}
$$

Devido à presença de formas não fechadas na componente $\log f_{\log T_i, \Delta_i|U_i}(\log t_i; \boldsymbol{\alpha}, \gamma, \eta^2)$, a avaliação de $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(it)})$ requer a utilização de métodos de integração numérica, como as regras de quadratura Gaussiana ou da amostragem de Monte Carlo.

Passo M

No passo M obtemos os parâmetros $\boldsymbol{\theta}$ atualizados por $\boldsymbol{\theta}^{(it+1)} = \arg_{\boldsymbol{\theta}} \max Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(it)})$. Apresentam-se abaixo os cálculos da derivada do valor esperado da log-verosimilhança dos dados completos com respeito aos parâmetros $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \gamma, \nu^2, \tau^2, \eta^2)$. Os cálculos apresentados fazem uso de um conjunto de propriedades de operações sobre matrizes e da fórmula de Leibniz descritas nos anexos C.2 e C.3 (p. 135 e 136), respetivamente.

$$E_c\left[\frac{\partial l(\boldsymbol{\theta};\boldsymbol{y}^o,\log\boldsymbol{t},\boldsymbol{\delta},U_i)}{\partial\boldsymbol{\beta}}\right] = 0+\sum_i^m E_c\left[\frac{\partial}{\partial\boldsymbol{\beta}}\left(-\frac{1}{2}\frac{(\boldsymbol{y}_i^o-\boldsymbol{X}_i\boldsymbol{\beta}-\boldsymbol{u}_i)^\top(\boldsymbol{y}_i^o-\boldsymbol{X}_i\boldsymbol{\beta}-\boldsymbol{u}_i)}{\tau^2}\right)\right]$$
$$= \sum_i^m E_c\left[\frac{1}{\tau^2}\boldsymbol{X}_i^\top(\boldsymbol{y}_i^o-\boldsymbol{X}_i\boldsymbol{\beta}-\boldsymbol{u}_i)\right], \tag{4.23}$$

$$E_c\left[\frac{\partial l(\boldsymbol{\theta};\boldsymbol{y}^o,\log\boldsymbol{t},\boldsymbol{\delta},U_i)}{\partial\tau^2}\right] = 0+\sum_i^m E_c\left[\frac{\partial}{\partial\tau^2}\left(-\frac{n_i}{2}\log(\tau^2)-\frac{(\boldsymbol{y}_i^o-\boldsymbol{X}_i\boldsymbol{\beta}-\boldsymbol{u}_i)^\top(\boldsymbol{y}_i^o-\boldsymbol{X}_i\boldsymbol{\beta}-\boldsymbol{u}_i)}{2\tau^2}\right)\right]$$
$$= \sum_i^m E_c\left[\frac{1}{2}\frac{1}{\tau^2}\left(-n_i+\frac{1}{\tau^2}(\boldsymbol{y}_i^o-\boldsymbol{X}_i\boldsymbol{\beta}-\boldsymbol{u}_i)^\top(\boldsymbol{y}_i^o-\boldsymbol{X}_i\boldsymbol{\beta}-\boldsymbol{u}_i)\right)\right], \tag{4.24}$$

$$E_c\left[\frac{\partial l(\boldsymbol{\theta};\boldsymbol{y}^o,\log\boldsymbol{t},\boldsymbol{\delta},U_i)}{\partial\nu^2}\right] = 0+\sum_i^m E_c\left[\frac{\partial}{\partial\nu^2}\left(-\frac{\log(\nu^2)}{2}-\frac{U_i}{2\nu^2}\right)\right]$$
$$= \sum_i^m E_c\left[\frac{1}{\nu^2}\left(-\frac{1}{2}+\frac{U_i}{2\nu^2}\right)\right], \tag{4.25}$$

$$E_c\left[\frac{\partial l(\boldsymbol{\theta};\boldsymbol{y}^o,\log\boldsymbol{t},\boldsymbol{\delta},U_i)}{\partial\eta^2}\right] = 0+\sum_i^m E_c\left[\frac{\partial}{\partial\eta^2}\Big(\delta_i\frac{1}{2}\left(-\log(\eta^2)-u_i(\log t_i)^2\right)\right.$$
$$\left.+(1-\delta_i)\log\left(1-\Phi\big(u_i(\log t_i)\big)\right)\Big)\right]$$
$$= \sum_i^m E_c\left[\delta_i\frac{1}{\eta^2}\frac{(u_i(\log t_i)^2-1)}{2}\right.$$
$$\left.+(1-\delta_i)\frac{1}{\eta^2}\frac{\varphi\big(u_i(\log t_i)\big)}{1-\Phi\big(u_i(\log t_i)\big)}\frac{u_i(\log t_i)}{2}\right] \tag{4.26}$$

$$E_c\left[\frac{\partial l(\boldsymbol{\theta};\boldsymbol{y}^o,\log\boldsymbol{t},\boldsymbol{\delta},U_i)}{\partial\gamma}\right] = 0+\sum_i^m E_c\left[\frac{\partial}{\partial\gamma}\left(\delta_i\frac{-u_i(\log t_i)^2}{2}+(1-\delta_i)\log\left(1-\Phi\big(u_i(\log t_i)\big)\right)\right)\right]$$
$$= \sum_i^m E_c\left[\delta_i\frac{U_i}{\sqrt{\eta^2}}u_i(\log t_i)+(1-\delta_i)\frac{\varphi\big(u_i(\log t_i)\big)}{1-\Phi\big(u_i(\log t_i)\big)}\frac{U_i}{\sqrt{\eta^2}}\right] \tag{4.27}$$

$$E_c\left[\frac{\partial l(\boldsymbol{\theta};\boldsymbol{y}^o,\log\boldsymbol{t},\boldsymbol{\delta},U_i)}{\partial\boldsymbol{\alpha}}\right]=0+\sum_i^m E_c\left[\frac{\partial}{\partial\boldsymbol{\alpha}}\left(\delta_i\frac{-u_i(\log t_i)^2}{2}+(1-\delta_i)\log\left(1-\Phi\big(u_i(\log t_i)\big)\right)\right)\right]$$
$$=\sum_i^m E_c\left[\delta_i\frac{\boldsymbol{W}_i}{\sqrt{\eta^2}}u_i(\log t_i)+(1-\delta_i)\frac{\varphi\big(u_i(\log t_i)\big)}{1-\Phi\big(u_i(\log t_i)\big)}\frac{\boldsymbol{W}_i}{\sqrt{\eta^2}}\right] \tag{4.28}$$

onde

$$E_c\left[\frac{\partial l(\boldsymbol{\theta};\boldsymbol{y}^o,\log\boldsymbol{t},\boldsymbol{\delta},U_i)}{\partial\theta}\right]=E\left[\frac{\partial l(\boldsymbol{\theta};\boldsymbol{y}^o,\log\boldsymbol{t},\boldsymbol{\delta},U_i)}{\partial\theta}\mid\boldsymbol{y}^o,\log\boldsymbol{t},\boldsymbol{\delta};\boldsymbol{\theta}^{(it)}\right].$$

A partir das equações (4.23)-(4.25) obtêm-se formas fechadas para os parâmetros $\boldsymbol{\beta}$, τ^2, e ν^2:

$$E_c\left[\frac{\partial l(\boldsymbol{\theta};\boldsymbol{y}^o,\log\boldsymbol{t},\boldsymbol{\delta},U_i)}{\partial\boldsymbol{\beta}}\right]=0\Leftrightarrow\ \hat{\boldsymbol{\beta}}=\frac{1}{\sum_i^m\boldsymbol{X}_i}\sum_i^m\Big(\boldsymbol{y}_i^o-E_c[\boldsymbol{U}_i]\Big),$$

$$E_c\left[\frac{\partial l(\boldsymbol{\theta};\boldsymbol{y}^o,\log\boldsymbol{t},\boldsymbol{\delta},U_i)}{\partial\tau^2}\right]=0\Leftrightarrow\ \hat{\tau}^2=\frac{1}{\sum_i^m n_i}\sum_i^m\Big((\boldsymbol{y}_i^o-\boldsymbol{X}_i\boldsymbol{\beta})^\top(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}-2E_c[U_i])+E_c[U_i^2]\Big),$$

$$E_c\left[\frac{\partial l(\boldsymbol{\theta};\boldsymbol{y}^o,\log\boldsymbol{t},\boldsymbol{\delta},U_i)}{\partial\nu^2}\right]=0\Leftrightarrow\ \hat{\nu}^2=\frac{1}{m}\sum_i^m E_c[U_i^2].$$

Para os restantes parâmetros, não conseguindo obter soluções de forma fechada a partir das equações (4.26)-(4.28), o passo M é implementado através da atualização de Newton-Raphson de um passo, i.e.,

$$\hat{\boldsymbol{\theta}}^{(it+1)}=\hat{\boldsymbol{\theta}}^{(it)}-\left(\frac{\partial S(\hat{\boldsymbol{\theta}}^{(it)})}{\partial\boldsymbol{\theta}}\right)^{-1}S(\hat{\boldsymbol{\theta}}^{(it)}),$$

onde $\hat{\boldsymbol{\theta}}^{(it)}$ denota os valores do parâmetros $\boldsymbol{\theta}$ na iteração atual, e $\partial S(\hat{\boldsymbol{\theta}}^{(it)})/\partial\boldsymbol{\theta}$ denota os blocos correspondentes da matriz Hessiana, avaliada em $\hat{\boldsymbol{\theta}}^{(it)}$. Os elementos do vetor *score* de $\boldsymbol{\theta}$ têm a forma

$$
\begin{aligned}
S(\gamma) &= \frac{1}{\sqrt{\eta^2}}\sum_{i}^{m}\left(\delta_i E_c\left[u_i(\log t_i)\, U_i\right] + (1-\delta_i)E_c\left[\frac{\varphi(u_i(\log t_i))}{1-\Phi(u_i(\log t_i))}U_i\right]\right) \\
&= \frac{1}{\sqrt{\eta^2}}\sum_{i}^{m}\left(\delta_i\frac{(\log t_i - \boldsymbol{W}_i^\top\boldsymbol{\alpha})E_c[U_i] - \gamma E_c[U_i^2]}{\sqrt{\eta^2}} + (1-\delta_i)E_c\left[\frac{\varphi(u_i(\log t_i))}{1-\Phi(u_i(\log t_i))}U_i\right]\right),
\end{aligned}
$$

$$
\begin{aligned}
S(\eta^2) &= \frac{1}{2\eta^2}\sum_{i}^{m}\left(\delta_i\left(E_c[u_i(\log t_i)^2] - 1\right) + (1-\delta_1)E_c\left[\frac{\varphi(u_i(\log t_i))}{1-\Phi(u_i(\log t_i))}u_i(\log t_i)\right]\right. \\
&= \frac{1}{2\eta^2}\sum_{i}^{m}\left(\delta_i\left(\frac{(\log t_i - \boldsymbol{W}_i^\top\boldsymbol{\alpha})^2 + \gamma^2 E_c[U_i^2] - 2\gamma(\log t_i - \boldsymbol{W}_i^\top\boldsymbol{\alpha})E[U_i]}{\eta^2} - 1\right)\right. \\
&\quad \left.\left. + (1-\delta_i)\frac{1}{\sqrt{\eta^2}}E_c\left[\frac{\varphi(u_i(\log t_i))}{1-\Phi(u_i(\log t_i))}(\log t_i - \boldsymbol{W}_i^\top\boldsymbol{\alpha} - \gamma U_i)\right]\right),
\end{aligned}
$$

$$
\begin{aligned}
S(\boldsymbol{\alpha}) &= \frac{1}{\sqrt{\eta^2}}\sum_{i}^{m}\left(\delta_i E_c\left[u_i(\log t_i)\right] + (1-\delta_i)E_c\left[\frac{\varphi(u_i(\log t_i))}{1-\Phi(u_i(\log t_i))}\right]\right)\boldsymbol{W}_i \\
&= \frac{1}{\sqrt{\eta^2}}\sum_{i}^{m}\left(\delta_i\frac{\log t_i - \boldsymbol{W}_i^\top\boldsymbol{\alpha} - E_c\left[U_i\right]}{\sqrt{\eta^2}} + (1-\delta_i)E_c\left[\frac{\varphi(u_i(\log t_i))}{1-\Phi(u_i(\log t_i))}\right]\right)\boldsymbol{W}_i,
\end{aligned}
$$

onde $u_i(\cdot)$ é definido como em (4.21), $\varphi(\cdot)$ é a função densidade de probabilidade da distribuição $N(0,1)$.

A utilização do algoritmo EM permitiu-nos obter, na presença de tempo-até-abandono censurados, pela primeira vez expressões com forma fechada para alguns dos parâmetros do modelo.

Capítulo 5

Conclusões e trabalho futuro

Os estudos longitudinais constituem uma importante ferramenta de investigação uma vez que fornecem conhecimento sobre a evolução e persistência de uma caraterística de interesse e dos fatores que contribuem para o seu desenvolvimento. A existência de medições repetidas ao longo do tempo sobre o mesmo indivíduo permite separar o efeito do coorte do efeito da idade [2]. Cada indivíduo funciona como o seu próprio controlo, sendo as alterações na caraterística de interessa observadas em cada indivíduo ao longo do tempo estimadas independentemente de qualquer variação da caraterística entre indivíduos. Assim, os estudos longitudinais possibilitam uma compreensão mais profunda sobre as relações causa-efeito das variáveis observadas. Contudo, como num estudo longitudinal a recolha de observações não é concentrada num instante mas dilatada no tempo, a presença de dados omissos por abandono dos participantes é comum. Quando a causa abandono está relacionado com a resposta longitudinal em estudo diz-se que a omissão ocorre de forma não aleatória. Nesses casos os dados observados não podem ser considerados uma amostra aleatória do conjunto de dados completo. Por isso, de um modo geral, métodos de análise válidos para os dados completos deixam de produzir inferências válidas.

Neste trabalho apresenta-se a função `trim()` para o software R. Esta função permite ao utilizador simular conjuntos de dados omissos por abandono em conjuntos de dados longitudinais completos. Esta função distingue-se de outras já disponibilizadas noutros *packages* no software por permitir ao utilizador controlar a proporção global de abandono observada, e pelo elevado nível de personalização do mecanismo de omissão. O modo de implementação

da função permite que pela utilização de uma semente de aleatoriedade se garanta que os dados incompletos gerados a partir de diferentes mecanismos de omissão apresentem a mesma progressão no tempo de dados omissos e de abandono de indivíduos, removendo essa fonte de variabilidade da análise. A função apresenta ainda tempos de computação muito céleres, com $Q_{0.25} = 0.006s$ e $Q_{0.75} = 0.765s$ por conjunto de dados para o estudo conduzido. A utilidade da função desenvolvida não se esgota no estudo de simulação desenvolvido no decurso deste trabalho. Pelo que trabalho futuro passa por a incorporar a função `trim()` num *package* do software R e, assim, disponibilizar publicamente esta ferramenta de simulação a outros investigadores.

O estudo de simulação conduzido visou avaliar de que forma quer caraterísticas de um conjunto de dados longitudinais gerado a partir de um modelo com efeito aleatório na interseção, quer caraterísticas dos dados omissos por abandono podem influenciar inferências baseadas exclusivamente nos dados observados. Os resultados indicam que o aumento do número m de participantes e do número n de ocasiões em que cada participante é observado conduz a uma diminuição da variabilidade da resposta média dos dados observados e a uma diminuição do EPM dos parâmetros. Estes resultados sugerem que para situações em que o modelo considerado seja adequado e que exista uma grande suscetibilidade para abandono informativo que o desenho do estudo inclua um aumento de número de participantes e/ou do número de vezes em que são observados, para atuar como contrapeso ao possível abandono de alguns dos participantes. No entanto é importante notar que os nossos resultados sugerem que o benefício na redução do EPM parece diminuir à medida que o número m de participantes no estudo também aumenta. Por isso, em estudos que envolvam um elevado número de indivíduos o adjuvante de incluir mais participantes pode não justificar o investimento. O aumento da proporção de abandono traduz-se num aumento da variabilidade da resposta média observada e num agravamento do EPM dos parâmetros estimados. O estudo de simulação comparou três mecanismos de omissão MCAR, MAR e MNAR. Nos casos em que a omissão ocorreu de forma totalmente aleatória, a diferença entre valor médio observado e o valor médio real dos dados completos é aproximadamente nula. Estes resultados reiteram a informação vinculada pela literatura de que na presença de um mecanismo MCAR os dados observados podem ser considerados um amostra aleatório dos dados completos. Ao nível dos

mecanismos MAR e MNAR é notória uma diferença entre valor médio dos dados observados e dos dados completos. Esta diferença reflete a dependência imposta entre a resposta longitudinal e o tempo-até-*dropout* nestes mecanismos de omissão pela função `trim()`. Ao nível do processo de estimação, não foi possível detetar diferenças expressivas entre os diferentes mecanismos de omissão. Neste contexto, é importante notar que no desenho do estudo se impôs a observação dos mesmos padrões globais de abandono entre os diferentes mecanismos e ainda que os conjuntos de dados fossem ajustados com a estrutura de correlação correta. Por outro lado, estes resultados podem advir do facto das progressões individuais serem relativamente monótonas. Estas caraterísticas podem explicar estes resultados, no entanto estas hipóteses deverão ser exploradas em estudos de simulação posteriores. O trabalho desenvolvido pode ser alargado para incluir modelos linear mistos que incluam estruturas de correlação mais complexas que incluam, por exemplo, um efeito efeito aleatório para o declive ou um processo estacionário Gaussiano. Este tipo de modelos permitirá avaliar de que forma a distribuição da variabilidade total entre a variabilidade entre indivíduos e dentro do individuo afetam a progressão da resposta média dos dados observados e inferências que ignorem o processo de omissão. O estudo simulação apresentado pressupõe que em cada conjunto de dados só está presente um mecanismo de omissão, para melhorar a comparabilidade do efeito dos mesmos. Em estudos de simulação posteriores poder-se-á avaliar de que forma diferentes combinações de mecanismos num mesmo conjunto de dados podem afetar os resultados obtidos.

