UNICAMP

UNIVERSIDADE ESTADUAL DE CAMPINAS

Instituto de Matemática, Estatística e Computação Científica

NURY BIBIANA RIAÑO GAONA

Modelagem conjunta de dados longitudinais e de sobrevivência com risco competitivo e tempo discreto: Aplicação em dados de evasão estudantil

Campinas

2022

Nury Bibiana Riaño Gaona

Modelagem conjunta de dados longitudinais e de sobrevivência com risco competitivo e tempo discreto: Aplicação em dados de evasão estudantil

Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestra em Estatística.

Orientador: Rafael Pimentel Maia

Este trabalho corresponde à versão final da Dissertação defendida pela aluna Nury Bibiana Riaño Gaona e orientada pelo Prof. Dr. Rafael Pimentel Maia.

Campinas

2022

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

R351m

Riaño Gaona, Nury Bibiana, 1990-
Modelagem conjunta de dados longitudinais e de sobrevivência com risco competitivo e tempo discreto : aplicação em dados de evasão estudantil / Nury Bibiana Riaño Gaona. – Campinas, SP : [s.n.], 2022.

Orientador: Rafael Pimentel Maia.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Desempenho acadêmico. 2. Modelos mistos (Estatística). 3. Modelos lineares (Estatística). 4. Método longitudinal. 5. Modelos lineares generalizados. I. Maia, Rafael Pimentel, 1983-. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Informações Complementares

Título em outro idioma: Joint modeling of longitudinal data and survival with competing risk and discrete time : application to student drop out data
Palavras-chave em inglês:
Academic performance
Mixed models (Statistics)
Linear models (Statistics)
Longitudinal method
Generalized linear models
Área de concentração: Estatística
Titulação: Mestra em Estatística
Banca examinadora:
Rafael Pimentel Maia [Orientador]
Benilton de Sá Carvalho
Cristian Marcelo Villegas Lobos
Data de defesa: 14-09-2022
Programa de Pós-Graduação: Estatística

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0001-5301-2679
- Currículo Lattes do autor: http://lattes.cnpq.br/4739904280836225

Dissertação de Mestrado defendida em 14 de setembro de 2022 e aprovada pela banca examinadora composta pelos Profs. Drs.

Prof(a). Dr(a). RAFAEL PIMENTEL MAIA

Prof(a). Dr(a). BENILTON DE SÁ CARVALHO

Prof(a). Dr(a). CRISTIAN MARCELO VILLEGAS LOBOS

A Ata da Defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria de Pós-Graduação do Instituto de Matemática, Estatística e Computação Científica.

Agradecimentos

Sinto-me feliz por culminar uma etapa tão importante da minha vida profissional. Quero agradecer a Universidade Estadual de Campinas pela hospitalidade e acolhimento que teve durante todo o mestrado, a todos os professores que me acompanharam durante o processo. Em especial, agradeço ao professor Caio pelos fundamentos teóricos da disciplina de Inferência que me ajudou a fortalecer os conceitos de estatística.

Agradeço ao meu orientador Rafael pela paciência, ensinanças e conversas que me levaram a concluir com sucesso a dissertação.

Agradeço ao meu namorado Javier por me acompanhar durante todo o processo de graduação e mestrado, pela coragem e perseverança em alcançar grandes conquistas, pois me deu força para alcançar também as minhas.

Agradeço a minha família por estar sempre presente na minha vida. A minha mãe por me levar pelo caminho da educação, ao meu pai por sempre me apoiar, ao meu irmão Jeison pela camaradagem, amizade e longas conversas que tivemos à distância, ao meu irmão Stiven pela amizade e entretenidas conversas sobre música.

Agradeço também as minhas famílias maiores, a família Mendieta e Gaona, porque existe sempre o verdadeiro amor nos laços que nos tem mantido próximos.

Agradeço a todos os meus amigos da Colômbia pelas conversas no *Zoom*, especialmente às minhas amigas de longa data, Indira e Angélica.

Agradeço aos meus amigos que conheci durante o programa de mestrado, porque alegraram os meus dias, à Ana, à Sara, ao Ivan e ao Cícero.

Agradeço ao meu país Colômbia e à Universidade Nacional da Colômbia pela educação pública e de qualidade que me proporcionaram. E agradeço também por um governo diferente, um governo de vida, baseado na paz, na justiça social e ambiental.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior — Brasil (CAPES) — Código de Financiamento 001.

Resumo

Neste trabalho propomos um modelo de risco competitivo conjunto com tempo discreto para modelar o tempo de permanência dos alunos do curso de bacharelado em Estatística da Universidade Estadual de Campinas (UNICAMP), considerando dois riscos competitivos, o aluno se forma, ou o aluno evade o curso. Também é modelado, via um modelo conjunto, o coeficiente de rendimento (CR) padronizado medido longitudinalmente, semestre a semestre.

O conjunto de dados é composto por pelos registros de todos os alunos ingressantes no curso de bacharelado em Estatística entre 2012 e 2017, com tempo de máximo de observação de 12 semestres. Estudantes com tempo de permanência superior à 12 semestres foram tratados como censura à direita. As informações foram obtidas a partir de um questionário demográfico aplicado pela Comissão Permanente de Vestibulares (COMVEST) na inscrição do vestibular, as notas do vestibular aplicadas também pelo COMVEST e informações acadêmicas fornecidas pela Diretoria Acadêmica (DAC) da UNICAMP.

Ao modelar conjuntamente o tempo de permanência dos alunos e o CR padronizado, é proposto um modelo conjunto com dois modelos marginais: um modelo de risco competitivo e um modelo linear misto.

Para ajustar o modelo conjunto serão utilizadas duas abordagens, o modelo de dois estágios em que o efeito aleatório é estimado usando os dados longitudinais no primeiro estágio, e um modelo bivariado, com efeito aleatório compartilhado, em que todos os efeitos são estimados simultaneamente a partir da distribuição conjunta, para eventualmente recuperar a curva de sobrevivência.

Por fim, para avaliar a qualidade de ajuste do modelo conjunto, utilizamos a estatística deviance e a análise gráfica dos resíduos: matingale, deviance ajustado e Cox-Snell. Os resultados do modelo conjunto indicam que o efeito do CR padronizado tem efeito significativo na curva de sobrevivência ou tempo de permanência, e com base na estatística deviance, observamos que o modelo bivariado apresenta melhor ajuste que o modelo de dois estágios.

Palavras-chave: Risco Competitivo, Modelagem Conjunta, Dados Longitudinais, Dados de Sobrevivência Tempo Discreto, Modelo Linear Misto.

Abstract

In this work, we propose a discrete-time joint competitive risk model to model the length of stay of students of the Bachelor of Statistics course at Universidade Estadual de Campinas (UNICAMP), considering two competitive risks, the student graduates, or the student evades the course. The standardized yield coefficient (CR) measured longitudinally, semester by semester, is also modeled via a joint model.

The data set is composed of the records of all students entering the bachelor course in statistics between 2012 and 2017, with a maximum observation time of 12 semesters. Students who stayed longer than 12 semesters were treated as right censoring. The information was obtained from a demographic questionnaire applied by the Permanent Commission of Vestibulares (COMVEST) at enrollment of the entrance exam, the marks of the entrance exam applied also by COMVEST and academic information provided by the Academic Board (DAC) of UNICAMP.

By jointly modeling the students' length of stay and the standardized CR, a joint model with two marginal models is proposed: a competitive risk model and a mixed linear model.

To adjust the joint model, two approaches will be used, the two-stage model in which the random effect is estimated using the longitudinal data in the first stage, and a bivariate model, with shared random effect, in which all the effects are estimated simultaneously from of the joint distribution, to eventually recover the survival curve.

Finally, to assess the goodness of fit of the joint model, we used the deviance statistic and the graphical analysis of residuals: matingale, adjusted deviance and Cox-Snell. The results of the joint model indicate that the effect of the standardized CR has a significant effect on the survival curve or length of stay, and based on the deviance statistic, we observed that the bivariate model presents a better fit than the two-stage model.

Keywords: Competitive Risk, Joint Modeling, Longitudinal Data, Discrete Time Survival Data, Linear Mixed Model

Lista de ilustrações

Lista de tabelas

Sumário

Introdução

A evasão (saída definitiva do curso de origem sem conclusão) nos programas de graduação das universidades são um problema de grande transcendência para o sistema de educação, porque representa um elevado custo econômico e social. Segundo o resumo técnico do censo da educação superior 2019 (INEP, 2021) a taxa de evasão para os alunos ingressantes em 2010 após dez anos de acompanhamento foi de 59% e de 38% ao final do terceiro ano. Para os os ingressante entre 2010 até 2015 ao final do quinto ano de acompanhamento de 50% a 54% dos ingressantes desistiram de seus cursos e ao final do terceiro ano de acompanhamento de 37% a 40% já haviam desistido de seus cursos. Adicionalmente, o Instituto de Excelência ao Serviço de Ensino Superior (SEMESP, 2021) no *Mapa do Ensino Superior* na edição para 2021, evidenciou que a região de Campinas teve uma taxa de evasão do 33.5% para cursos presenciais. Alguns fatores associados à evasão no ensino superior das ciências naturais e engenharias segundo (SACCARO; FRANÇA; JACINTO, 2019) é ser homem e ter mais idade, os efeitos foram avaliados a partir de estudos de análise de sobrevivência.

O objetivo deste trabalho é propor um modelo de riscos competitivos conjunto para estudar o tempo de permanência dos alunos ingressantes no curso de Estatística da UNICAMP e encontrar associações entre a chance de se formar ou evadir o curso de graduação em Estatística com as variáveis socio-demográficas, variáveis de desempenho no vestibular e a variável longitudinal, o desempenho acadêmico mensurado ao longo de 12 semestres.