Diggle et al. (2008) [1] propuseram um modelo simples totalmente paramétrico, o TGM, para descrever a distribuição conjunta da resposta longitudinal e do tempo até abandono. Este modelo destaca-se de outras abordagens de modelação descritas na literatura pela sua simplicidade e rapidez de computação. Neste trabalho propusemos novas estruturas de correlação com uma interpretação mais intuitiva com o objetivo de ultrapassar a interpretação puramente empírica do vetor de correlação cruzada entre as medições longitudinais e o tempo até abandono. Usamos o algoritmo EM para obter as estimativas de máxima verosimilhança dos parâmetros do modelo para uma das estruturas propostas. Esta abordagem permitiu-nos obter, na presença de tempos-até-*dropout*, censurados, pela primeira vez, relativamente ao trabalho inicial, expressões com forma fechada para alguns dos parâmetros do modelo. Trabalho futuro passa por aplicar o modelo desenvolvido a conjuntos de dados reais onde

se tenha observado abandono informativo. E, assim, avaliar os ganhos em termos de interpretabilidade obtidos pelas novas estruturas de correlação; e, ainda, avaliar os benefícios em termos de rapidez de computação pela obtenção pela primeira vez de formas fechadas para alguns dos parâmetros.

Anexo A

Código R

Os ficheiros .R com o código disponibilizado neste anexo podem ser encontrados em https://bit.ly/2Oe17ga.

A.1 Função `trim()`

```
trim = function(data.arg,
varying.arg,
times.arg,
droptime.arg = times.arg[1],
perc.drop,
mechanism.arg,
summary.arg,
lambdaf.arg,
nsim.arg = 1,
seed.arg = NULL,
savedir.arg = NULL,
savename.arg = NULL)

{

if(mechanism.arg == "mar" & droptime.arg == times.arg[1]){print("MAR
    mechanism does not allow dropout at the first measurement time.")}
else {

# Initiate the timer
start_time1 <- proc.time()[[3]]

# Define global variables
m <- nrow(data.arg)
n <- length(times.arg)
N <- m * n
m.drop <- round(m * perc.drop)
```

```
27
28 times.to.drop <- times.arg[which(times.arg%in%droptime.arg) : n]
29
30 # Select scenarios
31
32 num <- RcppAlgos::permuteCount(v = times.to.drop, m = m.drop,
       repetition = TRUE)
33
34 l2.num <- gmp:: log2.bigz (num)
35
36 mySamp <- gmp::urand.bigz(n = nsim.arg, size = l2.num, seed = seed.arg
       )
37
38 myScen <- RcppAlgos::permuteSample(v = times.to.drop, m = m.drop,
       repetition = TRUE, sampleVec = mySamp + 1)
39
40 myScen <- cbind(matrix(NA, nrow = nsim.arg, ncol = m-m.drop), myScen)
       # To add individuals who didn't dropout
41
42 n.scen <- num * choose(m, m-m.drop)
43
44 print(paste0("Total number of possible permutations: ", n.scen))
45
46 # Calculate the percentage of missing observations in each scenario
47
48 perc.miss <- apply(myScen, 1, function(x){
49
50 sum(n - match(x, times.arg) + 1, na.rm = T)/N
51
52 })
53
54 # Lambda & Probability matrix
55 if(mechanism.arg != "mcar"){
56
57 lambda.func <- function(y.arg, lambdaf.arg){
58
59 sapply(y.arg, function(x){
60
61 if ( length(which(lambdaf.arg$y == x)) != 0 ) {lambdaf.arg$rate[which(
       lambdaf.arg$y == x)]}
62
63 else if ( x < min(lambdaf.arg$y) || x > max(lambdaf.arg$y) ) { 0 }
64
65 else {
66
67 if(head(lambdaf.arg$y, 1) > tail(lambdaf.arg$y,1)) {sign <- -1} #
       Negative slope
68 else {sign <- +1} # Positive slope
```

```
69
70 pos <- tail(rank(sign*c(lambdaf.arg$y, x)), 1)
71
72 # Linear Interpolation: lambda = m * y + b
73
74 lambda <- c(lambdaf.arg$rate[pos-1], lambdaf.arg$rate[pos])
75 y <- c(lambdaf.arg$y[pos-1], lambdaf.arg$y[pos])
76
77 m <- (lambda[2] - lambda[1]) / (y[2] - y[1]) # Formula for Slope m
78 b <- lambda[1] - m*y[1] # Solve for Intercept b
79
80 m * x + b
81
82 }
83 })
84
85 }
86
87 pexp.func <- function(lambda.arg, col.arg, times.arg){
88
89 sapply(lambda.arg, function(x){
90
91 if (col.arg == 1) { pexp(times.arg[col.arg], rate = x) }
92
93 else { pexp(times.arg[col.arg] - times.arg[col.arg - 1], rate = x) }
94
95 })
96
97 }
98
99 # Past
100 if(mechanism.arg == "mar"){
101
102 summary.lambda <- matrix(0, ncol = n, nrow = m)
103 summary.prob <- matrix(0, ncol = n + 1, nrow = m)
104
105 for(col in 2:n){
106
107 summary.y <- apply(data.arg, 1, function(x){
108
109 if(summary.arg == "mean"){ mean( x[varying.arg[1:(col-1)]] ) }
110 else if (summary.arg == "one"){ mean( x[varying.arg[col-1]] ) }
111 else if (summary.arg == "median"){ median( x[varying.arg[1:(col-1)]]
    ) }
112 else if (summary.arg == "range"){ diff(range( x[varying.arg[1:(col-1)
    ]] )) }
113 else if (summary.arg == "gaus.wt" &&  col == 2){ x[varying.arg[1]] }
114 else if (summary.arg == "gaus.wt" &&  col != 2 ){
```

```
115
116 weights <- dnorm(times.arg[1:(col-1)], mean = times.arg[(col-1)], sd =
        (times.arg[(col-1)] - times.arg[1])/3)
117 t(as.matrix(weights/sum(weights)))%*%x[varying.arg[1:(col-1)]]
118
119 }
120
121 })
122
123 summary.lambda[, col] <- lambda.func(y.arg = summary.y,
124 lambdaf.arg = lambdaf.arg)
125
126 summary.prob[, col] <- pexp.func(lambda.arg = summary.lambda[, col],
127 col.arg = col,
128 times.arg = times.arg)
129
130
131 }
132
133 summary.prob[, n + 1] <- 1 - summary.prob[, n] # To add the
       probability of each patient to not dropout during the study
134
135 }
136
137 # All sequence
138 else if(mechanism.arg == "mnar1"){
139
140 summary.y <- apply(data.arg, 1, function(x){
141
142 if(summary.arg == "mean"){mean(x[varying.arg])}
143 else if (summary.arg == "median"){median(x[varying.arg])}
144 else if (summary.arg == "range"){diff(range(x[varying.arg]))}
145 })
146
147 summary.lambda <- lambda.func(y.arg = summary.y,
148 lambdaf.arg = lambdaf.arg)
149
150 summary.lambda <- matrix(summary.lambda, nrow = m, ncol = length(
       varying.arg), byrow = F)
151
152 summary.prob <- matrix(0, ncol = n + 1, nrow = m)
153
154 for(col in 1:n){
155
156 summary.prob[, col] <- pexp.func(lambda.arg = summary.lambda[, col],
157 col.arg = col,
158 times.arg = times.arg)
159 }
```

```
160
161 summary.prob[, n + 1] <- 1 - summary.prob[, n] # To add the
       probability of each patient to not dropout during the study
162
163 }
164
165 # Future
166 else if(mechanism.arg == "mnar2"){
167
168 summary.lambda <- matrix(0, ncol = n, nrow = m)
169 summary.prob <- matrix(0, ncol = n + 1, nrow = m)
170
171 for(col in 1:n){
172
173 summary.y <- apply(data.arg, 1, function(x){
174
175 if(summary.arg == "mean"){ mean(x[varying.arg[col:n]]) }
176 else if (summary.arg == "one"){ x[varying.arg[col]] }
177 else if (summary.arg == "median"){ median(x[varying.arg[col:n]]) }
178 else if (summary.arg == "range"){ diff(range(x[varying.arg[col:n]]))
       }
179 else if (summary.arg == "gaus.wt" &&  col == n){ x[varying.arg[n]] }
180 else if (summary.arg == "gaus.wt" &&  col != n){
181
182 weights <- dnorm(times.arg[col:n], mean = times.arg[col], sd = (time
       s.arg[n] - times.arg[col])/3)
183 t(as.matrix(weights/sum(weights)))%*%x[varying.arg[col:n]]
184
185 }
186 })
187
188 summary.lambda[, col] <- lambda.func(y.arg = summary.y,
189 lambdaf.arg = lambdaf.arg)
190
191 summary.prob[, col] <- pexp.func(lambda.arg = summary.lambda[, col],
192 col.arg = col,
193 times.arg = times.arg)
194
195 }
196
197 summary.prob[, n + 1] <- 1 - summary.prob[, n] # To add the
       probability of each patient to not dropout during the study
198
199 }
200
201 }
202
203 # Sort each scenario according to its probabilities
```

```
204 if(mechanism.arg != "mcar"){
205
206 sort.func <- function(scn.arg, prob.m.arg, times.arg){
207
208 scn <- scn.arg[order(x = scn.arg, na.last = TRUE)]
209
210 indic <- match(scn, c(times.arg, NA))
211
212 prob.m <- prob.m.arg[, indic]
213
214 m <- length(scn)
215
216 null.v <- rep(0, m)
217
218 for(col in 1:m) {
219
220 if(sum(prob.m[, col]) == 0){print("Error: all probabilities are equal
        to zero.")}
221
222 #sub <- sample(1:m, 1, prob = prob.m[, col])
223 sub <- which.max(prob.m[, col])
224
225 prob.m[sub, ] <- null.v
226 prob.m[sub, col] <- sub
227
228 }
229
230 pos <- colSums(prob.m)
231
232 scn <- scn[pos]
233
234 return(scn)
235
236 }
237
238 myScen <- apply(myScen, 1, function(x){ sort.func(scn.arg = x,
       prob.m.arg = summary.prob, times.arg = times.arg) })
239
240 myScen <- t(myScen)
241
242 } else if(mechanism.arg == "mcar"){
243
244 myScen <- apply(myScen, 1, function(x){ sample(x) })
245
246 myScen <- t(myScen)
247
248 }
249
```

```
250 dur_time1 <- proc.time()[[3]] - start_time1
251
252 # Trim data
253
254 all.data <- matrix(0, nrow = nsim.arg * m, ncol = ncol(data.arg) + 6)
255
256 for(sim in 1:nsim.arg){
257
258 start_time2 <- proc.time()[[3]]
259
260 inc.data <- data.arg
261
262 for(i in 1:m){
263
264 drop.time <- myScen[sim, i]
265
266 if(is.na(drop.time)){next}
267 else{
268
269 inc.data[i, varying.arg[which(times.arg%in%drop.time) : n]] <- NA
270
271 }
272
273 }
274
275 dur_time2 <- proc.time()[[3]] - start_time2
276
277 aux <- ncol(inc.data)
278 rows <- (-m + 1 + sim * m) : ( sim * m)
279
280 all.data[rows, 1:aux] <- as.matrix(inc.data)
281
282 all.data[rows, aux + 1] <- rep( sim, m) # Simulation number
283 all.data[rows, aux + 2] <- rep( m.drop/m , m) # Percentage of dropout
284 all.data[rows, aux + 3] <- rep( perc.miss[sim], m) # Percentage of
      missing observations
285 all.data[rows, aux + 4] <- rep( dur_time1, m) # Duration 1 (
      Initialization)
286 all.data[rows, aux + 5] <- rep( dur_time2, m) # Duration 2 (Each
      simulation)
287 all.data[rows, aux + 6] <- myScen[sim, ] # Subjects dropout time
288
289 }
290
291 all.data <- as.data.frame(all.data)
292
293 names(all.data) <- c(names(data.arg),"sim","perc.drop","perc.miss","
      dur1","dur2","dtime")
```

```
294
295 # Save all.data into a .txt
296 if(!is.null(savedir.arg)){
297
298 if(!is.null(savename.arg)){ file.name <- paste0(savedir.arg,
      savename.arg,".txt") }
299 else { file.name <- paste0(savedir.arg,"m",m,"n",n,mechanism.arg,"d",
      perc.drop,".txt")}
300
301 write.table(all.data,
302 file = file.name,
303 sep = "\t",
304 row.names = FALSE,
305 col.names = TRUE)
306
307 }
308
309 Print all duration time
310 print(paste("Total duration (s):",proc.time()[[3]] - start_time1))
311
312 return(all.data)
313
314 }
315 }
```

A.2 Estudo simulação

```
# To install the required packages
packages <- c("nlme", "ggplot2", "RcppAlgos", "gmp", "viridis")
invisible(lapply(packages, install.packages, character.only = TRUE))

start_time All <- proc.time()[[3]]

# To import the trim function
source("/Volumes/Simulation/Simulation/NAFunc.R")

# Directories
fig.dir <- c("~/Desktop/MScThesis/Thesis PT/Figures/") # Where to save figures
complete.dir <- c("/Volumes/Simulation/CompleteData/") # Where to save/load the complete data
incomplete.dir <- c("/Volumes/Simulation/IncompleteData/") # Where to save the trimmed data
error.dir <- c("/Volumes/Simulation/FittedData/") # Where to save the error data

# Simulation Parameters
seed <- 2019
nsim <- 500 # Number of simulations

## Model Parameters
beta <- c(1000, -3)
nu2 <- 100
tau2 <- 1

## Gen Data
m.v <- c(10, 30, 70, 120) # Number of subjects
n.v <- c(5, 15, 20, 50) # Number of observations / subject

## Trim Data
mec.v <- c("mcar", "mar", "mnar2")
perc.v <- c(0.1, 0.4, 0.6) # Desired percentage of dropout
summ <- c("one")

lambda.par = c(0, 4)
pexp(1, lambda.par)
quant.par = c(0, 1) # quantiles

# Plot Rate Vs Y & Prob Vs Y
{
{pdf(paste0(fig.dir,"C2simlambdafunc.pdf"), height = 10/2.54, width = 20/2.54)
```

```
41
42 {
43 par(mfrow = c(1, 2))
44
45 # Plot Rate Vs Y
46 plot(quant.par, lambda.par, type = 'l',
47 ylab = bquote(lambda[ij]),
48 xlab = bquote(F[2](bold(y)[i])), xaxt = "n")
49 sapply(quant.par, function(q){ axis(1, at = q, labels = bquote(Q[.(q)
      ])) })
50
51 y <- c()
52 rate <- c()
53 for(pos in 1:(length(quant.par)-1)){
54
55 y <- c(y, seq(quant.par[pos], quant.par[pos + 1], length.out = 10))
56 rate <- c(rate, seq(lambda.par[pos], lambda.par[pos + 1], length.out =
       10))
57
58 }
59
60 plot(y, pexp(1, rate), type = 'l',
61 ylab = bquote("F(1;"~lambda[ij]~")"),
62 xlab = bquote(F[2](bold(y)[i])), xaxt = "n")
63
64 sapply(quant.par, function(q){ axis(1, at = q, labels = bquote(Q[.(q)
      ])) })
65 }
66
67 dev.off()}
68 }
69
70 # Generate Complete Data
71
72 # Gen Complete Data
73 {
74
75 set.seed(seed)
76
77 print("Simulation Started")
78 start_time1 <- proc.time()[[3]]
79
80 for(m in m.v){ # Number of subects
81 for(n in n.v){ #number of observations per subject
82
83 start_time2 <- proc.time()[[3]]
84
85 N <- m*n
```

```
86  times <- c(0:(n-1))
87
88  data.wide <- matrix(NA, nrow = m * nsim, ncol = n + 3)
89
90  for(sim in 1:nsim){
91
92  rows <- 1:m + (sim-1)*m
93
94  data.wide[rows, 1] <- 1:m # Id
95
96  design.m <- matrix(c(rep(1, N), rep(times, times = m)), ncol = 2, nrow
        = N)
97
98  betaX <- beta
99  betaX[2] <- beta[2]*(min(n.v)-1)/(n-1)
100
101 mu_ij <- design.m%*%betaX
102
103 U_i <- rnorm(m, mean = 0, sd = sqrt(nu2))
104 U_i <- rep(U_i, each = n)
105
106 Z_ij <- rnorm(N, mean = 0, sd = sqrt(tau2))
107
108 Y <- mu_ij + U_i + Z_ij
109 Y <- matrix(Y, ncol = n, nrow = m, byrow = T)
110
111 data.wide[rows, 2:(n+1)] <- Y
112 dur_time <- proc.time()[[3]] - start_time2
113 data.wide[rows, n + 2] <- rep(dur_time, m)
114 data.wide[rows, n + 3] <- rep(sim, m)
115
116 }
117
118 data.wide <- as.data.frame(data.wide)
119 names(data.wide)[c(1, n+2, n+3)] <- c("id", "dur", "sim")
120 names(data.wide)[2:(n+1)] <- sapply(times, function(x){ paste0("Y.t",
      x) })
121
122 # Save data,wide into a .txt
123 write.table(data.wide,
124 file = paste0(complete.dir,"m",m,"n",n,".txt"),
125 sep = "\t",
126 row.names = FALSE,
127 col.names = TRUE)
128
129 }
130 }
131
```

```
print("Simulation Finished")
# Print all duration time
print(paste("Total simulation duration (s):",proc.time()[[3]] -
    start_time1))
}

# Plot the density of the Random Effects and Measurement Errors
{pdf( paste0 (fig.dir,"C2gendensity.pdf"), height = 16/2.54, width = 16
    /2.54)

{
par(mfrow = c(1,1))

# U_i density
x1 <- seq(from = -3*sqrt(nu2), to = 3*sqrt(nu2), length = 1000)
y1 <- dnorm(x1, mean = 0, sd = sqrt(nu2))

# Z_ij density
x2 <- seq(from = -3*sqrt(tau2), to = 3*sqrt(tau2), length = 1000)
y2 <- dnorm(x2, mean = 0, sd = sqrt(tau2))


plot(0, cex = 0, xlim = range(c(x1,x2)), ylim = range(c(y1,y2)), ylab
     = "Densidade de probabilidade", xlab = "Y")
curve(dnorm(x, mean = 0, sd = sqrt(nu2)), min(c(x1,x2)), max(c(x1,x2)
    ), add= TRUE) # U_i
curve(dnorm(x, mean = 0, sd = sqrt(tau2)), min(c(x1,x2)), max(c(x1,x2
    )), add= TRUE, lty = 2) # Z_ij

legend("topleft", lty = c(1, 2), bty = "n",
legend = sapply(
c(bquote(U[i] ~ "~ Normal(0," ~ nu^2 == .(nu2) ~ ")"),
bquote(Z[ij] ~ "~ Normal(0," ~ tau^2 == .(tau2) ~ ")")
), as.expression)
)
}

dev.off()}

# Spagheti Plots
{pdf( paste0 (fig.dir,"C2compspaghetti.pdf"), height = 30/2.54, width =
     30/2.54)

par(mfcol = c(length(n.v), length(m.v)))
par(mar = c(5.1, 4.1, 4.1, 5)) # (Bottom, Left, Top, Right)

for(m in m.v){

```

```
for(n in n.v){

data <- paste0("m", m, "n", n, ".txt")

data.wide <- read.table(paste0(complete.dir, data), header = TRUE)

sim <- sample(1:nsim, 1)

data.wide <- data.wide[data.wide$sim == sim, ]

# Data in long format
times <- 0:(n-1)
data.long <- reshape(data.wide, direction = "long", varying = list(2 :
    (n + 1)), times = times)
names(data.long)[5] <- "Y"

# Spagheti plots
ylim <- range(data.long$Y)
plot(0,0, ylim = ylim, xlim = range(times), xlab = "Tempo", ylab = "Y
   ", cex = 0)
title(paste0("m=", m, "  n=", n, " (c=",sim,")"), font.main = 1)
by(data.long, data.long$id, function(x){lines(x$time, x$Y)})

}
}

dev.off()}

# Trim Complete Data

lambda.func3 = function(data.arg,
varying.arg,
lambda.arg,
quant.arg)
{

data = data.arg[, varying.arg]

y <- quantile(data, probs = quant.arg, na.rm = TRUE)

res = list(y = y, rate = lambda.arg)

return(res)

}

# Trim Data
{
```

```
print("Simulation Started")
start_time <- proc.time()[[3]]

progress.ind <- 0

for(m in m.v){
for(n in n.v){

data.name <- paste0("m", m, "n", n, ".txt")
data <- read.table(paste0(complete.dir, data.name), header = TRUE)

progress.ind = progress.ind + 1
print("-----------------")
print(paste(progress.ind, "out of", length(m.v) * length(n.v) ))
print("-----------------")
print(paste("Data: ", data.name))

times <- c(0 : (n-1) )

for(mec in mec.v){

for(perc in perc.v){

all.data <- matrix(NA, nrow = m * nsim, ncol = n + 3 + 6)

for(sim in 1:nsim){

data.sim <- data[data$sim == sim, ]

rows <- 1:m + (sim-1)*m

lambdaf <- lambda.func3(data.arg = data.sim,
varying.arg = c(2 : (n+1) ),
lambda.arg = lambda.par,
quant.arg = quant.par) # quantiles

all.data[rows, ] <- as.matrix(trim(data.arg = data.sim,
varying.arg = c(2 : (n+1) ),
times.arg = times,
droptime.arg = times[2], # Time 1
perc.drop = perc,
seed.arg = seed + sim,
mechanism.arg = mec,
savedir.arg = NULL,
nsim.arg = 1,
summary.arg = summ,
lambdaf.arg = lambdaf
)
```

```
)

}

all.data <- as.data.frame(all.data)
names(all.data) <- c(names(data), "sim", "perc.drop", "perc.miss", "dur1", "dur2", "dtime")
all.data[, 1 + n + 3] <- NULL

# Save dt.wide into a .txt
write.table(all.data,
file = paste0(incomplete.dir,"m",m,"n",n,mec,"d",perc,".txt"),
sep = "\t",
row.names = FALSE,
col.names = TRUE)

}
}
}
}


print("Simulation Finished")
# Print all duration time
print(paste("Total simulation duration (s):",proc.time()[[3]] - start_time))
}

# Boxplot of trimming duration
{
m.times <- matrix(nrow = length(m.v) * length(n.v) * length(perc.v) * length(mec.v) * nsim, ncol = 5)
row <- 1
id <- 1

for(m in m.v){
for(n in n.v){
for(mec in mec.v){
for(perc in perc.v){

data  = read.table(paste0(incomplete.dir, "m", m, "n", n,mec,"d",perc,".txt"), header = TRUE)

m.times[row:(row + nsim - 1), c(1:4)] <- matrix(c(m, n, which(mec.v%in%mec), perc), nrow = nsim, ncol = 4, byrow = TRUE)

m.times[row:(row + nsim - 1), 5] <- as.vector(by(data, data$sim, function(x){x$dur1[1]+x$dur2[1]}))
```

```
310
311 row <- row + nsim
312
313 }
314 }
315 }
316 }
317
318 df.times <- as.data.frame(m.times)
319 names(df.times)[1:5] <- c("m", "n", "mec", "perc", "dur")
320 df.times$mec[df.times$mec == 1] <- "MCAR"
321 df.times$mec[df.times$mec == 2] <- "MAR"
322 df.times$mec[df.times$mec == 3] <- "MNAR"
323 df.times$mec <- as.factor(df.times$mec)
324 df.times$mec <- factor(df.times$mec, levels = levels(df.times$mec)[c(2
    ,1,3)])
325
326
327 {pdf(paste0(fig.dir,"C2trimtimes.pdf"), height = 20/2.54, width = 30/
    2.54)
328
329 nf <- layout(matrix(c(1, 1, 2, 3, 4, 5), 2, 3, byrow = TRUE), height =
     c(5, 5), TRUE)
330 #layout.show(nf)
331
332 hist(df.times$dur,
333 xlab = paste0("Tempo (s)"),
334 ylab = "Frequencia",
335 main = "")
336
337 boxplot(dur ~ mec, data = df.times,
338 xlab = "Tempo (s)",
339 ylab = "Mecanismo",
340 horizontal = TRUE)
341
342 boxplot(dur ~ perc, data = df.times,
343 xlab = "Tempo (s)",
344 ylab = "Proporcao de abandono",
345 horizontal = TRUE)
346
347 boxplot(dur ~ m, data = df.times,
348 xlab = "Tempo (s)",
349 ylab = "N. de individuos",
350 horizontal = TRUE)
351
352 boxplot(dur ~ n, data = df.times,
353 xlab = "Tempo (s)",
354 ylab = "N. de observacoes por individuo",
```

```
horizontal = TRUE)

dev.off()}
}

quantile(df.times$dur)

# Analyse Trimmed Data

# Fit trimmed data
{
print("Simulation Started")
start_time <- proc.time()[[3]]

n.row <- length(perc.v) * length(n.v) * length(m.v) * length(mec.v)
res1 <- matrix(NA, nrow = n.row, ncol = 4 + 4)
row <- 1

for(perc in perc.v){
for (n in n.v){

times <- 0 : (n - 1)
varying.arg <- 2 : (n + 1)

for(m in m.v){
for( mec in mec.v){

print( paste0 ("m", m, "n", n,mec ,"d", perc))

data <- read.table( paste0 (incomplete.dir, "m", m, "n", n,mec ,"d",
   perc,".txt"), header = TRUE)

res2 <- matrix(NA, nrow = nsim, ncol = 4)

for(sim in 1:nsim){

data.wide <- data[data$sim == sim, ]

data.long <- reshape(data.wide, direction = "long", varying = list(
   varying.arg), times = times)
names(data.long)[10] <- "Y"

fit <- nlme::lme(Y~time, random=~1|id, data = data.long[!is.na(data
   .long$Y),])

res2[sim, 1] <- coef(summary(fit))[1, 1] # Beta_0
res2[sim, 2] <- coef(summary(fit))[2, 1] # Beta_1
res2[sim ,3] <- as.numeric(nlme::VarCorr(fit)[1,1]) # nu^2
```

```
res2[sim ,4] <- sigma(fit)^2 # tau^2

}

betaX <- beta
betaX[2] <- beta[2]*(min(n.v)-1)/(n-1)

real.values = matrix(c(betaX[1], betaX[2], nu2, tau2), nrow = nsim,
    ncol = 4, byrow = TRUE)

error <- colMeans(abs(real.values - res2) / abs(real.values)) #
    percent error


res1[row, ] <- c(m, n, mec, perc, error)

row <- row + 1

}
}
}
}

res1 <- as.data.frame(res1)
names(res1) <- c("m","n","mec","perc","ebeta0","ebeta1","enu2","etau2"
    )

# Save dt.wide into a .txt
write.table(res1,
file = paste0(error.dir,"error",".txt"),
sep = "\t",
row.names = FALSE,
col.names = TRUE)

print("Simulation Finished")
# Print all duration time
print(paste("Total simulation duration (s):",proc.time()[[3]] -
    start_time))
}

# Estimates error
{
## Heat map
{
data  <- read.table(paste0(error.dir, "error.txt"), header = TRUE)
data$m <- as.factor(data$m)
data$n <- as.factor(data$n)

```

```
levels(data$mec)[levels(data$mec)=="mcar"]  <- "MCAR"
levels(data$mec)[levels(data$mec)=="mar"]   <- "MAR"
levels(data$mec)[levels(data$mec)=="mnar2"] <- "MNAR"

par.v <- c("ebeta0", "ebeta1", "enu2", "etau2")

for(par in par.v){

{pdf(paste0(fig.dir,"C2heatmap",par,".pdf"), height = 30/2.54, width
   = 30/2.54)

scale.factor <- 2 #10^scale.factor

par.pos <- which(names(data)%in%par)

if(par == "ebeta0"){title = bquote(beta[0])
} else if(par == "ebeta1") {title <- bquote(beta[1])
} else if(par == "enu2") {title <- bquote(nu^2)
} else if(par == "etau2") {title <- bquote(tau^2)}

hp <- ggplot2::ggplot(data, ggplot2::aes(x = m, y = n, fill = data[,
   par.pos] * 10 ^ scale.factor)) + ggplot2::facet_grid(perc ~ factor
   (mec, levels = c("MCAR", "MAR", "MNAR")))

hp <- hp + ggplot2::ggtitle(title) + ggplot2::theme(plot.title =
    ggplot2::element_text(hjust = 0.5)) # To center the title

hp <- hp + ggplot2::geom_raster()

hp <- hp + ggplot2::geom_text(ggplot2::aes(label = round(data[,
   par.pos] * 10 ^ scale.factor, 3)), size = 3, color = "white") # To
    add cell values

hp <- hp + ggplot2::theme(panel.background = ggplot2::element_rect(
   fill = "white"),
strip.background = ggplot2::element_blank(),
legend.position = "bottom")

hp <- hp + viridis::scale_fill_viridis()

hp <- hp + ggplot2::guides(fill = ggplot2::guide_colourbar(barwidth =
   40, barheight = 1, ticks = TRUE, title.position = "top",
   title.hjust = 0.5, title = c("Erro percentual medio (%)")))

print(hp)

dev.off()}
```

```
}
}
}

# Boxplots 1
{
## Save data
{
nrow <- length(perc.v) * length(n.v) * length(m.v) * length(mec.v) *
    nsim
ncol <- max(n.v) + 4

m.res <- matrix(NA, nrow = nrow, ncol = ncol)

rows <- 1:nsim

for(n in n.v){

varying.arg <- 2 : (n + 1)

for(m in m.v){

data.comp <- read.table( paste0 (complete.dir, "m", m, "n", n,".txt"),
    header = TRUE)

mean.comp <- do.call(rbind, by(data.comp,  data.comp$sim, function(x){
    colMeans(x[, varying.arg])}))

for( perc in perc.v){

for(mec in mec.v){

data.inc <- read.table( paste0 (incomplete.dir, "m", m, "n", n, mec,"d",
     perc,".txt"), header = TRUE)

mean.inc <- do.call(rbind, by(data.inc,  data.inc$sim, function(x){
    colMeans(x[, varying.arg], na.rm = TRUE)}))

m.res[rows, 1:4] <- matrix(c(m, n, perc, which(c("mcar","mar","mnar1",
    "mnar2")%in%mec)), ncol = 4, nrow = nsim, byrow = TRUE)

m.res[rows, (ncol - n + 1):ncol] <- mean.comp - mean.inc

rows <- rows + nsim
}

}