O modelo será ajustado com os registros de de 424 alunos ingressantes no curso de graduação em Estatística entre os anos 2012 e 2017. Os dados cosniderados foram as respostas ao questionário sociodemográfico aplicado pela Comissão Permanente de Vestibulares da Unicamp (COMVEST) na inscrição do vestibular, as notas da prova de segunda fase do vestibular nas disciplinas de português, matemática, geografia, historia, física, biologia e química, e informações acadêmicas fornecidas pela Diretoria Acadêmica (DAC). Foram considerados apenas o primeiro ingresso de cada estudante. A Tabela 1 mostra a distribuição dos alunos segundo o ano de ingresso e a situação dos alunos atualizados em março de 2022. Foi considerado o tempo de até 12 semestre desde o início da graduação. Observamos na Tabela 1 que o curso de graduação em Estatística em geral apresenta uma taxa de evasão alta. Para o ano 2012 temos a taxa mais alta de evasão com o 75%, seguida pelo ano de ingresso 2015 com 58%. Também observa-se que entre os anos de ingresso 2012 até 2017 se apresenta uma taxa de evasão maior que a taxa de formação com exceção do ano 2016.

Tabela 1 – Distribuição dos alunos ingressantes no curso de Estatística por ano de ingresso segundo a situação dos alunos ao longo de 12 semestres; dados atualizados em março 2022

Situação	Ano de Ingresso						
	2012	2013	2014	2015	2016	2017	Total
Ativo	4 5%	6 8%	6 8%	3 4%	8 12%	19 28%	46 11%
Evadido	57 75%	41 55%	35 49%	39 58%	21 31%	24 35%	217 51%
Formado	15 20%	27 36%	30 42%	25 37%	38 57%	26 38%	161 38%
Total Alunos	76 100%	74 100%	71 100%	67 100%	67 100%	69 100%	424 100%

Análise de sobrevivência denomina um conjunto de método estatísticos aplicados à estudos em que a variável resposta consiste do tempo, de um ponto inicial definido, até a ocorrência de um evento de interesse. O problema de riscos competitivos surge quando há dois ou mais eventos de interesse possíveis. Estes modelos devem considerar a presença de dados incompletos, também chamados dados censurados, ou seja, que não foram possíveis de medir em todo ponto do tempo pela natureza da pesquisa. O tempo de permanência dos estudantes é representado por uma variável aleatória discreta, mensurada em semestres. Assim, no capítulo 1 deste trabalho apresentamos uma breve revisão dos aspectos mais relevantes em análise de sobrevivência para risco competitivo para tempo discreto tomando como referência o livro de (TUTZ; SCHMID, 2016) e o artigo de (SCHMID; BERGER, 2021) que fornecem uma visão geral da metodologia estatística atualmente disponível.

Depois, no capítulo 2 é apresentada uma revisão de modelos longitudinais com o foco de estudar a variável CR padronizado mensurada em cada semestre para cada aluno. Os modelos longitudinais são tratados via modelo linear misto. Uma das vantagens do modelo linear misto é permite acomodar duas categorias de variabilidade, a variabilidade entre e dentro dos indivíduos.

Já no capítulo 3 abordamos duas metodologias para o ajuste do modelo conjunto, elas permitem incorporar a variável longitudinal no modelo de sobrevivência e mensurar seu efeito. Segundo (RIZOPOULOS, 2012) a ideia motivadora por trás dos modelos conjuntos é acoplar o modelo de sobrevivência, sendo de interesse primário, com um modelo adequado para as medições repetidas da covariável longitudinal. No livro de (RIZOPOULOS, 2012) encontramos que os modelos conjuntos foram desenvolvidos quando o tempo é modelado como uma variável continua, mas nem sempre é adequado. A vantagem dos modelos conjuntos para tempo discreto é que na maioria dos estudos o tempo é medido de forma discreta e usar as ferramentas com tempo discreto se volta intuitivo. Na hora de estimar os

parâmetros do modelo conjunto de riscos competitivos para tempo discreto, nos baseamos na tese de doutorado (QIU, 2012) e no artigo (QIU et al., 2016), em que se apresenta a estimação em dois estágios e a estimação com parâmetro compartilhado para o modelo conjunto. No que segue apresentamos a literatura sobre a modelagem conjunta continua, como proposito de aplicar algumas ferramentas e conceitos na modelagem conjunta discreta.

No trabalho de (LIMA, 2007) se apresentam quatro modelos conjuntos contínuos. O primeiro modelo, é o modelo desenvolvido por (WULFSOHN; TSIATIS, 1997), em que o modelo de Cox é usado para o processo de sobrevivência e o modelo linear misto para o processo longitudinal. A estimação da variável resposta longitudinal é introduzida no modelo de Cox como variável explicativa. Para estimar os parâmetros deste modelo é usado o algoritmo EM na função de máxima verossimilhança dos dados observados. O segundo modelo, é o modelo conjunto de (HENDERSON; DIGGLE; DOBSON, 2000), este modelo é dividido em duas partes, a parte de sobrevivência e a parte longitudinal, onde em cada parte se incorpora uma variável latente, ao juntar às duas variáveis latentes, elas geram um processo latente normal bi-variado, para estimar os parâmetros do modelo, é usado o algoritmo EM, do mesmo modo que no modelo anterior. O terceiro modelo, é o modelo conjunto de (DIGGLE; SOUSA; CHETWYND, 2008), neste modelo o processo de sobrevivência é transformado por meio da função logarítmica e leva-se em consideração o modelo normal multivariado para a variável (W, S) onde W é a variável longitudinal e $S = \ln T$ é a variável de sobrevivência, a estimação dos parâmetros baseia-se na distribuição normal multivariada. Por fim, o quarto modelo, o modelo conjunto em dois estágios de (VONESH; GREENE; SCHLUCHTER, 2006), sendo um modelo mais simples do que os outros três, porque não precisam de cômputos complexos. No primeiro estágio se estimam os parâmetros para o processo longitudinal e no segundo estágio se utiliza os valores ajustados da variável longitudinal como variável explicativa dependente do tempo no modelo de Cox.

No artigo de (IBRAHIM; CHU; CHEN, 2010) se mostra uma revisão dos modelos conjuntos aplicados num banco de dados sobre ensaios clínicos de câncer, em que biomarcadores longitudinais estão associados com o tempo de sobrevivência. Neste estudo se comparam três modelos o modelo de Cox sem covariável longitudinal, o modelo de Cox com covariável longitudinal e o modelo conjunto em dois estágios, apresentando uma melhor estimativa para a curva de sobrevivência o modelo em dois estágios, devido a que usa os efeitos estimados para estimar a curva de sobrevivência e não os dados observados da variável longitudinal, com grande variabilidade de erro pelas medidas dentro de cada indivíduo.

Outro trabalho interessante que fazem uso de modelos conjuntos é a tese de (MAIORANO, 2018), que propõe analisar dados para avaliação de desfechos clínicos do parto com modelagem conjunta para tempo contínuo. Neste estudo, utilizam-se as

generalizações da distribuição Weibull com três parâmetros para o processo de sobrevivência e os modelos lineares mistos para o processo longitudinal.

Adicionalmente, os modelos conjuntos para tempo contínuo têm sido estudados no artigo (PAPAGEORGIOU et al., 2019) que além das técnicas de estimação do modelo conjunto, apresenta o software disponível para aplicação dos modelos. Segundo (CROWTHER, 2014) a abordagem mais comum em análise conjunta é combinar os modelos lineares mistos com o modelo de riscos proporcionais, por efeitos aleatórios compartilhados, que ajuda a caracterizar a associação entre os dois processos. Na tese de doutorado de (CROWTHER, 2014) se pode encontrar a metodologia de parâmetro compartilhado incorporando o modelo de sobrevivência paramétrico flexível Royston-Parmar.

Além dos modelos conjuntos para um só risco, nosso interesse principal são os modelos de risco competitivo. Segundo (MONDAL, 2017) para os modelos de risco competitivo tem-se duas abordagens amplamente usadas, a primeira é estudar cada uma das causas ou riscos usando o modelo de riscos proporcionais de Cox, e a segunda abordagem é modelar a função de risco associada à função de incidência acumulada. O modelo baseado na função de incidência acumulada é chamado de modelo de riscos de sub-distribuição proporcional. Usando às duas abordagens dos riscos competitivos e trabalhando com dados longitudinais, se gera o modelo conjunto de riscos competitivos, em que, primeiro usamos o modelo linear misto para estimar os efeitos aleatórios da variável longitudinal e logo introduzimos os efeitos no modelo de riscos competitivos.

Logo, no capítulo 4 apresentamos os resultados do ajuste do modelo conjunto de dados longitudinais e de sobrevivência com risco competitivo e tempo discreto ao banco de dados educacionais do curso de graduação em estatística com uma janela de tempo para o ano de ingresso de 2012 até 2017. Por fim, no capítulo 5 são apresentadas as principais conclusões do trabalho bem como perspectivas de trabalhos futuros.

1 Análise de Sobrevivência

Apresentamos neste capítulo uma revisão da teoria de análise de sobrevivência abordando o tempo como uma variável aleatória discreta. A principal característica dos métodos de análise de sobrevivência é estudar o tempo até a ocorrência de um evento de interesse, muitas vezes chamado de tempo de falha. Um aspecto comum em estudos desta natureza é a presença de dados censurados, ou seja, com informação incompleta da variável tempo de um ou mais indivíduos no estudo.

Os dados coletados podem ser classificados como: *dados censurados*, observações que são incompletas de uma maneira conhecida, o que, em linhas gerais, significa que os indivíduos não experimentam a falha na janela de tempo do estudo; e os *dados completos*, são informações de indivíduos que experimentaram a falha antes do estudo terminar.