```

```
}
}


m.res <- as.data.frame(m.res)
names(m.res)[1:4] <- c("m", "n", "perc", "mec")

m.res$mec[m.res$mec == 1] <- "MCAR"
m.res$mec[m.res$mec == 2] <- "MAR"
m.res$mec[m.res$mec == 4] <- "MNAR"

m.res$m[m.res$m == m.v[1]] <- paste0("m=", m.v[1])
m.res$m[m.res$m == m.v[2]] <- paste0("m=", m.v[2])
m.res$m[m.res$m == m.v[3]] <- paste0("m=", m.v[3])
m.res$m[m.res$m == m.v[4]] <- paste0("m=", m.v[4])

m.res$n[m.res$n == n.v[1]]  <- paste0("n=", n.v[1])
m.res$n[m.res$n == n.v[2]]  <- paste0("n=", n.v[2])
m.res$n[m.res$n == n.v[3]]  <- paste0("n=", n.v[3])
m.res$n[m.res$n == n.v[4]]  <- paste0("n=", n.v[4])

}

## Plot all
{
m.res.long <- reshape(m.res, direction = "long", varying = list(5 : (
    max(n.v) + 4)), times = 0:(max(n.v) - 1))
names(m.res.long)[6] <- "mean.y"
m.res.long <- m.res.long[!is.na(m.res.long$mean.y), ]

for(perc in perc.v){

{pdf(paste0(fig.dir,"C2bpall",perc*100,".pdf"), height = 30/2.54,
    width = 30/2.54)

p.col <- length(m.v) # col number in plot
p.row <- length(n.v) # row number in plot
l <- 5 # number of the last time points to plot

data <- m.res.long[m.res.long$perc == perc & m.res.long$time >= (max(
    n.v) - l), ]

bp <- ggplot2::ggplot(data, ggplot2::aes(x = as.factor(time), y =
    mean.y)) + ggplot2::geom_boxplot(ggplot2::aes(fill = factor(mec,
    levels = c("MCAR", "MAR", "MNAR")))) + ggplot2::facet_grid(factor(
    n, levels = c("n=5", "n=15", "n=20", "n=50")) ~ factor(m, levels =
     c(paste0("m=", m.v[1]), paste0("m=", m.v[2]), paste0("m=", m.v[3]
    ), paste0("m=", m.v[4]))))
```

```
565
566 bp <- bp + ggplot2::theme(panel.background = ggplot2::element_rect(
    fill = "white"),
567 strip.background = ggplot2::element_blank(),
568 panel.grid.major.x = ggplot2::element_blank(),
569 panel.grid.major.y = ggplot2::element_line(linetype = "solid", colour
     = "grey"),
570 panel.border = ggplot2::element_rect(colour = "black", fill = NA),
571 legend.position = "bottom",
572 legend.key = ggplot2::element_rect(fill = "white", color = NA))
573
574 bp <- bp + ggplot2::ylab(bquote(bar(Y)[j]^Comp~-~bar(Y)[j]^Obs))
575
576 group.colors <- c(MCAR = "#31688EFF", MAR = "#35B779FF", MNAR ="#FDE72
    5FF")
577
578 bp <- bp + ggplot2::scale_fill_manual(values = group.colors, name = "
    Mecanismo")
579
580 bp <- bp + ggplot2::ggtitle(paste0(perc*100, "% Abandono")) + ggplot2:
    :theme(plot.title = ggplot2::element_text(hjust = 0.5)) # To
    center the title
581
582 bp <- bp + ggplot2::scale_x_discrete(name = "Tempo", labels=c(bquote(t
    [n-4]), bquote(t[n-3]), bquote(t[n-2]), bquote(t[n-1]), bquote(t[n
    ])))
583
584 print(bp)
585
586 dev.off()}
587
588 }
589 }
590
591 ## Plot one
592 {
593 perc <- 0.6
594 m <- 30
595 n <- 15
596
597 {pdf(paste0(fig.dir,"C2bpone.pdf"), height = 10/2.54, width = 30/2.54
    )
598
599 m.res.long <- reshape(m.res, direction = "long", varying = list(5 : (
    max(n.v) + 4)), times = 0:(max(n.v) - 1))
600 names(m.res.long)[6] <- "mean.y"
601 m.res.long <- m.res.long[!is.na(m.res.long$mean.y), ]
602
```

```
603 data <- m.res.long[m.res.long$m == paste0("m=",m) & m.res.long$n ==
        paste0("n=",n) & m.res.long$perc == perc, ]
604
605 bp <- ggplot2::ggplot(data, ggplot2::aes(x = as.factor(time), y =
        mean.y)) + ggplot2::geom_boxplot(ggplot2::aes(fill = factor(mec,
        levels = c("MCAR", "MAR", "MNAR")))) + ggplot2::facet_grid(n ~ m)
606
607 group.colors <- c(MCAR = "#31688EFF", MAR = "#35B779FF", MNAR ="#FDE72
        5FF")
608
609 bp <- bp + ggplot2::scale_fill_manual(values = group.colors, name = "
        Mecanismo")
610
611 bp <- bp + ggplot2::theme(panel.background = ggplot2::element_rect(
        fill = "white"),
612 strip.background = ggplot2::element_blank(),
613 panel.grid.major.x = ggplot2::element_blank(),
614 panel.grid.major.y = ggplot2::element_line(linetype = "solid", colour
        = "grey"),
615 panel.border = ggplot2::element_rect(colour = "black", fill = NA),
616 legend.position = "bottom",
617 legend.key = ggplot2::element_rect(fill = "white", color = NA))
618
619 bp <- bp + ggplot2::ylab(bquote(bar(Y)[j]^Comp~-~bar(Y)[j]^Obs))
620
621 bp <- bp + ggplot2::scale_x_discrete(name = "Tempo", labels=c(0:(n-1))
        )
622
623 print(bp)
624
625 dev.off()}
626 }
627
628 ## Plot across %Dropout
629 {
630 m <- 30
631 n <- 5
632
633 {pdf(paste0(fig.dir,"C2bpperc.pdf"), height = 10/2.54, width = 25/2.5
        4)
634
635 m.res.long <- reshape(m.res, direction = "long", varying = list(5 : (
        max(n.v) + 4)), times = 0:(max(n.v) - 1))
636 names(m.res.long)[6] <- "mean.y"
637 m.res.long <- m.res.long[!is.na(m.res.long$mean.y), ]
638
639 data <- m.res.long[m.res.long$m == paste0("m=",m) & m.res.long$n ==
        paste0("n=",n), ]
```

```
640
641 bp <- ggplot2::ggplot(data, ggplot2::aes(x = as.factor(time), y =
      mean.y)) + ggplot2::geom_boxplot( ggplot2::aes(fill = factor(mec,
      levels = c("MCAR", "MAR", "MNAR")))) + ggplot2::facet_grid(m ~
      perc)
642
643 bp <- bp + ggplot2::theme(panel.background = ggplot2::element_rect(
      fill = "white"),
644 strip.background = ggplot2::element_blank(),
645 panel.grid.major.x = ggplot2::element_blank(),
646 panel.grid.major.y = ggplot2::element_line(linetype = "solid", colour
       = "grey"),
647 panel.border = ggplot2::element_rect(colour = "black", fill = NA),
648 legend.position = "bottom",
649 legend.key = ggplot2::element_rect(fill = "white", color = NA))
650
651 bp <- bp + ggplot2::ylab(bquote(bar(Y)[j]^Comp~-~bar(Y)[j]^Obs))
652
653 group.colors <- c(MCAR = "#31688EFF", MAR = "#35B779FF", MNAR ="#FDE72
      5FF")
654
655 bp <- bp + ggplot2::scale_fill_manual(values = group.colors, name = "
      Mecanismo")
656
657 bp <- bp + ggplot2::scale_x_discrete(name = "Tempo", labels=c(0:(n-1))
      )
658
659 print(bp)
660
661 dev.off()}
662 }
663 }
664
665 # Boxplots 2
666 {
667 ## Save data
668 {
669 nrow <- length(perc.v) * length(n.v) * length(m.v) * length(mec.v) *
      nsim
670 ncol <- max(n.v) - 1 + 4
671
672 m.res3 <- matrix(NA, nrow = nrow, ncol = ncol)
673
674 rows <- 1:nsim
675
676 for(n in n.v){
677
678 for(m in m.v){
```

```
679
680 for( perc in perc.v){
681
682 for(mec in mec.v){
683
684 data <- read.table( paste0 (incomplete.dir, "m", m, "n", n, mec,"d",
      perc,".txt"), header = TRUE)
685
686 data.comp <- data[is.na(data$dtime), ] # Completers
687 data.drop <- data[!is.na(data$dtime), ] # Dropouts
688
689 fake.id <- c(0, c(rep(99, n-1), NA), 0, nsim+1, 0, 0, 0, 0, n-1) # To
      ensure that all possible time distances are considered in the wide
       format
690 data.drop <- rbind(fake.id, data.drop)
691
692 data.drop$id2 <- data.drop$id
693 data.drop$id <- 1:nrow(data.drop) # Otherwise non-unique id to reshape
       into long format
694 data.comp$id2 <- data.comp$id
695 data.comp$id <- 1:nrow(data.comp)
696
697 data.drop.long <- reshape(data.drop, direction = "long", varying =
      list(2 : (n + 1)), times = 0:(n - 1))
698 data.comp.long <- reshape(data.comp, direction = "long", varying =
      list(2 : (n + 1)), times = 0:(n - 1))
699
700 names(data.drop.long)[11] <- "Y"
701 names(data.comp.long)[11] <- "Y"
702
703 data.drop.long <- data.drop.long[!is.na(data.drop.long$Y), ]
704
705 data.drop.long$timetodrop <- data.drop.long$time - data.drop.long$d
      time
706 data.comp.long$timetodrop <- data.comp.long$time - (n - 1)
707 data.comp.long <- data.comp.long[data.comp.long$timetodrop != 0, ]
708
709 data.drop.long <- data.drop.long[order(data.drop.long$timetodrop), ] #
       To ensure that the times appear in order, from tmin to tmax, when
       reshaping to wide
710 data.comp.long <- data.comp.long[order(data.comp.long$timetodrop), ]
711
712 data.drop.long$time <- NULL
713 data.comp.long$time <- NULL
714
715 data.drop.wide <- reshape(data.drop.long, direction = "wide", timevar
      = "timetodrop", idvar = "id", v.names = "Y")
716 data.comp.wide <- reshape(data.comp.long, direction = "wide", timevar
```

```
    = "timetodrop", idvar = "id", v.names = "Y")
717
718 data.drop.wide$id <- NULL
719 names(data.drop.wide)[8] <- "id"
720 data.comp.wide$id <- NULL
721 names(data.comp.wide)[8] <- "id"
722
723 data.drop.wide <- data.drop.wide[!data.drop.wide$id == 0, ]
724
725 m.res3[rows, 1:4] <- matrix(c(m, n, perc, which(c("mcar","mar","mnar1"
    ,"mnar2")%in%mec)), ncol = 4, nrow = nsim, byrow = TRUE)
726
727 varying.arg <- (ncol(data.drop.wide) - (n - 1) + 1) : ncol(data
    .drop.wide)
728
729 mean.drop <- do.call(rbind, by(data.drop.wide,  data.drop.wide$sim,
    function(x){colMeans(x[, varying.arg], na.rm = TRUE)}))
730 mean.comp <- do.call(rbind, by(data.comp.wide,  data.comp.wide$sim,
    function(x){colMeans(x[, varying.arg], na.rm = TRUE)}))
731
732 m.res3[rows, (ncol - (n - 1) + 1):ncol] <- mean.drop - mean.comp
733
734 rows <- rows + nsim
735
736 }
737
738 }
739 }
740 }
741
742
743 m.res3 = as.data.frame(m.res3)
744 names(m.res3)[1:4] <- c("m", "n", "perc", "mec")
745
746 m.res3$mec[m.res3$mec == 1] <- "MCAR"
747 m.res3$mec[m.res3$mec == 2] <- "MAR"
748 m.res3$mec[m.res3$mec == 4] <- "MNAR"
749
750 m.res3$m[m.res3$m == m.v[1]] <- paste0("m=", m.v[1])
751 m.res3$m[m.res3$m == m.v[2]] <- paste0("m=", m.v[2])
752 m.res3$m[m.res3$m == m.v[3]] <- paste0("m=", m.v[3])
753 m.res3$m[m.res3$m == m.v[4]] <- paste0("m=", m.v[4])
754
755 m.res3$n[m.res3$n == 5]   <- "n=5"
756 m.res3$n[m.res3$n == 15]  <- "n=15"
757 m.res3$n[m.res3$n == 20]  <- "n=20"
758 m.res3$n[m.res3$n == 50]  <- "n=50"
759
```

```
}

## Plot all
{
m.res3.long <- reshape(m.res3, direction = "long", varying = list(5:ncol(m.res3)), times = -(max(n.v)-1):-1)
names(m.res3.long)[6] <- "mean.y"
m.res3.long <- m.res3.long[!is.na(m.res3.long$mean.y), ]

for(perc in perc.v){

{pdf(paste0(fig.dir,"C2bptimetodropall2",perc*100,".pdf"), height = 30/2.54, width = 30/2.54)

p.col <- length(m.v) # col number in plot
p.row <- length(n.v) # row number in plot
l <- 4 # number of the last time points to plot

data <- m.res3.long[m.res3.long$perc == perc & m.res3.long$time >= tail(-(max(n.v)-1):-1, l)[1], ]

bp <- ggplot2::ggplot(data, ggplot2::aes(x = as.factor(time), y = mean.y)) + ggplot2::geom_boxplot(ggplot2::aes(fill = factor(mec, levels = c("MCAR", "MAR", "MNAR")))) + ggplot2::facet_grid(factor(n, levels = c("n=5", "n=15", "n=20", "n=50")) ~ factor(m, levels = c(paste0("m=", m.v[1]), paste0("m=", m.v[2]), paste0("m=", m.v[3]), paste0("m=", m.v[4]))))

bp <- bp + ggplot2::theme(panel.background = ggplot2::element_rect(fill = "white"),
strip.background = ggplot2::element_blank(),
panel.grid.major.x = ggplot2::element_blank(),
panel.grid.major.y = ggplot2::element_line(linetype = "solid", colour = "grey"),
panel.border = ggplot2::element_rect(colour = "black", fill = NA),
legend.position = "bottom",
legend.key = ggplot2::element_rect(fill = "white", color = NA))

bp <- bp + ggplot2::ylab(bquote(bar(Y)[k]^Aban~-~bar(Y)[k]^Comp))

group.colors <- c(MCAR = "#31688EFF", MAR = "#35B779FF", MNAR ="#FDE725FF")

bp <- bp + ggplot2::scale_fill_manual(values = group.colors, name = "Mecanismo")

bp <- bp + ggplot2::ggtitle(paste0(perc*100, "% Abandono")) + ggplot2::theme(plot.title = ggplot2::element_text(hjust = 0.5)) # To
```

```
    center the title
795
796 bp <- bp + ggplot2::scale_x_discrete(name = "Tempo ate evento")
797
798 print(bp)
799
800 dev.off()}
801
802 }
803 }
804
805 ## Plot one
806 {
807 perc <- 0.6
808 m <- 30
809 n <- 15
810
811 {pdf( paste0 (fig.dir,"C2bptimetodropone2.pdf"), height = 10/2.54,
    width = 30/2.54)
812
813 m.res3.long <- reshape(m.res3, direction = "long", varying = list(5 :
    ncol(m.res3)), times = -(max(n.v)-1):-1)
814 names(m.res3.long)[6] <- "mean.y"
815 m.res3.long <- m.res3.long[!is.na(m.res3.long$mean.y), ]
816
817 data <- m.res3.long[m.res3.long$m == paste0("m=",m) & m.res3.long$n ==
     paste0("n=",n) & m.res3.long$perc == perc, ]
818
819 bp <- ggplot2::ggplot(data, ggplot2::aes(x = as.factor(time), y =
    mean.y)) + ggplot2::geom_boxplot( ggplot2::aes(fill = factor(mec,
    levels = c("MCAR", "MAR", "MNAR")))) + ggplot2::facet_grid(factor(
    n, levels = c("n=5", "n=15", "n=20", "n=50")) ~ factor(m, levels =
     c( paste0("m=", m.v[1]), paste0("m=", m.v[2]), paste0("m=", m.v[3]
    ), paste0("m=", m.v[4]))))
820
821 bp <- bp + ggplot2::theme(panel.background = ggplot2::element_rect(
    fill = "white"),
822 strip.background = ggplot2::element_blank(),
823 panel.grid.major.x = ggplot2::element_blank(),
824 panel.grid.major.y = ggplot2::element_line(linetype = "solid", colour
     = "grey"),
825 panel.border = ggplot2::element_rect(colour = "black", fill = NA),
826 legend.position = "bottom",
827 legend.key = ggplot2::element_rect(fill = "white", color = NA))
828
829 bp <- bp + ggplot2::ylab(bquote(bar(Y)[k]^Aban~-~bar(Y)[k]^Comp))
830
831 group.colors <- c(MCAR = "#31688EFF", MAR = "#35B779FF", MNAR ="#FDE72
```

```
    5FF")
832
833 bp <- bp + ggplot2::scale_fill_manual(values = group.colors, name = "
    Mecanismo")
834
835 bp <- bp + ggplot2::scale_x_discrete(name = "Tempo ate evento")
836
837 print(bp)
838
839 dev.off()}
840 }
841
842 }
843
844 # Plot % of missing observations
845 {
846 for(perc in perc.v){
847
848 {pdf(paste0(fig.dir,"C2percmissobs",perc*100,".pdf"), height = 30/2.5
    4, width = 30/2.54)
849
850 par(mfcol = c(length(n.v), length(m.v)), oma = c(0, 0, 2, 0))
851
852 for(m in m.v){ # Column
853
854 for(n in n.v){ # Row
855
856 data.mcar  <- read.table(paste0(incomplete.dir, "m", m, "n", n,"mcar",
    "d",perc,".txt"), header = TRUE)
857 data.mar   <- read.table(paste0(incomplete.dir, "m", m, "n", n,"mar","
    d",perc,".txt"), header = TRUE)
858 data.mnar  <- read.table(paste0(incomplete.dir, "m", m, "n", n,"mnar2"
    ,"d",perc,".txt"), header = TRUE)
859
860 mcar.perc <- as.vector(by(data.mcar, data.mcar$sim, function(x){
    x$perc.miss[1]}))
861 mar.perc  <- as.vector(by(data.mar, data.mar$sim, function(x){
    x$perc.miss[1]}))
862 mnar.perc <- as.vector(by(data.mnar, data.mnar$sim, function(x){
    x$perc.miss[1]}))
863
864 perc.df <- data.frame(perc = c(mcar.perc, mar.perc, mnar.perc),
865 mec = rep(c("MCAR","MAR","MNAR"), each = nsim))
866
867
868 boxplot(perc.df$perc ~ perc.df$mec,
869 xlab = "Mecanismo de Omissao",
870 ylab = "Prop. Observacoes Omissas"
```

```
)

title(paste0("m=", m, " n=", n), font.main = 1)

}

}

mtext(paste0("Abandono ", perc*100, "%"), outer = TRUE, cex = 1,
    font.main = 1)

dev.off()}

}
}

# Plot Evolution of % Missing Information & % Dropout
{

for(perc in perc.v){

{pdf(paste0(fig.dir,"C2missevolu",perc*100,".pdf"), height = 30/2.54,
     width = 30/2.54)

par(oma = c(0, 0, 2, 0))
layout(matrix(c(1:16, 19, 17, 18, 20), length(n.v)+1, length(m.v), by
    row = TRUE), height = c(rep((30-1)/4,4), 1.5)/2.54, TRUE)

for(n in n.v){ # Row

for(m in m.v){ # Column

data.mcar <- read.table(paste0(incomplete.dir, "m", m, "n", n, "mcar"
    ,"d", perc,".txt"), header = TRUE)
data.mar  <- read.table(paste0(incomplete.dir, "m", m, "n", n, "mar" ,
    "d", perc,".txt"), header = TRUE)
data.mnar <- read.table(paste0(incomplete.dir, "m", m, "n", n, "mnar2"
     ,"d", perc,".txt"), header = TRUE)

times <- 0:(n-1)
varying.arg <- 2:(n+1)

M <- m * nsim

N.v <- m * 1:n * nsim
N.v <- sapply(1:n, function(x){sum(N.v[1:x])})

na.mcar   <- apply(data.mcar[,varying.arg],  2, function(x){sum(is.na(
```

```
    x))})
drop.mcar <- na.mcar / M
perc.mcar <- sapply(1:n, function(x){sum(na.mcar[1:x])}) / N.v

na.mar   <- apply(data.mar[,varying.arg],  2, function(x){sum(is.na(x)
    )})
drop.mar <- na.mar / M
perc.mar <- sapply(1:n, function(x){sum(na.mar[1:x])}) / N.v

na.mnar   <- apply(data.mnar[,varying.arg],  2, function(x){sum(is.na(
    x))})
drop.mnar <- na.mnar / M
perc.mnar <- sapply(1:n, function(x){sum(na.mnar[1:x])}) / N.v

par(mar = c(5,4,4,2+3)) # To add some extra space to the right to add
     the new y axis

ylim1 <- range(c(drop.mcar, drop.mar, drop.mnar))

plot(0, xlim = range(times), ylim = ylim1, cex = 0, yaxt = "n", xlab
    = "Tempo", ylab = " ", bty = "n")
title(paste0("m=", m, " n=", n), font.main = 1)

col1 <- c("#31688EFF") # Blue
col2 <- c("#35B779FF") # Green

lines(times, drop.mcar, lwd = 1, col = col1)
lines(times, drop.mar,  lwd = 1, col = col1)
lines(times, drop.mnar, lwd = 1, col = col1)

points(times, drop.mcar, pch = 4, col = col1)
points(times, drop.mcar, pch = 3, col = col1)
points(times, drop.mcar, pch = 21, col = col1)

axis(side = 2)
mtext(side = 2, line = 3, "Prop. abandono", cex=0.6)

par(new = TRUE)

ylim2 <- range(c(perc.mcar, perc.mar, perc.mnar))

plot(0, xlim = range(times), ylim = ylim2, type = "l", xaxt = "n",
    yaxt = "n", xlab ="", ylab = "",  bty = "n")

lines(times, perc.mcar,  lwd = 1, type = "b", pch = 4,  lty = 2, col
    = col2)
lines(times, perc.mar,   lwd = 1, type = "b", pch = 3,  lty = 2, col
    = col2)
```

```
953 lines(times, perc.mnar,  lwd = 1, type = "b", pch = 21, lty = 2, col
      = col2)
954
955
956
957 axis(side = 4)
958 mtext(side = 4, line = 3, "Prop. dados omissos", cex=0.6)
959
960 }
961
962 }
963
964 # Legend 1
965 par(mar = c(0, 4.1, 0, 2.1)) # (Bottom, Left, Top, Right)
966 plot(0, 0, cex = 0, axes = F, ann = FALSE)
967 legend("top",
968 horiz = FALSE,
969 bty = "n",
970 lty = c(2, 2, 2),
971 legend = c("Prop. dados omissos MCAR", "Prop. dados omissos MAR", "
      Prop. dados omissos MNAR"),
972 col = rep(col2, 3),
973 pch = c(4, 3, 21),
974 pt.cex = 2,
975 seg.len = 3
976 )
977
978 # Legend 2
979
980 plot(0, 0, cex = 0, axes = F, ann = FALSE)
981 legend("top",
982 horiz = FALSE,
983 bty = "n",
984 lty = c(1, 1, 1),
985 legend = c("Prop. abandono MCAR", "Prop. abandono MAR", "Prop.
      abansono MNAR"),
986 col = rep(col1, 3),
987 pch = c(4, 3, 21),
988 pt.cex = 2,
989 seg.len = 3
990 )
991
992 mtext(paste0("Abandono ", perc*100, "%"), outer = TRUE, cex = 1,
      font.main = 1)
993
994 dev.off()}
995
996 }
```

```
997
998 }
999
1000 # Other Plots
1001
1002 {
1003 # Plot Gaussian Weights
1004 {
1005 x1 <- seq(-3, 0, length = 100)
1006 x2 <- seq(0,  3, length = 100)
1007
1008 {pdf( paste0 (fig.dir,"C2weightsplot.pdf"), height = 20/2.54, width = 3
      0/2.54)
1009
1010 {
1011
1012 par(mar = c(5.1, 5.1 , 4.1, 2.1))
1013
1014 plot(x1, dnorm(x1), type="l", lty=1, xlim=range(c(x1,x2)), ylim=c(0,
      dnorm(0)*1.3), xaxt="n", yaxt="n", xlab="Tempo", ylab="")
1015
1016 title(ylab="Pesos (nao normalizados)",line = 4)
1017
1018 axis(1,at=c(-2,-1.5,-1,-0.5,0),labels=c(parse(text='t[i1]'),parse(
      text='t[i2]'),parse(text='...'),parse(text='t[i(j-2)]'),parse(text
      ='t[i(j-1)]')))
1019
1020 axis(2,at=c(dnorm(-2),dnorm(-1.5),dnorm(-0.5),dnorm(0)),labels=c(
      parse(text='w[i0]'),parse(text='w[i1]'),parse(text='w[i(j-1)]'),
      parse(text='w[ij]')),las=1)
1021
1022 axis(3,at=c(0,-2),labels=expression(mu,paste(mu-3,sigma)))
1023 abline(v=c(-2,0),lty=1,lwd=1,col="grey")
1024
1025
1026 lines(x2, dnorm(x2), lty = 2, col = "grey")
1027
1028 # segments(x0, y0, x1, y1)
1029 # Horizontal segments
1030 segments( x0 = min(x1)-10, y0 = dnorm(-2), x1 = -2, y1 = dnorm(-2),
      lty = 2)
1031 segments( x0 = min(x1)-1, y0 = dnorm(-1.5), x1 = -1.5, y1 = dnorm(-1.
      5), lty = 2)
1032 segments( x0 = min(x1)-1, y0 = dnorm(-0.5), x1 = -0.5, y1 = dnorm(-0.
      5), lty = 2)
1033 segments( x0 = min(x1)-1, y0 = dnorm(0), x1 = 0, y1 = dnorm(0), lty =
       2)
1034
```

```
# Vertical segments
segments( x0 = 0, y0 = 0, x1 = 0, y1 = dnorm(0), lty = 2)
segments( x0 = -0.5, y0 = 0, x1 = -0.5, y1 = dnorm(-0.5), lty = 2)
segments( x0 = -1.5, y0 = 0, x1 = -1.5, y1 = dnorm(-1.5), lty = 2)
segments( x0 = -2, y0 = 0, x1 = -2, y1 = dnorm(-2), lty = 2)

# arrows(xHigh,yHigh,xLow,yLow,...)

arrows(0, dnorm(0)*1.2, -2, dnorm(0)*1.2, angle = 90, length = 0.05,
    code = 3)
text(-1,dnorm(0)*1.2,"99.73%",pos=3)

text(0,dnorm(0)*1.1,expression(paste("N(",mu,",",sigma^2,")")),cex=1)

#text(1,0.35,expression(paste(mu==t[ij])),cex=1)
text(2,0.35,expression(paste(sigma==over(t[i(j-1)] - t[i1], 3))),cex=
    2)
text(2,0.15,expression(paste(w[ij]^"'"==over(w[ij], sum(w[ik], k==0,
    j)))),cex=2)

}

dev.off()}

}

# Lambda plot examples
{
# 1

{pdf( paste0 (fig.dir,"C2lambdaplot1.pdf"), height = 10/2.54, width = 1
    0/2.54)

lambdaf1 <- list(y = c(0, 10), rate = c(0.1, 0.9))

plot(x = lambdaf1$y, y = rev(lambdaf1$rate), ylim = c(0.8*min(lambdaf
    1$rate), 1.2*max(lambdaf1$rate)),
type = "l",
xlab = bquote(F[2](bold(y)[ij])), ylab = bquote(lambda[ij]))
points(x = lambdaf1$y, y = rev(lambdaf1$rate), pch = 20)

dev.off()}

# 2

{pdf( paste0 (fig.dir,"C2lambdaplot2.pdf"), height = 10/2.54, width = 1
    0/2.54)

```

```
lambdaf2 <- list(y = c(10, 30, 50, 100), rate = c(10, 10, 30, 30))

plot(x = lambdaf2$y, y = lambdaf2$rate, type = "l", ylim = c(0.8*min(
   lambdaf2$rate), 1.2*max(lambdaf2$rate)),
xlab = bquote(F[2](bold(y)[ij])), ylab = bquote(lambda[ij]))
points(x = lambdaf2$y, y = lambdaf2$rate, pch = 20)

dev.off()}

# 3

lambdaf3 <- lambda.func2(data.arg = matrix(c(0:10,10)+10, nrow = 2),
varying.arg = c(1:5),
lambda.range.arg = c(5, 50),
evolution = "exponential",
sign.arg = "inverse",
nodes = 5)

{pdf(paste0(fig.dir,"C2lambdaplot3.pdf"), height = 10/2.54, width = 1
   0/2.54)

plot(x = lambdaf3$y, y = lambdaf3$rate, type = "l", ylim = c(0.8*min(
   lambdaf3$rate), 1.2*max(lambdaf3$rate)),
xlab = bquote(F[2](bold(y)[ij])), ylab = bquote(lambda[ij]))
points(x = lambdaf3$y, y = lambdaf3$rate, pch = 20)

dev.off()}
}
}

print(paste("Total duration (min):", (proc.time()[[3]] - start_time
   All)/60 ))
```

Anexo B

Gráficos complementares

B.1 Estudo de simulação

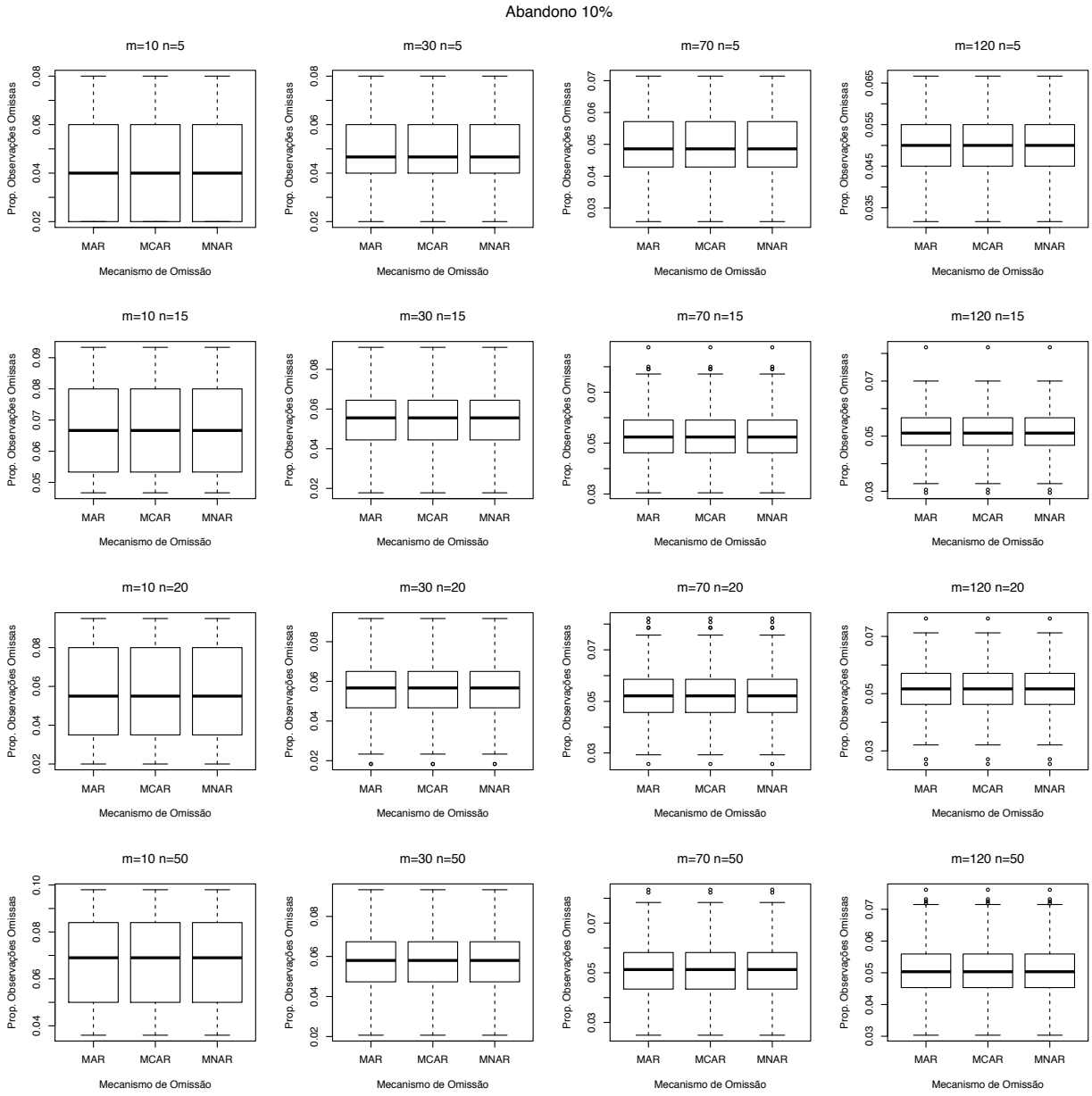


Figura B.1: Gráfico de caixa da proporção de observações omissas em cada um dos cenários em função do mecanismo de omissão, na presença de 10% de abandono dos indivíduos.