Há vários tipos de mecanismos que geram a censura: censura à direita, à esquerda e intervalar; nesta dissertação abordaremos só a censura à direita. O mecanismo de *censura à direita*, no que lhe concerne, também se divide em três tipos: tipo 1, tipo 2 e aleatória. Nos três tipos, os tempos observados dos indivíduos no estudo são menores que o tempo de ocorrência do evento. Os *dados com censura tipo 1* ocorrem quando o experimento termina após um período pré-estabelecido de tempo e alguns indivíduos ainda estão em risco; os *dados com censura tipo 2* ocorrem quando o experimento termina após ter ocorrido o evento de interesse em um número pré-estabelecido de indivíduos; já os *dados com censura aleatória* ocorrem quando, no decorrer do experimento, alguns indivíduos são retirados sem ter ocorrido a falha. A Figura 1 ilustra os tipos de censura à direita (COLOSIMO; GIOLO, 2006).

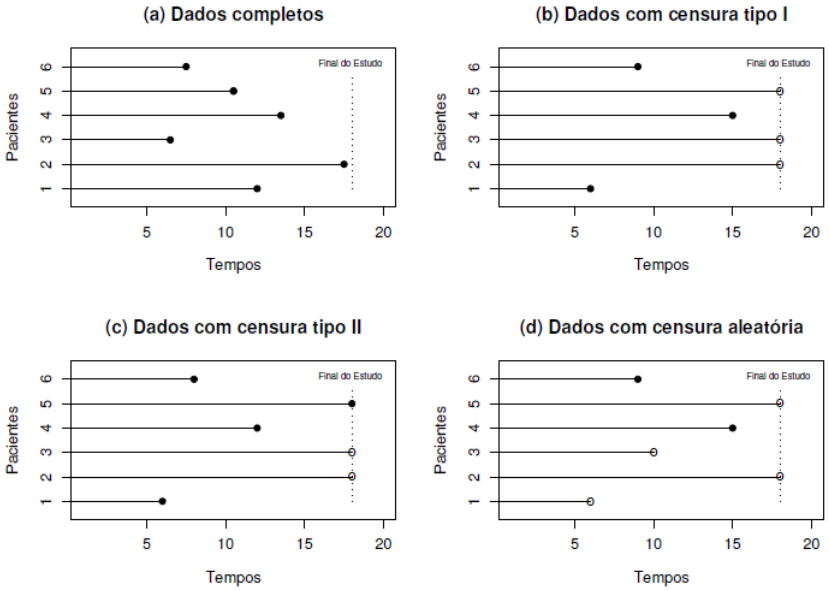
1.1 Caracterizando o tempo de sobrevivência

Seja T uma variável aleatória discreta representando o tempo de sobrevivência. Os dados de sobrevivência para o indivíduo i $(i = 1, 2, ..., n)$ sob estudo, são apresentados, em geral, pela dupla (t_i, δ_i) sendo $T_i = t_i$ o tempo de falha ou censura observado e δ_i a variável indicadora de falha ou censura, isto é,

$$\delta_i = \begin{cases} 1, & \text{se } t_i \text{ é um tempo de falha} \\ 0, & \text{se } t_i \text{ é um tempo censurado} \end{cases}$$

A seguir, definimos as funções básicas de análise de sobrevivência. Sejam os tempos de falha ordenados $t_0 < t_1 < t_2 ... < t_k$. A Primeira função, **a função de probabilidade** é definida como a probabilidade de uma observação falhar no tempo exato $T = t_j$.

Figura 1 – Ilustração de alguns mecanismos de censura em que • representa a falha e ◦ a censura. (a) todos os indivíduos experimentaram o evento antes do final do estudo, (b) alguns pacientes não experimentaram o evento até o final do estudo, (c) o estudo foi finalizado após a ocorrência de um número pré-estabelecido de falhas e (d) o acompanhamento de alguns indivíduos foi interrompido por alguma razão e alguns pacientes não experimentaram o evento até o final do estudo.



[h!]

Fonte: (COLOSIMO; GIOLO, 2006) (p.10)

$$f(t_j) = P(T = t_j) = F_T(t_j) - F_T(t_j^-) \qquad j = 1, 2, ..., k,$$

em que $F_T(t) = P(T \leqslant t)$ representa a **função de distribuição acumulada** e $F_T(t^-) = \lim_{t \to t_j^-} F(t_j) = F(t_{j-1})$.

A **função de sobrevivência** é definida como a probabilidade de que uma observação não falhe até um certo tempo t, ou seja, a probabilidade de que uma observação sobreviva ao tempo t. Em termos probabilísticos isto se escreve como:

$$S(t) = P(T > t) = 1 - F(t) = \sum_{j:t_j > t} f(t_j).$$

Outra função importante é a chamada **função risco** definida como a probabilidade de que o indivíduo falhe no tempo $T = t_i$, dado que o indivíduo sobreviveu até o tempo $T = t_i$.

$$\lambda_j = P(T = t_j \mid T \geqslant t_j) = \frac{f(t_j)}{S(t_{j-1})}.$$

A função de sobrevivência para o tempo $T = t_0$ como:

$$S(t_0) = P(T > t_0) = 1 - P(T \leqslant t_0) = 1.$$

Para o tempo $T = t_1$,

$$\begin{aligned} S(t_1) &= P(T > t_1) = P(T > t_1, T > t_0) \\ &= P(T > t_1 | T > t_0) \overbrace{P(T > t_0)}^{1} \\ &= 1 - P(T \leqslant t_1 | T > t_0) = 1 - P(T = t_1 | T \geqslant t_1) \\ &= 1 - \lambda_1. \end{aligned}$$

Enfim, podemos conceber, em termos gerais, que a função de sobrevivência para o tempo $T = t_j$ é:

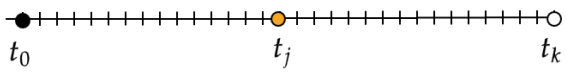
$$S(t_j) = P(T > t_j) = \prod_{s=1}^{j}(1 - \lambda_s) \qquad j = 1, 2, ..., k.$$

Da mesma forma, a função de probabilidade pode ser escrita como:

$$f(t_j) = P(T = t_j) = \lambda_j S(t_{j-1}) = \lambda_j \prod_{s=1}^{j-1}(1 - \lambda_s).$$

1.2 Estimando a curva de sobrevivência - Estimador de Kaplan-Meier

Considere uma amostra aleatória de n indivíduos com k distintos tempos de falhas $(k < n)$ $0 = t_0 < t_1 < t_2, \ldots, t_k < \infty$ e seja d_j o número de falhas observadas no tempo t_j.



Defina ainda a função de risco no tempo t_j com $j = 0, 1, ..., k$ por $\lambda_j = P(T = t_j \mid T \geqslant t_j)$. Na ausência de dados censurados a função de verossimilhança em termos da função de risco é dada por:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\lambda}) &= \prod_{j=1}^{k} f(t_j)^{d_j} = \prod_{j=1}^{k} (\lambda_j S(t_{j-1}))^{d_j} \\ &= \prod_{j=1}^{k} \left[\lambda_j \left(\prod_{s=1}^{j-1} (1 - \lambda_s) \right) \right]^{d_j} \\ &= \prod_{j=1}^{k} \lambda_j^{d_j} (1 - \lambda_j)^{n - \sum_{s \leqslant j} d_s} \end{aligned}$$

em que $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)^T$.

Na presença de censura à direita consideremos ainda que: m_j seja o número de indivíduos censurados no intervalo $[t_j, t_{j+1})$ nos tempos $t_{j1}, \ldots, t_{jm_j}$ e que n_j seja o número de indivíduos sob o risco em t_j, ou seja, os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior t_j. Para os indivíduos que tiveram o tempo de falha observado a contribuição para a função de verossimilhanças é dada pela função de probabilidade $f(\cdot)$, já para os indivíduos censurados à direita a contribuição é dada pela função de sobrevivência $S(\cdot)$ aplicada no tempo de censura. A função de verossimilhança para dados censurados à direita é:

$$\mathcal{L}(\boldsymbol{\lambda}) = \prod_{j=1}^{k} f(t_j)^{d_j} \prod_{l=1}^{m_j} S(t_{jl}) = \prod_{j=1}^{k} (\lambda_j S(t_{j-1}))^{d_j} \prod_{l=1}^{m_j} S(t_{jl}). \tag{1.1}$$

Se consideramos ainda que a contribuição de indivíduos censurados no intervalo $[t_j, t_{j+1})$ seja, $S(t_j)$ então a função de verossimilhança para dados censurados à direita é:

$$\mathcal{L}(\boldsymbol{\lambda}) = \prod_{j=1}^{k} (\lambda_j S(t_{j-1}))^{d_j} S(t_j)^{m_j} = \prod_{j=1}^{k} \lambda_j^{d_j} (1-\lambda_j)^{n_j - d_j}.$$

Agora, maximizando a função acima em relação a λ_j podemos mostrar que o estimador que maximiza a função de verossimilhança dada na equação 1.1 é:

$$\hat{\lambda}_j = \frac{d_j}{\sum_{i=j}^{k}(m_i + d_i)} = \frac{d_j}{n_j}.$$

O estimador de Kaplan-Meier (K-M) para a função de sobrevivência é então dado por:

$$\hat{S}(t_j) = \prod_{j=1}^{k} (1 - \hat{\lambda}_j),$$

também chamado estimador produto limite. Para mais detalhes, ver (KAPLAN; MEIER, 1958; KALBFLEISCH; PRENTICE, 2011; COLOSIMO; GIOLO, 2006).

O estimador de K-M é um dos métodos não paramétricos mais utilizados para estimar a função de sobrevivência, embora, existem outros estimadores não paramétricos, por exemplo, o estimador da tabela de vida e o estimador de Nelson-Aalen. Estes estimadores oferecem uma estimação mais ajustada para alguns problemas particulares, revisar exemplos em (KALBFLEISCH; PRENTICE, 2011; AALEN, 1978).

1.3 Análise Univariada com Covariáveis

Dado o vetor de variáveis explanatórias $\mathbf{X}$ de dimensão $n \times p$ com n o número de indivíduos e e p o número de variáveis explanatórias com seu respectivo vetor de

parâmetros γ de dimensãoo $p \times 1$, nós temos, que a variável aleatória T o tempo pode ser descrito com quatro funções que capturam as suas características, a saber, a função de densidade $P(T = t \mid \mathbf{x})$, $t = t_1, \ldots, t_n$, a função de distribuição acumulada $F(t \mid \mathbf{x}) = P(T \leqslant t \mid \mathbf{x})$, a função de risco $\lambda(t \mid \mathbf{x}) = P(T = t \mid T \leqslant t, \mathbf{x})$ e a função de sobrevivência $S(t \mid \mathbf{x}) = P(T > t \mid \mathbf{x}) = \prod_{s:s \leqslant t} (1 - \lambda(s \mid \mathbf{x}))$.

A Tabela 2 apresenta os principais modelos para tempo discreto que associal a função de risco a uma combinação de variáveis explanatórias (preditor linear) em cada ponto no tempo:

Tabela 2 – Modelos para a função de risco com tempo discreto

Modelo Logístico: $\lambda(t\|\mathbf{x}) = \dfrac{\exp(\gamma_{0t} + \mathbf{x}^T\boldsymbol{\gamma})}{1 + \exp(\gamma_{0t} + \mathbf{x}^T\boldsymbol{\gamma})}$
Modelo Probito: $\lambda(t\|\mathbf{x}) = \phi(\gamma_{0t} + \mathbf{x}^T\boldsymbol{\gamma})$
Modelo Gompertz: $\lambda(t\|\mathbf{x}) = 1 - \exp(-\exp(\gamma_{0t} + \mathbf{x}^T\boldsymbol{\gamma}))$
Modelo Gumbel: $\lambda(t\|\mathbf{x}) = \exp(-\exp(-(\gamma_{0t} + \mathbf{x}^T\boldsymbol{\gamma})))$
Modelo Exponencial: $\lambda(t\|\mathbf{x}) = \exp(\gamma_{0t} + \mathbf{x}^T\boldsymbol{\gamma})$

Fonte: (TUTZ; SCHMID, 2016)

Consequentemente, é possível obter a função de probabilidade e sobrevivência em cada ponto do tempo. Observação, note que todos os modelos mencionados para a função de risco com tempo discreto devem cumprir propriedades de proporcionalidade definidas por **razão de continuação proporcional** (do inglês*continuation ratio* definido por (TUTZ; SCHMID, 2016)) facilitando na hora de estimar os parâmetros. Em termos gerais a razão de continuação proporcional se escreve como:

$$\psi(t|\mathbf{x}) := \frac{P(T = t \mid \mathbf{x})}{P(T > t \mid \mathbf{x})}.$$

No caso específico do Modelo Logístico a razão de continuação é:

$$\psi(t|\mathbf{x}) := \frac{P(T = t \mid \mathbf{x})}{P(T > t \mid \mathbf{x})} = e^{\gamma_{0t}}(e^{\gamma_1})^{x_1} \ldots (e^{\gamma_p})^{x_p}.$$

Agora, a razão das razões de continuação para dois indivíduos com o mesmo valor das covariaveis e em tempos distintos é dada por:

$$\frac{\psi(t \mid \mathbf{x})}{\psi(s \mid \mathbf{x})} = \exp(\gamma_{0t} - \gamma_{0s}).$$

Já a razão das razões de continuação para dois indivíduos distintos deixando em um mesmo ponto no tempo t:

$$\frac{\psi(t \mid \mathbf{x})}{\psi(t \mid \tilde{\mathbf{x}})} = \exp((\mathbf{x} - \tilde{\mathbf{x}})^T\gamma).$$

Se a função de continuação se calcula em dois tempos diferentes deixando fixas as covariáveis ela não vai depender das covariáveis, e se calculamos a função de razão de continuação no mesmo tempo e as covariáveis mudam ela não vai depender do tempo.

1.3.1 Método de Máxima Verossimilhança

Com o método de máxima verossimilhança estimaremos os parâmetros do modelo de regressão com covariáveis fixas e covariáveis mudando no tempo. Em forma geral temos o modelo:

$$\lambda(t \mid \mathbf{x}) = h(\mathbf{x}_{it}^T \boldsymbol{\beta})$$

em que $\mathbf{x}_{it}^T = (0, \ldots, 0, 1, 0, \ldots, 0, \mathbf{x}_i^T)$ com 1 sendo a t-ésima posição no vetor e $\boldsymbol{\beta}^T = (\gamma_{01}, \ldots, \gamma_{0q}, \boldsymbol{\gamma}^T)$.

Seja δ_i a indicadora de não censura para o i-ésimo indivíduo, isto é, recebe o valor 1 se o indivíduo i falhou e 0 caso contrário então a contribuição do i-ésimo indivíduo à função de verossimilhança é dada por

$$\mathcal{L}_i(\boldsymbol{\lambda}) = \lambda(t_i|x_i)^{\delta_i}(1 - \lambda(t_i \mid x_i))^{1-\delta_i} \prod_{j=1}^{t_i-1}(1 - \lambda(j \mid x_i)).$$

Defina a variável y_{is} por

$$y_{is} = \begin{cases} 1, & \text{se o indivíduo } i \text{ falha no tempo } s \\ 0, & \text{caso contrário.} \end{cases}$$

Então a contribuição do i-ésimo indivíduo a função de verossimilhança pode ser reescrita por

$$\mathcal{L}_i(\boldsymbol{\lambda}) = \prod_{s=1}^{t_i} \lambda(t_s|x_i)^{y_{is}}(1 - \lambda(t_s|x_i))^{1-y_{is}}.$$

Portanto, a função de verossimilhança será dada por

$$\mathcal{L}(\boldsymbol{\lambda}) = \prod_{i=1}^{n}\prod_{s=1}^{t_i} \lambda(t_s|x_i)^{y_{is}}(1 - \lambda(t_s|x_i))^{1-y_{is}}. \tag{1.2}$$

Observando a função de verossimilhança dada pela equação 1.2 nota-se que é equivalente à função de verossimilhança de um modelo de regressão de resposta binária $y_{11}, \ldots, y_{nt_n}$, onde se tem em total $t_1 + \cdots + t_n$ observações binárias é dizer t_i observações binárias para cada indivíduo i. As respostas binárias dependem do número de censuras e do tempo de vida, depois, o modelo de regressão para o risco se pode transformar em um modelo de regressão com resposta binária e usar todas as ferramentas que já conhecemos para o modelo de regressão binário

$$P(y_{ij} = 1 \mid \mathbf{x}_i) = h(\mathbf{x}_{ij}^T \beta).$$

O modelo de regressão logística descreve a transição entre tempos, como a função de verossimilhança se calculo para cada instante no tempo e cada indivíduo, dessa maneira podemos construir as matrizes que nos permitem distinguir entre dados censurados e não censurados.

Se $T = t_i$, $\delta_i = 1$ a observação binária t_i e a matriz de desenho para as variáveis estão dadas por:

Tabela 3 – Matriz para um indivíduo não censurado

Observações binárias	Matriz de desenho
0	$1 \quad 0 \quad 0 \quad \dots \quad 0 \quad x_{i1}^T$
0	$0 \quad 1 \quad 0 \quad \dots \quad 0 \quad x_{i2}^T$
$\vdots$	
0	$0 \quad 0 \quad 0 \quad \dots \quad 1 \quad x_{it_i}^T$
1	$0 \quad 0 \quad 0 \quad \dots \quad 1 \quad x_{it_i}^T$

Fonte:(TUTZ; SCHMID, 2016)[pág.54]

Quando $T = t_i$, $\delta_i = 0$ a observação t_i e a matriz de desenho para as variáveis estão dadas por

Tabela 4 – Matriz para um indivíduo censurado

Observações binárias	Matriz de desenho
0	$1 \quad 0 \quad 0 \quad \dots \quad 0 \quad x_{i1}^T$
0	$0 \quad 1 \quad 0 \quad \dots \quad 0 \quad x_{i2}^T$
$\vdots$	
0	$0 \quad 0 \quad 0 \quad \dots \quad 1 \quad x_{it_i}^T$
0	$0 \quad 0 \quad 0 \quad \dots \quad 1 \quad x_{it_i}^T$

Fonte:(TUTZ; SCHMID, 2016)[pág.54]

Em resumo, para estimar os parâmetros do modelo de sobrevivência para tempo discreto, se usa a coincidência da função de verossimilhança com um de resposta binária e todas a ferramentas do modelo de regressão logística (Probito, Gomperts, Gumbel, Exponencial) para a estimação de parâmetros, para mais detalhes ver (TUTZ; SCHMID, 2016)[pág.51].

1.4 Análise de Risco Competitivo

O problema de risco competitivo surge em estudos em que a variável de interesse é o tempo até a falha e a falha acontece por uma de duas ou mais causas possíveis. Por exemplo, quando se estuda alguma doença, os indivíduos podem vir a óbito por mais de uma causa e se quer conhecer qual causa é mais provável de gerar a falha ou o evento

em estudo. Na aplicação de foco desse trabalho, a variável de interesse é o tempo de permanência do estudantes no curso de Estatística com dois possíveis desfechos: conclusão ou evasão.