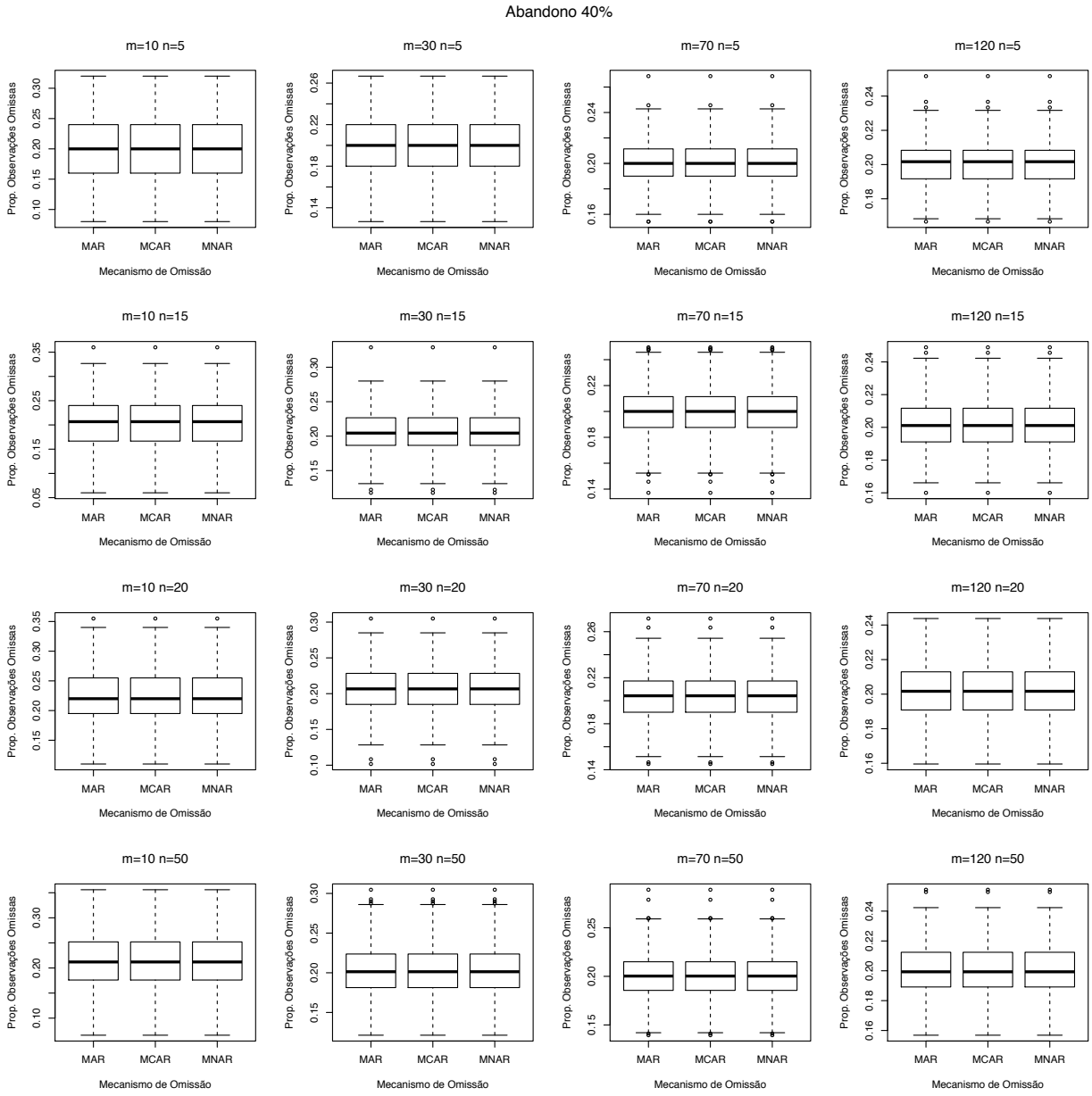


Figura B.2: Gráfico de caixa da proporção de observações omissas em cada um dos cenários em função do mecanismo de omissão, na presença de 40% de abandono dos indivíduos.

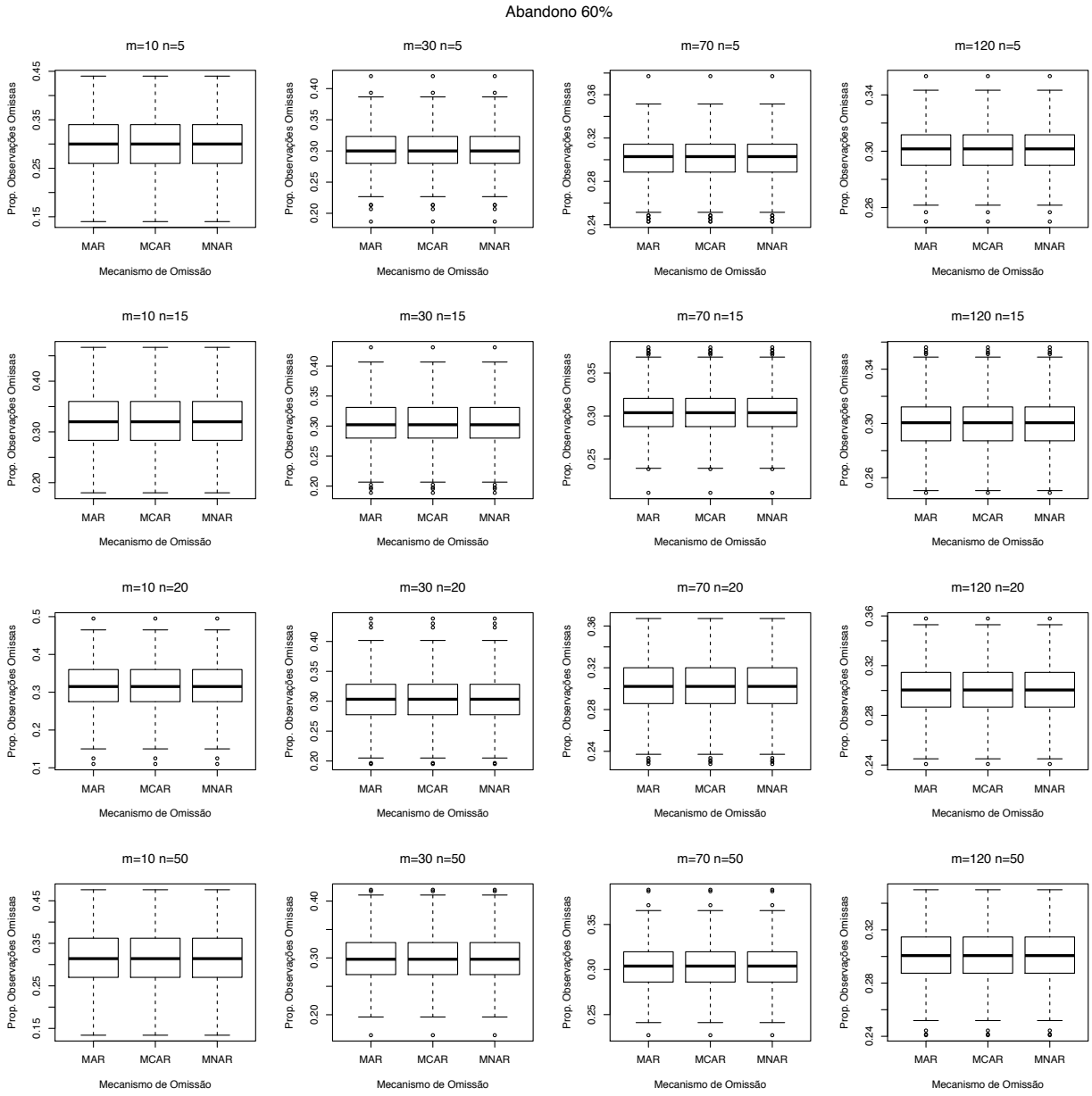


Figura B.3: Gráfico de caixa da proporção de observações omissas em cada um dos cenários em função do mecanismo de omissão, na presença de 60% de abandono dos indivíduos.

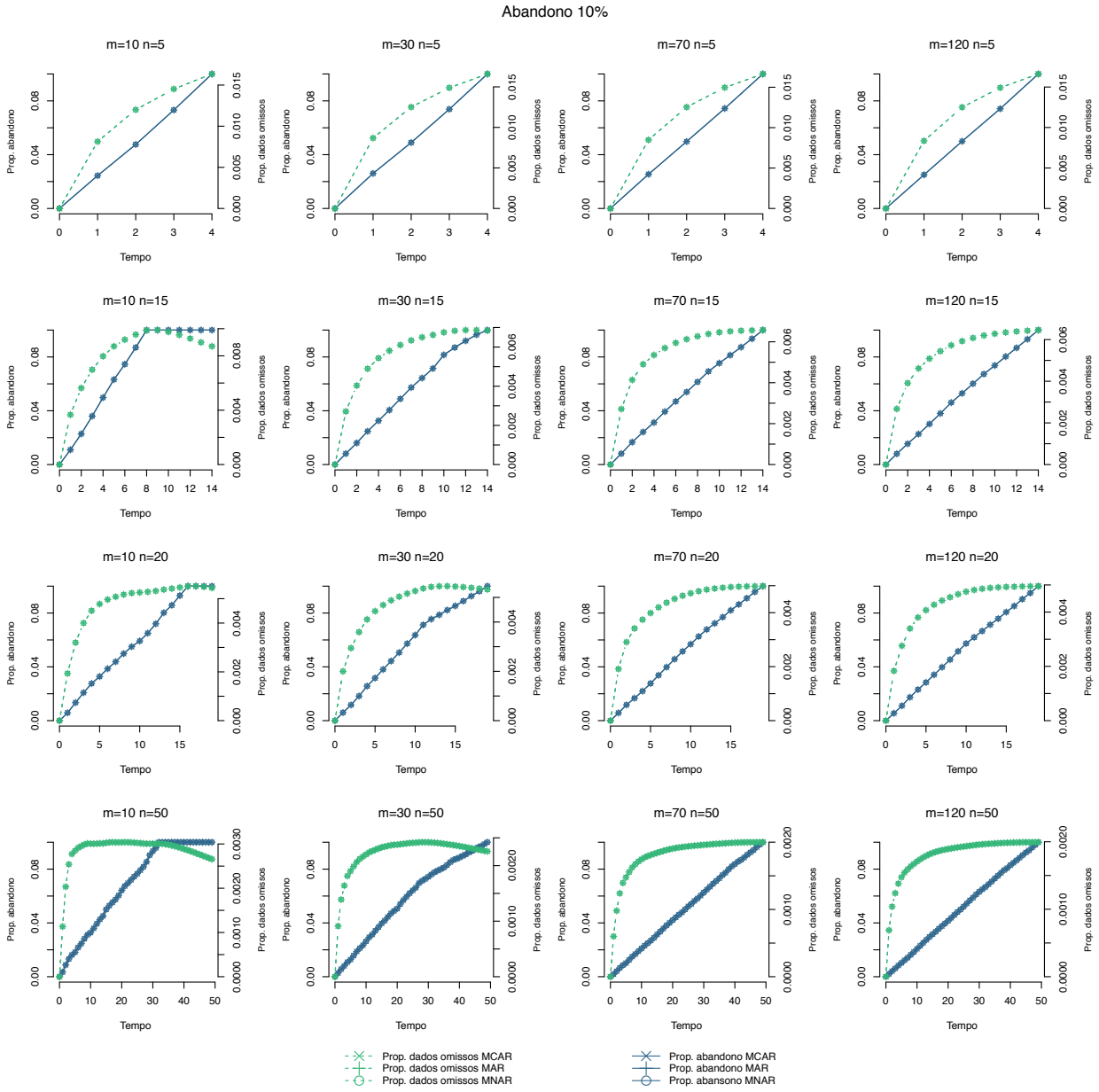


Figura B.4: Evolução da proporção de dados omissos e da proporção de indivíduos que abandonam o estudo em função do tempo, na presença de 10% de abandono dos indivíduos.

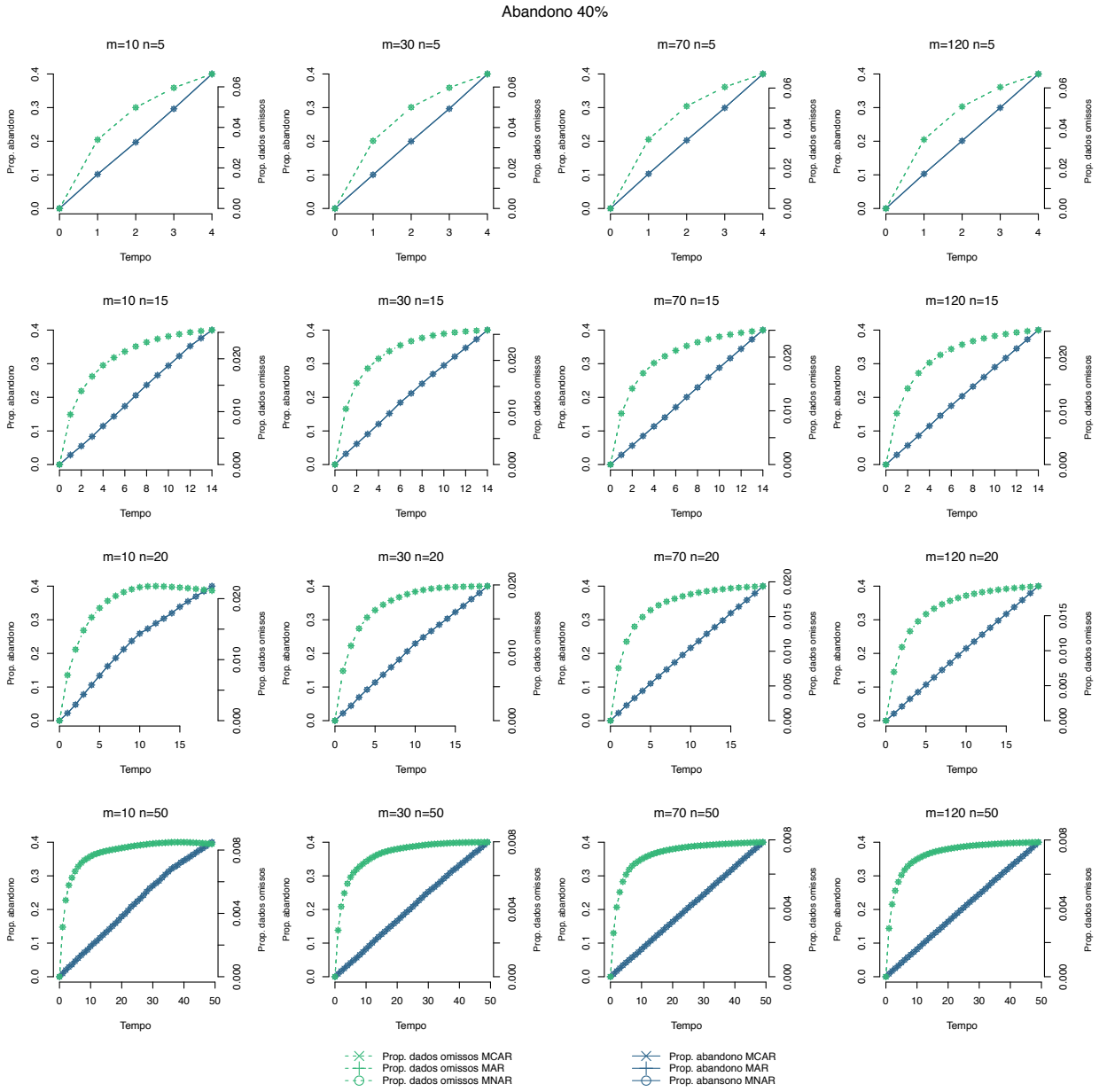


Figura B.5: Evolução da proporção de dados omissos e da proporção indivíduos que abandonam o estudo em função do tempo, na presença de 40% de abandono dos indivíduos.

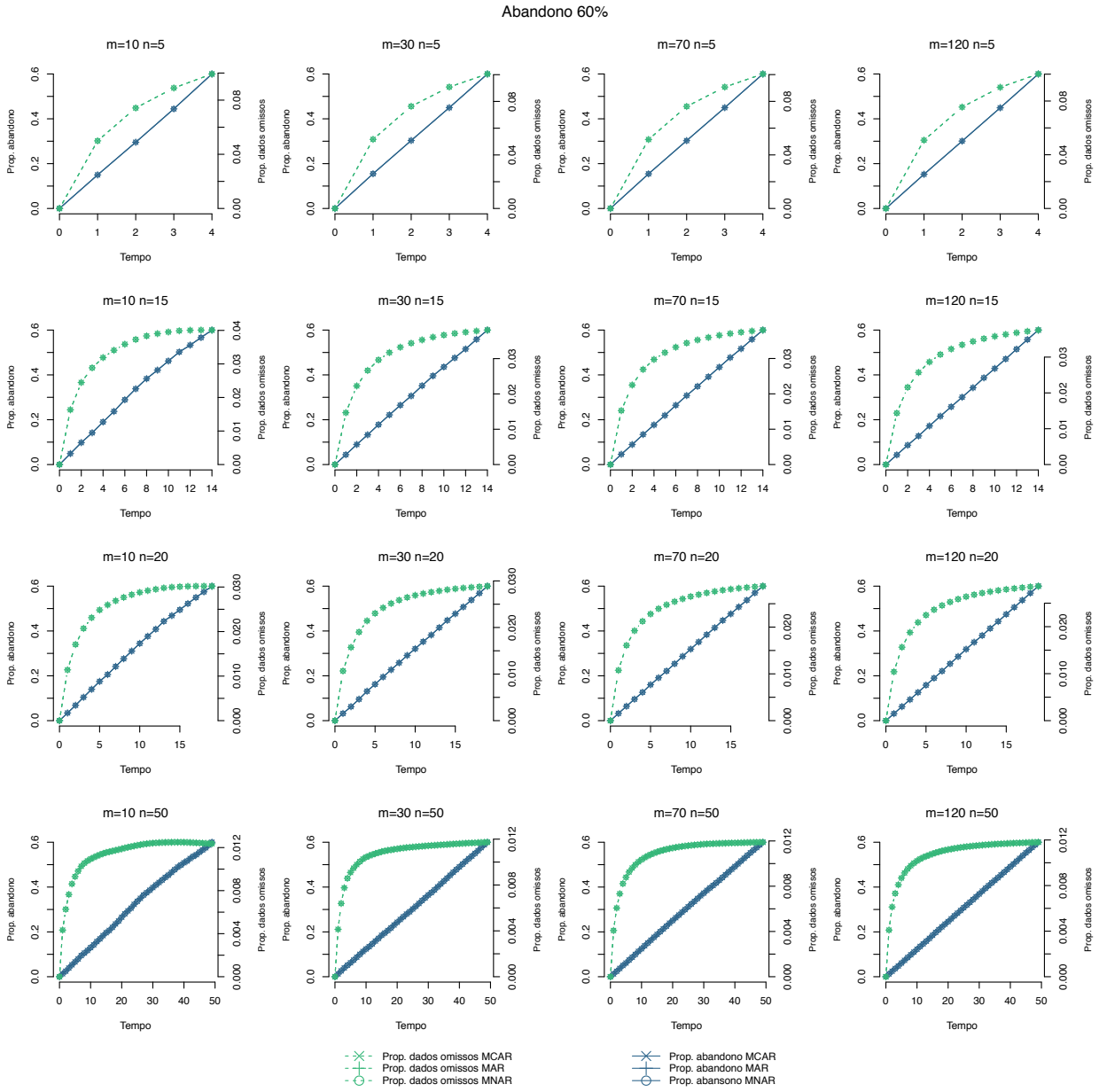


Figura B.6: Evolução da proporção de dados omissos e da proporção indivíduos que abandonam o estudo em função do tempo, na presença de 60% de abandono dos indivíduos.