Uma abordagem para o problema de riscos competitivos é via **função de risco de causa específica** que designa uma função de risco discreta para cada uma das causas. Seja $R \in \{1, \ldots, m\}$ uma variável aleatório que denote a causa do evento acontecer e $T \in \{1, \ldots, q+1\}$ o tempo discreto em que acontece o evento, a função de risco específica para cada causa r é definida por

$$\lambda_r(t \mid \mathbf{x}) = P(T = t, R = r \mid T \geqslant t, \mathbf{x}) \tag{1.3}$$

logo, as m funções de risco $\lambda_1(t \mid \mathbf{x}), \ldots, \lambda_m(t \mid \mathbf{x})$ podem ser combinadas para obter a função de risco total definida como

$$\lambda(t \mid \mathbf{x}) = \sum_{r=1}^{m} \lambda_r(t \mid \mathbf{x}) = P(T = t \mid T \geqslant t, \mathbf{x}).$$

Para o modelo de riscos competitivos também podemos obter as funções de sobrevivência, probabilidade e probabilidade do evento específico,

$$S(t \mid \mathbf{x}) = P(T > t \mid \mathbf{x}) = \prod_{s=1}^{t}(1 - \lambda(s \mid \mathbf{x}))$$
$$P(T = t \mid \mathbf{x}) = \lambda(t \mid \mathbf{x}) S(t-1 \mid \mathbf{x})$$
$$P(T = t, R = r \mid \mathbf{x}) = \lambda_r(t \mid \mathbf{x}) S(t-1 \mid \mathbf{x}).$$

Após definir a funções básicas de sobrevivência para riscos competitivos, estimamos os parâmetros do modelo de regressão baseados na função de distribuição multinomial. Para mais detalhes do modelo de risco competitivo ver (TUTZ; SCHMID, 2016)[Cap. 8].

1.4.1 Modelo de regressão multinomial

Nesta seção escrevemos a função de risco para mais de duas causas do evento acontecer e descrevemos o processo para estimar os parâmetros do modelo, sendo o tempo até a ocorrência do evento a variável resposta, baseados no modelo multinomial. Primeiro consideremos o modelo de risco competitivo de causa específica definido por:

$$\lambda_r(t \mid \mathbf{x}) = \frac{\exp(\gamma_{0tr} + \mathbf{x}^{\mathbf{T}}\gamma_{\mathbf{r}})}{1 + \sum_{l=1}^{m} \exp(\gamma_{0tl} + \mathbf{x}^T\gamma_l)}$$

em que $t = 1, \ldots, q$ e $r = 1, \ldots, m$. $\gamma_{01j}, \ldots, \gamma_{0qj}$ são os parâmetros que caracterizam a função de causa específica de base (*baseline*). $\sum_{j=0}^{m} \lambda_j(t|\mathbf{x}) = 1$.

Segundo (SCHMID; BERGER, 2021) o modelo de risco de causa específica para tempo discreto é equivalente a um modelo de regressão logístico multinomial com $m + 1$ categorias, com uma categoria de referência.

Para estimar os parâmetros do modelo de riscos competitivos se utiliza a função de distribuição multinomial, para mais detalhes da estimação de máxima verossimilhança do modelo de regressão multinomial conferir (CZEPIEL, 2002).

1.4.2 Seleção de preditores

Para a escolha dos preditores usamos o método passo a passo (*stepwise*) baseado no critério de informação de Akaike, o AIC. O método passo a passo baseia-se em subtrair ou adicionar variáveis no modelo base, deixando as variáveis mais significativas segundo o critério do AIC.

$$AIC = 2k - 2\log(L)$$

em que k é o número de parâmetros no modelo estatístico e L é o valor máximo da função de verossimilhança para o modelo estimado. O modelo melhor ajustado é o modelo com menor AIC.

1.4.3 Análise de resíduos

Como apresentamos na seção 1.3.1, a estimativa dos parâmetros pelo método de máxima verossimilhança para o modelo de risco discreto se basea na representação binária e multinomial das transições entre categorias. Porém, quando se quer analisar os resíduos, deve se levar em conta que os dados originais consistem em n observações independentes (t_i, δ_i, x_i), que provem de dados de sobrevivência. Portanto, os desvios e resíduos obtidos diretamente do ajuste de regressão de um modelo logístico ou multinomial não são apropriados para avaliar o ajuste do modelo. Em (TUTZ; SCHMID, 2016)[pág.80], se apresenta a construção dos resíduos de Pearson, deviance e martingale, e as estatísticas de qualidade de ajuste deviance e Pearson para o modelo de sobrevivência que tem em conta os dados censurados e não censurados.

Para um modelo sem dados censurados, temos que o tempo discreto $T_i \in 1, \ldots, q$ em que $T_i = (T_{i1}, \ldots, T_{im}) \sim Multinomial(n_i, (\pi_1, \ldots, \pi_m))$ em que i indica o indivíduo em estudo, q o número de períodos do estudo e T_{is} recebe o valor 0 se o indívio estava em risco no tempo s e 1 se o indíviduo i falhou no tempo s períodos dado n_i observações no mesmo indivíduo. A distribuição multinomial T_i é dada por:

$$f(t_{i1}, \ldots, t_{im}) = \begin{cases} \dfrac{n_i!}{t_{i1}!, \ldots, t_{im}!} \pi_1^{t_{i1}} \ldots \pi_m^{t_{im}} & t_{im} \in \{0, \ldots, n_i\}, \quad \sum_{r=1}^{m} n_r = n. \\ 0 & \text{caso contrário} \end{cases}$$

em que $\pi = (\pi_1, \ldots, \pi_k)^T$ é o vetor de probabilidades, tal que, $\pi \in [0,1]$ e $\sum_i \pi_i = 1$.

Primeiro analisemos o caso sem censura e com n indivíduos no estudo, em que cada um dos indivíduos tem n_i observações. Com $\hat{\beta}$ denotando o estimador de máxima verossimilhança, obtemos o risco ajustado do modelo de risco competitivo $\hat{\lambda}(t \mid \mathbf{x}_i) = \sum_{r=1}^{m} \hat{\lambda}_r(t \mid \mathbf{x}_i) = \sum_{r=1}^{m} h(\hat{\gamma}_{r0t} + \mathbf{x}_i^T \hat{\boldsymbol{\gamma}}_r)$ e as probabilidades ajustadas:

$$\begin{aligned}\hat{\pi}_{it} = \hat{P}(T = t \mid \mathbf{x}_i) &= \hat{\lambda}(t \mid \mathbf{x}_i) \prod_{s=1}^{t-1} (1 - \hat{\lambda}(s \mid \mathbf{x}_i)) \\ &= \sum_{r=1}^{m} \hat{\lambda}_r(t \mid \mathbf{x}_i) \prod_{s=1}^{t-1} \left(1 - \sum_{r=1}^{m} \hat{\lambda}_r(s \mid \mathbf{x}_i)\right).\end{aligned}$$

Além disso, seja D que denota a estatística *deviance*, uma medida apropriada para medir o ajuste do modelo

$$D = 2 \sum_{i=1}^{n} n_i \sum_{t=1}^{k} p_{it} \log\left(\frac{p_{it}}{\hat{\pi}_{it}}\right)$$

com os resíduos de desvio quadrático correspondentes dados por

$$r_{D,i}^2 = 2 n_i \sum_{t=1}^{k} p_{it} \log\left(\frac{p_{it}}{\hat{\pi}_{it}}\right)$$

quanto menor é a soma dos resíduos *deviance*, melhor o ajuste do modelo.

Quando trabalhamos com covariáveis continuas, na maior parte das vezes o $n_i = 1$, para o caso dos resíduos deviance ao quadrado, eles se escrevem em termos do modelo logístico com $(y_{i1}, \ldots, y_{it_i}) = (0, \ldots, 0, 1)$ que denota a transição entre períodos.

$$r_{D,i}^2 = 2 \sum_{s=1}^{t_i} \left\{ y_{is} \log\left(\frac{y_{is}}{\hat{\lambda}_{is}}\right) + (1 - y_{is}) \log\left(\frac{1 - y_{is}}{1 - \hat{\lambda}_{is}}\right) \right\},$$

o que mostra que em cada ponto de tempo a discrepância entre os dados e o ajuste é medida por $\log(y_{is}/\hat{\lambda}_{is})$ se $y_{is} = 1$ e por $\log((1 - y_{is})/(1 - \hat{\lambda}_{is}))$ se $y_{is} = 0$. Uma transformação dos resíduos deviance, são os resíduos deviance ajustados que estão próximos de uma distribuição normal no caso de ter um modelo bem ajustado.

$$\begin{aligned} d_i = &\sum_{s=1}^{t_i} \left\{ sinal(y_{is} - \hat{\lambda}_{is}) \sqrt{y_{is} \log\left(\frac{y_{is}}{\hat{\lambda}_{is})}\right) + (1 - y_{is}) \log\left(\frac{1 - y_{is}}{1 - \hat{\lambda}_{is})}\right)} \right\} \\ &+ \sum_{s=1}^{t_i} \left\{ \frac{1 - 2\hat{\lambda}_{is}}{\sqrt{\hat{\lambda}_{is}(1 - \hat{\lambda}_{is}) \times 36}} \right\}. \end{aligned} \tag{1.4}$$

Ao considerar essa classe de resíduo, o ajuste do modelo pode ser avaliado inspecionando gráficos quantil-quantil normal. No caso de censura os resíduos deviance ajustados tem a mesma forma que na equação 1.4.