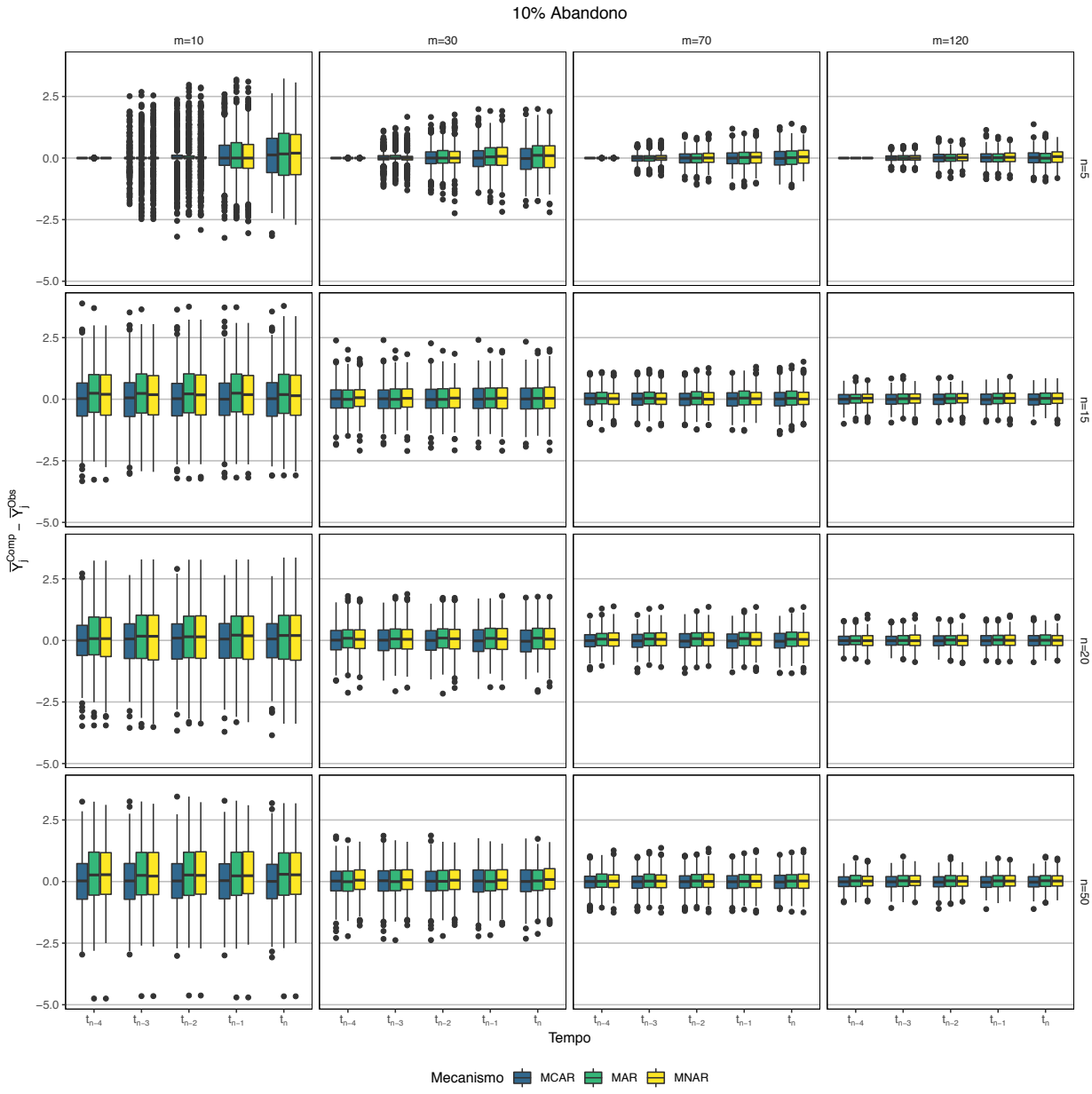


Figura B.7: Diagramas de caixa da diferença da média de $\boldsymbol{Y}_j$, no tempo j, entre os dados completos e os dados observados nas 500 simulações, nas últimas 5 ocasiões de observação, e na presença de 10% de abandono dos indivíduos.

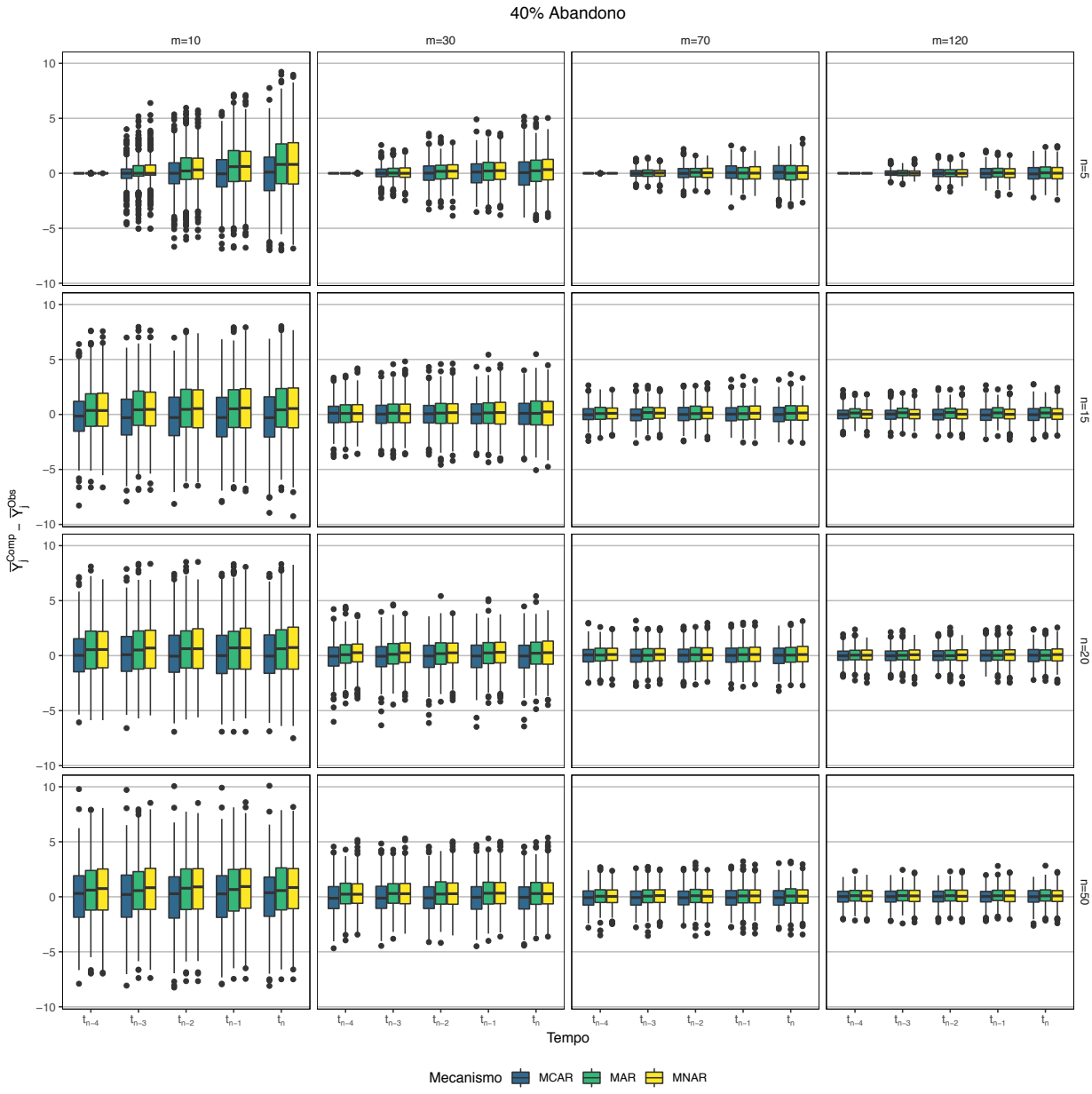


Figura B.8: Diagramas de caixa da diferença da média de $\boldsymbol{Y}_j$, no tempo j, entre os dados completos e os dados observados nas 500 simulações, nas últimas 5 ocasiões de observação, e na presença de 40% de abandono dos indivíduos.

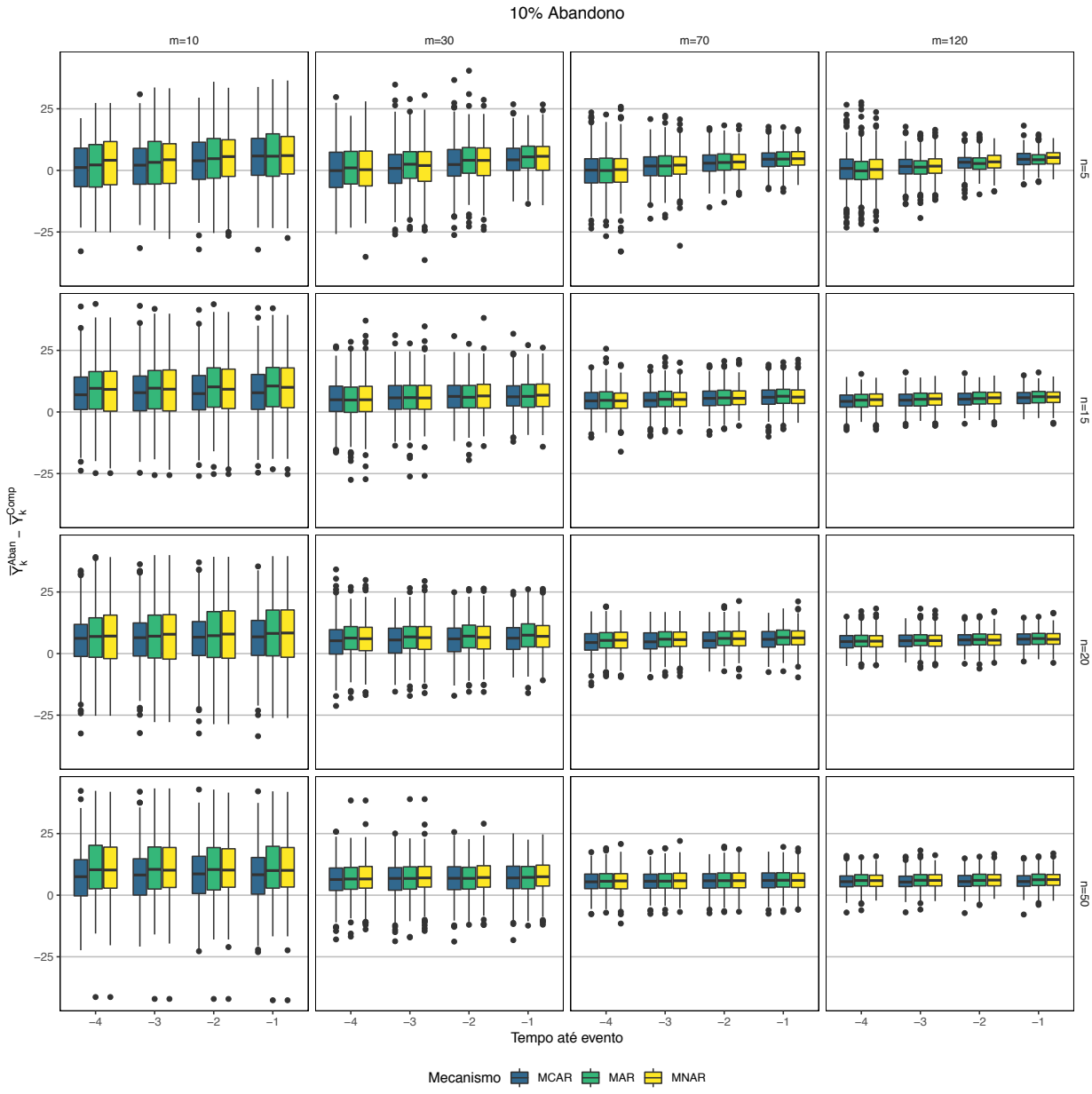


Figura B.9: Diagramas de caixa da diferença da média de $\boldsymbol{Y}_k$, no tempo até evento k, entre os indivíduos que completaram e os que abandonaram o estudo nas 500 simulações, nos últimos 4 instantes, e na presença de 10% de abandono dos indivíduos.

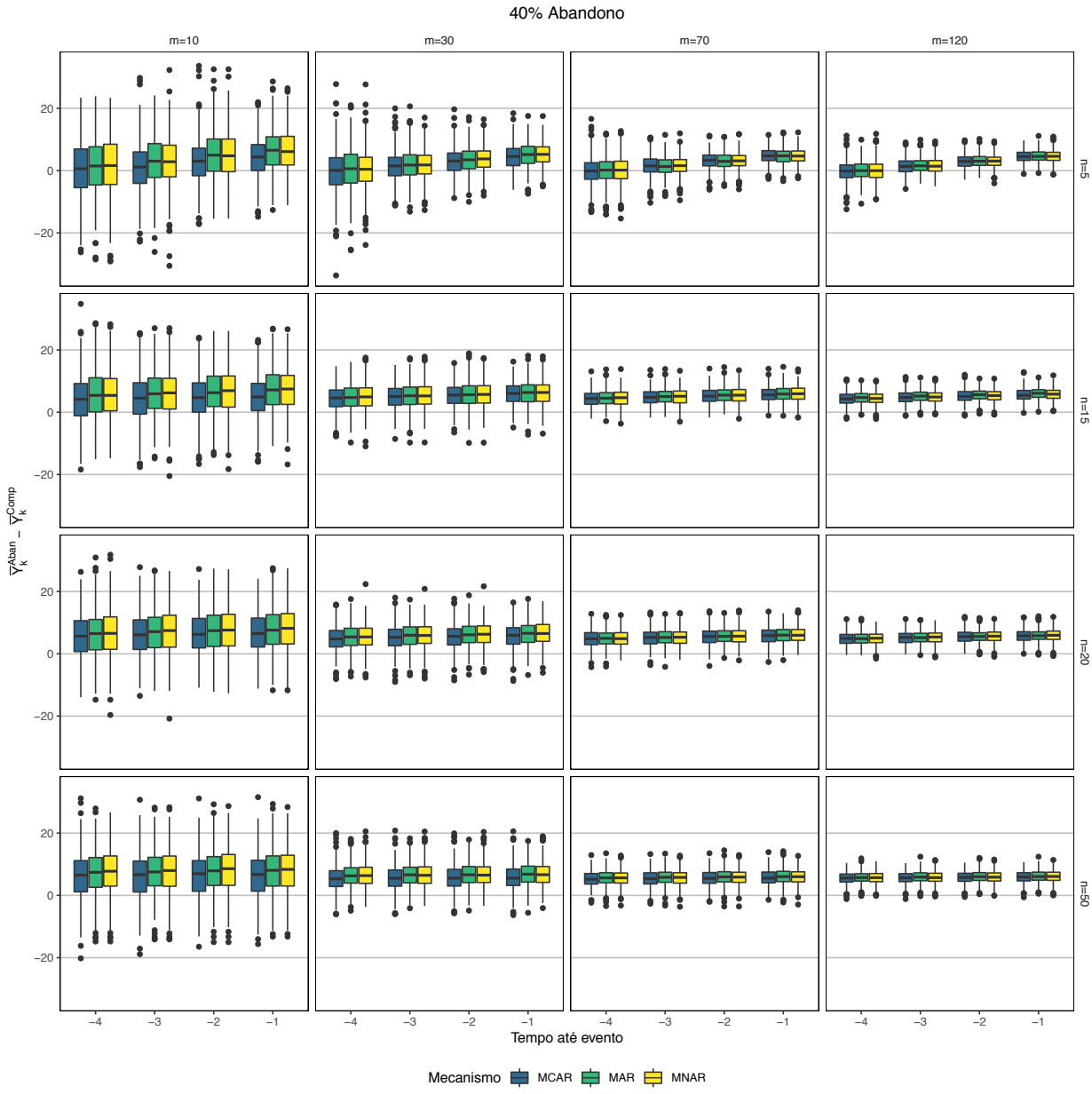


Figura B.10: Diagramas de caixa da diferença da média de $\boldsymbol{Y}_k$, no tempo até evento k, entre os indivíduos que completaram e os que abandonaram o estudo nas 500 simulações, nos últimos 4 instantes, e na presença de 40% de abandono dos indivíduos.