O resíduo *martingale* se define por:

$$m_i = \sum_{s=1}^{t_i} (y_{is} - \hat{\lambda}_{is}), \quad i = 1, \ldots, n$$

e o resíduo *Cox-Snell* se define por:

$$\hat{\Lambda}_i = \sum_{s=1}^{t_i} \hat{\lambda}_{is}, \quad i = 1, \ldots, n.$$

Por fim, podemos dizer que um modelo é bem ajustado se inclui todos os preditores relevantes, os resíduos de martingale são aleatórios e não correlacionados com os valores das covariáveis continuas; o ajuste do estimador de Kaplan-Meier para os resíduos Cox-Snell apresentam uma tendência a uma distribuição exponencial padrão; a distribuição dos resíduos deviance ajustados estão perto da distribuição normal e a estatística deviance é menor em comparação com outros modelos ajustados. Para mais detalhes constatar a seção de avaliação do modelo no livro (TUTZ; SCHMID, 2016)[Cap. 4].

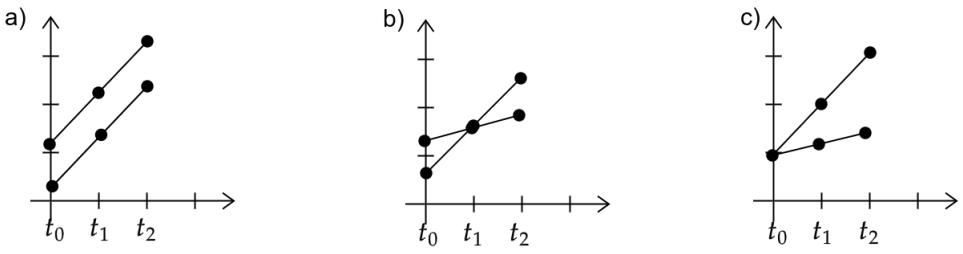
2 Análise de Dados Longitudinais

Os dados longitudinais são observações feitas em instantes de tempo diferentes da variável em estudo, em cada indivíduo sob investigação. Segundo (SINGER; NOBRE; ROCHA, 2018), os estudos longitudinais são um caso especial dos estudos conhecidos como medidas repetidas, com interesse quando o objetivo do pesquisador é avaliar mudanças globais quanto individuais ao longo do tempo. Neste contexto, as observações de um mesmo indivíduo não podem ser tratadas como independentes, então uma maneira de avaliar as variações das medidas de um mesmo indivíduo e modelar a correlação entre elas, é introduzir no modelo os chamados efeitos aleatórios.

2.1 Modelo Linear Misto

Os modelos lineares mistos abordam as estruturas lineares das observações medidas no mesmo indivíduo. A Figura 2 apresenta as classes de estrutura entre indivíduos. Se exibem três medidas da variável de interesse, para cada indivíduo em estudo. Na Figura 2 a) as medidas têm um comportamento linear, onde o coeficiente angular da reta não muda entre indivíduos, ou seja, se pode usar um modelo linear misto com intercepto aleatório para modelar os dados. Na Figura 2 b), as medidas têm um comportamento linear, porém o coeficiente angular da reta e o intercepto variam entre indivíduos, neste caso, se pode usar um modelo linear misto com intercepto e coeficiente angular aleatório. Na Figura 2 c) as medidas têm um comportamento linear, em a coeficiente angular, mas o intercepto é fixo entre indivíduos, assim, se pode usar um modelo linear misto com apenas o coeficiente angular aleatório.

Figura 2 – Ilustração de dados longitudinais. No desenho observamos os perfis para dois indivíduos, onde as medidas ao longo do tempo têm o comportamento linear. a) Muda o intercepto e permanece constante a inclinação da linha. b) Muda o intercepto e a inclinação da linha. c) O intercepto permanece constante e muda a inclinação da linha.



Fonte: Produzido pelo autor

Segundo (HARVILLE, 1977), um problema de longa data dos modelos lineares mistos é a estimativa dos componentes de variância dos efeitos aleatórios. No artigo (HARVILLE, 1977) apresenta um resumo das estimações de máxima verossimilhança para estimar os componentes de variância, em particular para o modelo chamado de modelagem em dois estágios. Em seguida se apresenta o modelo linear misto em dois estágios.

- **Etapa 1**: Ajustar um modelo de regressão para cada indivíduo.
- **Etapa 2**: Explicar a variabilidade dos coeficientes de regressão, específicos de cada indivíduo, usando variáveis conhecidas (efeitos fixos).

Seja W_{ij} a resposta para o i-ésimo indivíduo medido no tempo, $i = 1, \ldots, n$ e $j = 1, \ldots, n_i$. Assim, $\mathbf{W}_i = (W_{i1}, \ldots, W_{in_i})^T$ é o vetor de respostas para o indivíduo i, também chamado de perfil de resposta individual.

Desta forma o modelo $\mathbf{W}_i = \boldsymbol{\beta}_i \mathbf{Z}_i + \boldsymbol{\epsilon}_i$ descreve a variabilidade dentro de cada indivíduo, em que $\mathbf{Z}_i$ é uma matriz de ordem $n_i \times q$ de covariáveis conhecidas, $\boldsymbol{\beta}_i$ é um vetor q-dimensional com os coeficientes de regressão sujeito-específico e $\boldsymbol{\epsilon}_i$, é assumido como $N(0, \boldsymbol{\Sigma}_i)$, em que $\boldsymbol{\Sigma}_i$ é uma matriz variância e covariância. Em resumo, a etapa 1 descreve a variabilidade dentro do indivíduo.

Na etapa 2, se modela a variabilidade entre os indivíduos. Essa variabilidade pode ser modelada, se os β_i estiverem relacionados com covariáveis conhecidas. $\mathbf{K}_i$: é uma matriz $q \times p$ de variáveis conhecidas $\boldsymbol{\beta}$: é um vetor p-dimensional de parâmetros de regressão desconhecidos $\mathbf{b}_i \sim N(0, \mathbf{D})$, onde $\mathbf{D}$ é uma matriz de ordem de variâncias e covariâncias $q \times q$.

$$\text{Estagio } 1 : W_i = \mathbf{Z}_i \boldsymbol{\beta}_i + \epsilon_i$$
$$\text{Estagio } 2 : \beta_i = \mathbf{K}_i B + \mathbf{b}_i.$$

2.1.1 Modelo de Laird-Ware

Os autores (LAIRD; WARE, 1982) expõem que o modelo em dois estágios têm várias características desejáveis, como a não exigência de dados balanceado e a modelagem explícita entre e dentro do indivíduo. Mas também argumentam que muitos estatísticos da época desconheciam os recentes desenvolvimentos metodológicos, que permitem uma abordagem unificada para a formulação dos modelos lineais mistos, desde este artigo, o modelo unificado se conhece como o modelo linear misto de Laird-Ware.

$$\mathbf{W}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i b_i + \boldsymbol{\epsilon}_i$$

para $i = 1, 2, \ldots, n$, β vetor de efeitos fixos; $b_i \sim N(0, \sigma_b^2 D)$ efeitos aleatórios; $\epsilon \sim N(0, \sigma^2 I)$ erro aleatório; $cov(\epsilon_i, b_i) = 0$; W_i é um vetor de repostas $n_i \times 1$ do i-ésimo indivíduo, também

chamado de grupo ou sujeito; X_i é a matriz de desenho das variáveis explanatórias, também chamadas de covariáveis ou efeitos fixos; Z_i é uma matriz de desenho dos efeitos aleatórios. Segundo (SALAZAR; CORREA, 2016) os efeitos fixos modelam características populacionais e os efeitos aleatórios modelam características individuais.

2.2 Estimação

Segundo (LAIRD; WARE, 1982) não é possível utilizar os métodos clássicos para estimar os parâmetros do modelo linear misto, como, por exemplo, o método de mínimos quadrados ordinários, usados na regressão linear quando a variância entre os indivíduos é constante; o método de mínimos quadrados ponderados, usado quando na regressão linear quando não é possível estabilizar a variância, então designamos um peso diferente para cada um dos indivíduos com o fim estimar os parâmetros o método de mínimos quadrados generalizados, usado quando o modelo de regressão linear precisa de outros modelos estatísticos diferentes da distribuição normal para estimar o erro, finalmente temos o método de máxima verossimilhança que não se pode usar diretamente porque temos duas fontes de variabilidade, uma entre indivíduos e a outra dentro dos indivíduos. Segundo (DEMIDENKO, 2013) menciona que, as metodologias já ditas acima tendem a levar a estimadores superestimados, portanto, que se tem que fazer uma modificação do método de máxima verossimilhança, para poder estimar às duas fontes de variabilidade. Este método é chamado de máxima verossimilhança restrita ou máxima verosimilhança residual. (LAIRD; WARE, 1982) utilizam uma modificação da função de verossimilhança usual, usando mínimos quadrados residuais para obter o método de máxima verossimilhança restrita (MVR) que ajuda a ter estimadores não superestimados. Logo de definir o modelo linear misto, o objetivo é estimar β e predizer as variáveis aleatórias b_i, supondo que a distribuição dos efeitos aleatórios seja normal.