Anexo C

Fundamentos estatísticos e algébricos

C.1 Teoria multivariada Gaussiana

Um tratamento mais cuidado das propriedades abaixo descritas pode ser encontrado, por exemplo, em Mardia et al. (1979) [42].

Considere um vetor aleatório $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ que segue a distribuição multivariada Gaussiana, $\boldsymbol{Y} \sim MVN(\boldsymbol{\mu}, \boldsymbol{V})$.

1. $\boldsymbol{Y}$ tem uma função densidade de probabilidade da forma

$$f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{V}) = \frac{1}{(2\pi)^{n/2} \mid \boldsymbol{V} \mid^{1/2}} exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^\top \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\right\},$$

   onde $-\infty < y_j < \infty$, $j = 1, \ldots, n$.

2. Cada Y_j segue uma distribuição Gaussiana univariada,

$$Y_j \sim N(\mu_j, \sigma_j^2),$$

   onde $j = 1, \ldots, n$.

3. Se $\boldsymbol{Z} = (Y_1, \ldots, Y_{n_z})^\top$ com $n_z < n$, então $\boldsymbol{Z}$ segue uma distribuição Gaussiana multivariada com

$$\boldsymbol{Z} \sim MVN\left(\begin{bmatrix}\mu_1\\ \vdots \\ \mu_{n_z}\end{bmatrix}, \begin{bmatrix} v_{1,1} & \cdots & v_{1,n_z} \\ \vdots & \ddots & \vdots \\ v_{n_z,1} & \cdots & v_{n_z,n_z}\end{bmatrix}\right).$$

, onde a matriz de (co)variâncias é a sub-matriz superior esquerda n_z por n_z de $\boldsymbol{V}$.

4. Se $\boldsymbol{Z}_1 = (Y_1, \ldots, Y_{n_z})^\top$ e $\boldsymbol{Z}_2 = (Y_{n_z+1}, \ldots, Y_n)^\top$ com $n_z < n$, então $[\boldsymbol{Z}_1 \mid \boldsymbol{Z}_2 = \boldsymbol{z}_2]$ segue a distribuição multivariada Gaussiana com o vetor valor médio

$$\boldsymbol{\mu}_{\boldsymbol{Z}_1|\boldsymbol{Z}_2} = \boldsymbol{\mu}_1 + \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}(\boldsymbol{z}_2 - \boldsymbol{\mu}_2),$$

e matriz de (co)variâncias

$$\boldsymbol{V}_{\boldsymbol{Z}_1|\boldsymbol{Z}_2} = \boldsymbol{V}_{11} - \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}\boldsymbol{V}_{12}^\top,$$

onde $\mu_1 = (\mu_1, \ldots, \mu_{n_z})$, $\mu_2 = (\mu_{n_z+1}, \ldots, \mu_n)$ e $\boldsymbol{V}$ é particionada como

$$\underset{n\times n}{\boldsymbol{V}} = \begin{bmatrix} \underset{n_z\times n_z}{V_{11}} & \underset{n_z\times(n-n_z)}{V_{12}} \\ \underset{(n-n_z)\times n_z}{V_{12}^\top} & \underset{(n-n_z)\times(n-n_z)}{V_{22}}\end{bmatrix}.$$

5. Se $\boldsymbol{Z}_1 = (Y_1, \ldots, Y_{n_z})^\top$ e $\boldsymbol{Z}_2 = (Y_{n_z+1}, \ldots, Y_n)^\top$ com $n_z < n$, então $[\boldsymbol{Z}_2 \mid \boldsymbol{Z}_1 = \boldsymbol{z}_1]$ segue uma distribuição multivariada Gaussiana com o vetor médio

$$\underset{(n-n_z)\times 1}{\boldsymbol{\mu}_{\boldsymbol{Z}_2|\boldsymbol{z}_1}} = \boldsymbol{\mu}_2 + \boldsymbol{V}_{12}^\top\boldsymbol{V}_{11}^{-1}(\boldsymbol{z}_1 - \boldsymbol{\mu}_1),$$

e a matriz (co)variância

$$\underset{(n-n_z)\times(n-n_z)}{\boldsymbol{V}_{\boldsymbol{Z}_2|\boldsymbol{z}_1}} = \boldsymbol{V}_{22} - \boldsymbol{V}_{12}^\top\boldsymbol{V}_{11}^{-1}\boldsymbol{V}_{12},$$

onde $\mu_1 = (\mu_1, \ldots, \mu_{n_z})$, $\mu_2 = (\mu_{n_z+1}, \ldots, \mu_n)$ e $\boldsymbol{V}$ é particionado como

$$\underset{n\times n}{\boldsymbol{V}} = \begin{bmatrix} \underset{n_z\times n_z}{V_{11}} & \underset{n_z\times(n-n_z)}{V_{12}} \\ \underset{(n-n_z)\times n_z}{V_{12}^\top} & \underset{(n-n_z)\times(n-n_z)}{V_{22}} \end{bmatrix}.$$

C.2 Operações sobre matrizes

Operações elementares

Para as matrizes $\boldsymbol{A}$, $\boldsymbol{B}$, e $\boldsymbol{C}$ verificam-se as seguintes propriedades [43]:

1. $\boldsymbol{A}^\top = \boldsymbol{A}$, se $\boldsymbol{A}$ é uma matriz simétrica;
2. $\left(\boldsymbol{A}^\top\right)^{-1} = \left(\boldsymbol{A^{-1}}\right)^\top$, se $\boldsymbol{A}$ é uma matriz não singular;
3. $(\boldsymbol{ABC}\cdots)^\top = \cdots \boldsymbol{C}^\top\boldsymbol{B}^\top\boldsymbol{A}^\top$.

Derivação

Sejam $\boldsymbol{X}$ uma matriz, $\boldsymbol{a}$ e $\boldsymbol{b}$ vetores, provam-se as seguintes propriedades:

1. $\dfrac{\partial \ln|\det(\boldsymbol{X})|}{\partial \boldsymbol{X}} = \left(\boldsymbol{X}^{-1}\right)^\top = \left(\boldsymbol{X}^\top\right)^{-1}$ [44];
2. Como $\dfrac{\boldsymbol{X}^{-1}}{\partial x} = -\boldsymbol{X}^{-1}\dfrac{\partial \boldsymbol{X}}{\partial x}\boldsymbol{X}^{-1}$, tem-se que $\dfrac{\partial \boldsymbol{a}^\top\boldsymbol{X}^{-1}\boldsymbol{b}}{\partial \boldsymbol{X}} = -\boldsymbol{X}^{-\top}\boldsymbol{a}\boldsymbol{b}^\top\boldsymbol{X}^{-\top}$ [45];
3. $\dfrac{\boldsymbol{b}^\top\boldsymbol{a}}{\partial \boldsymbol{b}} = \dfrac{\boldsymbol{a}^\top\boldsymbol{b}}{\partial \boldsymbol{b}} = \boldsymbol{a}$ [46];
4. $\dfrac{\boldsymbol{a}^\top\boldsymbol{X}\boldsymbol{a}}{\partial \boldsymbol{a}} = 2\boldsymbol{X}\boldsymbol{a}$, se $\boldsymbol{X}$ é uma matriz simétrica [46];
5. $\dfrac{\partial}{\partial b}(\boldsymbol{a}-\boldsymbol{X}\boldsymbol{b})\,\boldsymbol{W}\,(\boldsymbol{a}-\boldsymbol{X}\boldsymbol{b})^\top = -2\boldsymbol{X}^\top\boldsymbol{W}\,(\boldsymbol{a}-\boldsymbol{X}\boldsymbol{b})$, se $\boldsymbol{W}$ é uma matriz simétrica [46];
6. Se $\boldsymbol{U} = \mathcal{F}(\boldsymbol{X})$, então

$$\frac{\partial \mathcal{G}\,(\boldsymbol{U})}{\partial \boldsymbol{X}} = \frac{\partial \mathcal{G}\,(\mathcal{F}(\boldsymbol{X}))}{\partial \boldsymbol{X}}.$$

Aplicando a Regra da Cadeia, pode ser escrito na forma

$$\frac{\partial \mathcal{G}(\boldsymbol{U})}{\partial \boldsymbol{X}} = \frac{\partial \mathcal{G}(\boldsymbol{U})}{\partial x_{ij}} = \sum_{k=1}^{M}\sum_{l=1}^{N} \frac{\partial \mathcal{G}(\boldsymbol{U})}{\partial u_{kl}} \cdot \frac{\partial u_{kl}}{\partial x_{ij}},$$

ou na forma matricial

$$\frac{\partial \mathcal{G}(\boldsymbol{U})}{\partial X_{ij}} = Tr\left[\left(\frac{\partial \mathcal{G}(\boldsymbol{U})}{\partial \boldsymbol{U}}\right)^{\top} \cdot \frac{\partial \boldsymbol{U}}{\partial X_{ij}}\right].$$

C.3 Outros

Fórmula de Leibniz

A fórmula de Leibniz afirma que para um integral da forma

$$\int_{a(x)}^{b(x)} f(x,t)\, dt,$$

onde $-\infty < a(x),\ b(x) < \infty$, a sua derivada pode ser expressa como [47]:

$$\frac{\partial}{\partial x}\int_{a(x)}^{b(x)} f(x,t)dt = \int_{a(x)}^{b(x)} \frac{\partial}{\partial x} f(x,t)\, dt + \left(\frac{\partial}{\partial x} b(x)\right) \cdot f(x,b(x)) - \left(\frac{\partial}{\partial x} a(x)\right) \cdot f(x,a(x)).$$

Bibliografia

[1] Peter J. Diggle, Inês Sousa, and Amanda G. Chetwynd. Joint modelling of repeated measurements and time-to-event outcomes: The fourth Armitage lecture. *Statistics in Medicine*, 2008.

[2] Peter J Diggle, Patrick Heagerty, Patrick J Heagerty, Kung-Yee Liang, Scott Zeger, and Others. *Analysis of longitudinal data*. Oxford University Press, 2002.

[3] Garrett M Fitzmaurice and Nan M Laird. A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, 80(1):141–151, 1993.

[4] M Helena Gonçalves, M Salomé Cabral, Maria Carme Ruiz de Villa, Eduardo Escrich, and Montse Solanas. Likelihood approach for count data in longitudinal experiments. *Computational Statistics & Data Analysis*, 51(12):6511–6520, 2007.

[5] Vandna Jowaheer and Brajendra C Sutradhar. Analysing longitudinal count data with overdispersion. *Biometrika*, 89(2):389–399, 2002.

[6] G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs. *Longitudinal Data Analysis*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, 2008.

[7] M Salomé Cabral and M Helena Gonçalves. *Análise de Dados Longitudinais*, volume 793. Sociedade Portuguesa de Estatística, 2011.

[8] Nan M Laird, James H Ware, et al. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982.

[9] Geert Verbeke and Geert Molenberghs. *Linear mixed models for longitudinal data*. Springer Science & Business Media, 2009.

[10] José Pinheiro and Douglas Bates. *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media, 2006.

[11] Cristina Rocha and Ana Luísa Papoila. *Análise de Sobrevivência*. Sociedade Portuguesa de Estatística, 2009.

[12] John P Klein and Melvin L Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2006.

[13] Samuel Kotz, Narayanaswamy Balakrishnan, and Norman L Johnson. *Continuous multivariate distributions, Volume 1: Models and applications*, volume 1. John Wiley & Sons, 2004.

[14] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[15] DG Kleinbaum and M Klein. Survival analysis: A self-learning text, 2005. *New York, Springer-Verlag*, 2011.

[16] DR Cox and D Oakes. Analysis if survival data, 1984.

[17] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[18] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[19] Dimitris Rizopoulos. *Joint models for longitudinal and time-to-event data: With applications in R*. Chapman and Hall/CRC, 2012.

[20] Geert Molenberghs, Bart Michiels, Michael G Kenward, and Peter J Diggle. Monotone missing data and pattern-mixture models. *Statistica Neerlandica*, 52(2):153–161, 1998.

[21] Simon Garnier. *viridis: Default Color Maps from 'matplotlib'*, 2018. R package version 0.5.1.

[22] Inês Sousa. A review on joint modelling of longitudinal measurements and time-to-event. *Revstat Stat J*, 9:57–81, 2011.

[23] Peter Diggle and Michael G Kenward. Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 43(1):49–73, 1994.

[24] Roderick JA Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134, 1993.

[25] Roderick JA Little. A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81(3):471–483, 1994.

[26] Margaret C Wu and Raymond J Carroll. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, pages 175–188, 1988.

[27] Dean Follmann and Margaret Wu. An approximate generalized linear model with random effects for informative missing data. *Biometrics*, pages 151–168, 1995.

[28] Joseph W Hogan and Nan M Laird. Mixture models for the joint distribution of repeated measures and event times. *Statistics in medicine*, 16(3):239–257, 1997.

[29] Paraskevi Pericleous. *Parametric joint modelling for longitudinal and survival data*. PhD thesis, University of East Anglia, 2016.

[30] Alberto Garcia-Hernandez and Dimitris Rizopoulos. % jm: A sas macro to fit jointly generalized mixed models for longitudinal data and time-to-event responses. *Journal of Statistical Software*, 84(1):1–29, 2018.

[31] Michael J Crowther. Stjm: Stata module to fit shared parameter joint models of longitudinal and survival data. 2013.

[32] Pete Philipson, Peter Diggle, Ines Sousa, Ruwanthi Kolamunnage-Dona, Paula Williamson, and Robin Henderson. joiner: Joint modelling of repeated measurements and time-to-event data. 2012.

[33] Dimitris Rizopoulos. JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, 35(9):1–33, 2010.

[34] Dimitris Rizopoulos. The R package JMbayes for fitting joint models for longitudinal and time-to-event data using mcmc. *Journal of Statistical Software*, 72(7):1–45, 2016.

[35] Cécile Proust-Lima, Viviane Philipps, and Benoit Liquet. Estimation of extended mixed models using latent classes and latent processes: the r package lcmm. *arXiv preprint arXiv:1503.00890*, 2015.

[36] Anastasios A Tsiatis and Marie Davidian. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, pages 809–834, 2004.

[37] Grigorios Papageorgiou, Katya Mauff, Anirudh Tomer, and Dimitris Rizopoulos. An overview of joint modeling of time-to-event and longitudinal outcomes. *Annual review of statistics and its application*, 2019.

[38] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[39] Michael S Wulfsohn and Anastasios A Tsiatis. A joint model for survival and longitudinal data measured with error. *Biometrics*, pages 330–339, 1997.

[40] Nan Laird, Nicholas Lange, and Daniel Stram. Maximum likelihood computations with repeated measures: application of the em algorithm. *Journal of the American Statistical Association*, 82(397):97–105, 1987.

[41] Marie Davidian and David M Giltinan. Some simple methods for estimating intraindividual variability in nonlinear mixed effects models. *Biometrics*, pages 59–73, 1993.

[42] Kantilal Varichand Mardia, John T. Kent, and John M. Bibby. *Multivariate analysis*. Probability and mathematical statistics. Acad. Press, London, 1979.

[43] Bernard Kolman and David Ross Hill. *Elementary linear algebra*. Pearson Education, 2004.

[44] M Brookes. The matrix reference manual, 2019.

[45] S M Selby. *Standard Mathematical Tables*. CRC Press, 1974.

[46] K B Petersen and M S Pedersen. *The Matrix Cookbook*. Technical University of Denmark, 2012.

[47] Murray H Protter, B Charles Jr, et al. *Intermediate calculus*. Springer Science & Business Media, 2012.