2.2.1 Máxima Verossimilhança Restrita

Nesta seção apresentamos o método de máxima verossimilhança restrita de (DEMIDENKO, 2013). O modelo linear geral, pode ser expresso como:

$$\mathbf{W} \sim N(\mathbf{X}\boldsymbol{\beta}, V(\theta))$$

com $\mathbf{X}$ sendo uma matriz $n \times m$ de posto completo e V uma matriz de covariância, que dependem do parâmetro θ. No método de estimação de MVR maximizamos o logaritmo da função de verossimilhança para o vetor de resíduos $\hat{\xi} = \mathbf{W} - \mathbf{X}\hat{\boldsymbol{\beta}}$, com $\hat{\beta} = (\mathbf{X}'V^{-1}\mathbf{X})^{-1}\mathbf{X}'V^{-1}\mathbf{W}$. É importante lembrar que $\mathbf{W}$ tem distribuição normal multivariada, $\hat{\boldsymbol{\beta}}$ e $\hat{\xi}$ são funções lineares em função de $\mathbf{W}$ também com distribuição normal

multivariada. Podemos observar, também, que $\boldsymbol{\beta}$ e ξ são independentes:

$$\begin{aligned} cov(\mathbf{X}V^{-1}\mathbf{W}, \hat{\xi}) &= \mathbf{X}'V^{-1}V[I - V^{-1}\mathbf{X}(\mathbf{X}'V^{-1}\mathbf{X})^{-1}\mathbf{X}'] \\ &= \mathbf{X}' - \mathbf{X}'V^{-1}\mathbf{X}(\mathbf{X}'V^{-1}\mathbf{X})^{-1}\mathbf{X}' = 0 \end{aligned}$$

implicando que a função de verossimilhança de $\mathbf{W}$ é o produto das funções de verossimilhança de $\hat{\xi}$ e $\hat{\boldsymbol{\beta}}$. Mas $\hat{\boldsymbol{\beta}} \sim N(\hat{\boldsymbol{\beta}}, (\mathbf{X}'V^{-1}\mathbf{X})^{-1})$, portanto, a função de verossimilhança do vetor residual $\hat{\xi}$,

$$\begin{aligned} l(\hat{\xi}, \theta) &= l(W, \theta) - l(\hat{\boldsymbol{\beta}}, \theta) \\ &= -\frac{1}{2}\{\ln|\mathbf{X}'V^{-1}\mathbf{X}| + ln|V| + (\mathbf{W}\mathbf{X}\boldsymbol{\beta})'V^{-1}(\mathbf{W} - \mathbf{X}\boldsymbol{\beta}) - (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'V^{-1}X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\} \end{aligned}$$

como

$$(\mathbf{W} - \mathbf{X}\boldsymbol{\beta})'V^{-1}(\mathbf{W} - \mathbf{X}\beta) = (\mathbf{W} - \mathbf{X}\beta)'V^{-1}(\mathbf{W} - \mathbf{X}\beta) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'V^{-1}\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

o logaritmo da função de verosimilhança corresponde a

$$l(\hat{\xi}, \theta) = -\frac{1}{2}[ln|\mathbf{X}'V\mathbf{X}| + ln|V| + \hat{\xi}'\hat{\xi}]$$

a maximização desta função é equivalente a

$$l_R(\boldsymbol{\beta}, \theta) = -\frac{1}{2}[ln|\mathbf{X}'V^{-1}\mathbf{X}| + ln|V| + (\mathbf{W} - \mathbf{X}\boldsymbol{\beta})'V^{-1}(\mathbf{W} - \mathbf{X}\boldsymbol{\beta})]$$

porque a maximização de l_R por β é $\hat{\xi} = W - X\hat{\beta}$. A função l_R é chamada de logaritmo da função de verossimilhança residual.

Nota: a diferença entre l_R e a função de log-verossimilhança usual é o termo $-\frac{1}{2}ln|\mathbf{X}'V\mathbf{X}|$. Como o l_R não é uma função de log-verossimilhança, desse modo a matriz de covariância, para θ não pode ser derivado como a inversa da esperança da segunda derivada. Entretanto, assintoticamente a função de MV e MVR e a sua correspondente matriz de covariância vão coincidir.

2.3 Diagnóstico

Segundo (SINGER; NOBRE; ROCHA, 2018) a análise de resíduos e análise de sensibilidade são ferramentas importantes para avaliar o ajuste de qualquer modelo estatístico de um determinado conjunto de dados, para verificar a validade das suposições adotadas e consequentemente garantir a confiabilidade das inferências nele baseadas. Nos modelos lineares mistos, a análise dos resíduos precisa ter em conta mais fontes de variação que nos modelos lineares clássicos, associados ao erro ($\boldsymbol{\epsilon}$) e os efeitos aleatórios (b). Para fazer a análise dos residuais do modelo longitudinal, consideramos (SINGER; NOBRE; ROCHA, 2018)[p.64-74] que propõe o uso de três categorias de resíduos:

i) **Resíduos marginais**: $\hat{\xi} = y - \mathbf{X}\hat{\boldsymbol{\beta}}$, preditores dos erros marginais $\xi = Y - E(Y) = Y - \mathbf{X}\boldsymbol{\beta}$.

ii) **Resíduos condicionais**, $\hat{\boldsymbol{\epsilon}} = y - X\hat{\boldsymbol{\beta}} - Z\hat{b}$: preditores dos erros condicionais $\epsilon = y - E(y|b) = y - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}b$.

iii) **Resíduos de efeitos aleatórios**: $\mathbf{Z}\hat{b}$, preditores dos efeitos aleatórios, $\mathbf{Z}b = E(\mathbf{W}|b) - E(Y)$.

Para avaliar a linearidade dos efeitos no modelo linear mistos se constroem dois gráficos, o primeiro é os resíduos marginais padronizados contra os valores de cada variável, o segundo é os resíduos marginais padronizados versus os valores ajustados. Além disso, construímos o gráfico dos resíduos marginais padronizados em relação aos índices das observações intra-unidades amostrais, este gráfico ajuda na detecção de observações discrepantes. Para avaliar a homoscedasticidade dos erros condicionais, se sugere a construção de gráficos dos resíduos condicionais padronizados contra os valores padronizados e para detectar observações discrepantes nos resíduos condicionais se usa gráfico dos resíduos condicionais padronizados em comparação aos índices de observações. Os gráficos quantil-quantil dos resíduos condicionais minimamente confundidos padronizados ou só padronizados quando não há confundimento podem ser utilizados para avaliação da normalidade. Por fim, para os resíduos de efeitos aleatórios se deve validar que seguem uma distribuição gaussiana q-dimensional usando o gráfico quantil-quantil baseado na distribuição qui-quadrado, para identificar pontos discrepantes se faz o gráfico de resíduos de efeitos mistos em relação ao os índices das unidades amostrais.

3 Modelagem conjunta

Anteriormente, nos capítulos 1 e 2 apresentamos uma revisão dos conceitos da análise de sobrevivência para tempo discreto e de dados longitudinais. Agora, o nosso interesse, é encontrar uma metodologia, que permita analisar o efeito de uma variável longitudinal na curva de sobrevivência para riscos competitivos, esta metodologia é chamada de modelagem conjunta.

Para estudar a modelagem conjunta discreta, exploraremos um pouco a literatura sobre a modelagem conjunta continua, como proposito de aplicar algumas ferramentas e conceitos na modelagem conjunta discreta.

No trabalho de (LIMA, 2007) são apresentados quatro modelos conjuntos contínuos. O primeiro modelo, é o modelo desenvolvido por (WULFSOHN; TSIATIS, 1997), em que o modelo de Cox é usado para o processo de sobrevivência e o modelo linear misto para o processo longitudinal. A estimação da variável resposta longitudinal é introduzida no modelo de Cox, como variável explicativa, para estimar os parâmetros deste modelo é usado o algoritmo EM na função de máxima verossimilhança dos dados observados. O segundo modelo, é o modelo conjunto de (HENDERSON; DIGGLE; DOBSON, 2000), este modelo é dividido em duas partes, a parte de sobrevivência e a parte longitudinal, em que em cada parte se incorpora uma variável latente, ao juntar às duas variáveis latentes, elas geram um processo latente normal bi-variado, para estimar os parâmetros do modelo, é usado o algoritmo EM, do mesmo modo que no modelo anterior. O terceiro modelo, é o modelo conjunto de (DIGGLE; SOUSA; CHETWYND, 2008), neste modelo o processo de sobrevivência é transformado por meio da função logarítmica e leva-se em consideração o modelo normal multivariado para a variável (W, S) em que W é a variável longitudinal e $S = \ln T$ é a variável de sobrevivência, a estimação dos parâmetros baseia-se na distribuição normal multivariada. Por fim, o quarto modelo, o modelo conjunto em dois estágios de (VONESH; GREENE; SCHLUCHTER, 2006), sendo um modelo mais simples do que os outros três, porque não precisam de cômputos complexos. No primeiro estágio se estimam os parâmetros para o processo longitudinal e no segundo estágio se utiliza os valores ajustados da variável longitudinal como variável explicativa dependente do tempo no modelo de Cox.

No artigo de (IBRAHIM; CHU; CHEN, 2010) se mostra uma revisão dos modelos conjuntos aplicados num banco de dados sobre ensaios clínicos de câncer, em que biomarcadores longitudinais estão associados com o tempo de sobrevivência. Neste estudo se comparam três modelos: o modelo de Cox sem covariável longitudinal, o modelo de Cox com covariável longitudinal e o modelo conjunto em dois estágios, apresentando uma melhor estimativa para a curva de sobrevivência o modelo em dois estágios, devido a que usa os efeitos estimados para estimar a curva de sobrevivência e não os dados observados da variável longitudinal, com grande variabilidade de erro pelas medidas dentro de cada indivíduo.

Outro trabalho interessante que aplicam os modelos conjuntos é a tese de (MAIORANO, 2018), que propõe analisar dados para avaliação de desfechos clínicos do parto com

modelagem conjunta para tempo contínuo. Neste estudo, utilizam-se as generalizações da distribuição Weibull com três parâmetros para o processo de sobrevivência e os modelos lineares mistos para o processo longitudinal.

Adicionalmente, os modelos conjuntos, para tempo contínuo, têm sido estudados no artigo (PAPAGEORGIOU et al., 2019) onde além das técnicas de estimação do modelo conjunto, se apresenta o software disponível para aplicação dos modelos. Segundo (CROWTHER, 2014) a abordagem mais comum em análise conjunta é juntar os modelos lineares mistos com o modelo de riscos proporcionais, por efeitos aleatórios compartilhados, que ajuda a caracterizar a associação entre os dois processos. Na tese de doutorado de (CROWTHER, 2014) se pode encontrar a metodologia de parâmetro compartilhado incorporando o modelo de sobrevivência paramétrico flexível Royston-Parmar.

Além dos modelos conjuntos para um só risco, nosso interesse principal são os modelos de risco competitivo. Segundo (MONDAL, 2017) para os modelos de risco competitivo tem-se duas abordagens amplamente usadas, a primeira é estudar cada uma das causas ou riscos usando o modelo de riscos proporcionais de Cox, e a segunda abordagem é modelar a função de risco associada à função de incidência acumulada. O modelo baseado na função de incidência acumulada é chamado de modelo de riscos de sub-distribuição proporcional. Usando às duas abordagens dos riscos competitivos e trabalhando com dados longitudinais, se gera o modelo conjunto de riscos competitivos, no qual, primeiro usamos o modelo linear misto para estimar os efeitos aleatórios da variável longitudinal e logo introduzimos os efeitos no modelo de riscos competitivos.

Baseados nos modelos conjuntos para tempo contínuo, surgiu a ideia de trabalhar os modelos conjuntos para tempo discreto, porque na maior parte dos estudos o tempo é medido discretamente, o que queremos neste trabalho é aprofundar na modelagem conjunta, para risco competitivo para tempo discreto. Para isso, primeiro revisamos a literatura de risco competitivo para tempo discreto do livro de (TUTZ; SCHMID, 2016) e o artigo de (SCHMID; BERGER, 2021) que fornecem uma visão geral da metodologia estatística atualmente disponível para a análise de dados para tempo discreto e para evento com causas competitivas.

No trabalho de (WEN; CHEN, 2020) se propõe o modelo risco discreto suficiente, para modelar a função de sobrevivência. O modelo suficiente se entende para a modelagem de riscos competitivos, usando o modelo de regressão logística com a estimação explicita da variância, que usam funções polinomiais ou regressão de *splines* para modelar a variável longitudinal. Daí, surge uma estatística completa e suficiente que se introduz no modelo de sobrevivência, é dai o nome de modelo de risco discreto suficiente.

Neste trabalho, para estimar os parâmetros do modelo conjunto de riscos competitivos para tempo discreto, nos baseamos nas ferramentas utilizadas na modelagem para tempo contínuo, na tese de doutorado (QIU, 2012) e no artigo (QIU et al., 2016) quem apresenta a estimação em dois estágios e a estimação com parâmetro compartilhado para o modelo conjunto de um risco. Como se observou no capítulo 1.4 o modelo de risco competitivo pode ser escrito como:

$$\lambda(t|\mathbf{x}) = \sum_{r=1}^{m} \lambda_r(t|\mathbf{x}) = P(T = t|T \leqslant t, \mathbf{x}) = \sum_{r=1}^{m} \frac{\exp(\gamma_{0tr+\mathbf{X}^T\gamma_r})}{1 + \sum_{i=1}^{m} \exp(\gamma_{0t_i+\mathbf{X}^T\gamma_i})}.$$

3.1 Parâmetro compartilhado

Seja w_{ij} uma medida longitudinal no tempo t_{ij} para o indivíduo i, $i = 1, \ldots, n$ e $j = 1, \ldots, n_i$, com w_{ij} e t_{ij} variáveis continuas, dessa maneira temos que a dupla (w_{ij}, t_{ik}) representa os dados longitudinais e $\tilde{T} = mín(T_i, C_i)$ os dados de sobrevivência. Logo, definimos uma nova variável u_i que não pode ser medida diretamente, ou seja, é uma variável não observada também chamada de variável latente, no modelo que definiremos é chamada de parâmetro compartilhado. O primeiro que faremos é definir a função de probabilidade de $w_{ij}|u_i$ por meio do modelo linear misto.

$$w_{ij} = \beta_0 + \beta_1 t_{ij} + Z_{ij}u_i + \epsilon_{ij}$$

em que β_1 e β_2 são efeitos fixos; $u_i = (u_{i0}, u_{i1})$ é um vetor de efeitos aleatórios sendo u_{i0} o intercepto e u_{i1} a inclinação sob cada indivíduo; $(u_{i0}, u_{i1}) \sim N_2(0, G)$; Z_{ij} é um vetor desenho de dimensão 1×2. Se $Z_{ij} = (1, 0)$ é um modelo com intercepto, se $Z_{ij} = (1, 1)$ implica que o intercepto e a inclinação são requeridos para o modelo; $\epsilon_{i,k} \sim N(0, \sigma^2)$; e $cov(u_i, \epsilon_{i,k}) = 0$. Baixo o suposto de u_i e $\epsilon_{i,k}$ são independentes temos que $w_{ik}|u_i$ e $w_{il}|u_i$ também são independentes. Definimos a distribuição da variável condicional $w_{ik}|u_i$ como:

$$g_w(w_{ij}|u_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{w_{ij} - \beta_0 - \beta_1 t_{ij} - Z_{ij}u_i}{2\sigma^2}\right)$$

A função de distribuição conjunta condicional de $w_i = (w_{i1}, \ldots, w_{in_i})$ dado u_i é

$$g_w(w_i|u_i) = g_w(w_{i1}|u_i), \ldots, g_w(w_{in_i}|u_i) = \prod_{j=1}^{n_i} g_w(w_{ij}|u_i)$$

agora, se os indivíduos são independentes temos:

$$g_w(w|u) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} g_w(w_{ij}|u_i).$$

com a matriz de efeitos aleatórios $u = (u_1, \ldots, u_n)$. Em síntese, escrevemos o modelo de riscos competitivos utilizando a mesma variável latente do modelo longitudinal como:

$$\lambda(t|x) = \sum_{r=1}^{m} \frac{\exp(\gamma_{0tr} + X^T\gamma_r + u_i\eta)}{1 + \sum_{i+1}^{m} \exp(\gamma_{0t_i} + X^T\gamma_i + u_i\eta)}$$

com η o parâmetro para o vetor de efeitos aleatórios u_i. A função de probabilidade para o modelo de riscos competitivos se escreve como:

$$g_y(y_{ij}|u_i) = (1 - \lambda_j)^{y_{ij0}} \lambda_{1j}^{y_{ij1}}, \ldots, \lambda_{mj}^{y_{ijm}}$$

$$g_y(y_i|u_i) = \prod_{j=1}^{t_i} g_y(y_{ij}|u_i).$$

Supondo que os indivíduos são independentes, então a função de distribuição conjunta se expressa:

$$g_y(y|u) = \prod_{i=1}^{n} g_y(y_i|u_i) = \prod_{i=1}^{n} \prod_{j=1}^{t_i} g_y(y_{ij}|u_i).$$

3.2 Estimação duas etapas

Primeiro estimamos os parâmetros para o modelo linear misto da variável longitudinal W_i, com $i = 1, \ldots, n$, e obtemos os efeitos aleatórios preditos $\hat{u}_i$.

$$w_{ij} = \beta_0 + \beta_1 t_{ij} + Z_{ij} u_i + \epsilon_{ij}.$$

Os dados de sobrevivência $y_i = (y_{i1} \ldots y_{it_i})$ são modelados usando a distribuição conjunta $g_w(wi|\hat{u}_i, X_{ij})$ da variável longitudinal w_i, é dizer os valores estimados $\hat{u}_i$ no segundo estágio. Em suma, a função de risco para a causa r, $r = 1, \ldots .m$ se expressa como:

$$\lambda_r(t|x) = \frac{\exp(\gamma_{0tr} + X^T \gamma_r + \hat{u}_i \eta)}{1 + \sum_{i+1}^{m} \exp(\gamma_{0t_i} + X^T \gamma_i + \hat{u}_i \eta)}. \tag{3.1}$$

Embora o modelo conjunto de parâmetro compartilhado e o método de duas etapas tenham a mesma forma para o modelo de dados longitudinais e dados de sobrevivência, os efeitos aleatórios u_i são caracterizados e estimados de forma diferente. Na abordagem de duas etapas, u_i são estimados usando apenas dados longitudinais no primeiro estágio enquanto na abordagem de parâmetro compartilhado todos os parâmetros são estimados simultaneamente a partir da probabilidade conjunta, consequentemente segundo (QIU, 2012) o modelo estimado pelo parâmetro compartilhado ajuda a diminuir o viés dos efeitos aleatórios mais do que o modelo em dois estágios.

Para revisar detalhes do método de estimação em dois estágios e parâmetro compartilhado podemos ver a tese (QIU, 2012)[pag.41-47], nela encontramos a metodologia para o modelo logístico, para o modelo multinomial é necessário trocar a função de verossimilhança da distribuição logística para um risco, pela função de verossimilhança da distribuição multinomial para mais de um risco.

3.3 Diagnóstico

Na modelagem conjunta se deve avaliar os dois processos do modelo, para o processo longitudinal se usa os residuais marginais e condicionais, e para o processo de sobrevivência se usam os residuais matingale, e deviance, ver em (RIZOPOULOS, 2012) para modelos de tempo contínuo e em (TUTZ; SCHMID, 2016) para modelos de tempo discreto.

4 Aplicação

4.1 Descrição do estudo

A evasão (saída definitiva do curso de origem sem conclusão) nos programas de graduação das universidades em geral, são um problema de grande transcendência para o sistema de educação, porque representa um elevado custo econômico e social. Segundo o instituto de Excelência ao Serviço de Ensino Superior (SEMESP, 2021) no *Mapa do Ensino Superior* na edição para 2021, evidenciou que a região de Campinas tem uma taxa de evasão do 33.5% para cursos presenciais. Diante desse contexto, o objetivo desta aplicação é analisar o evento sair do bacharelado em estatística, considerando duas causas, o estudante evade o programa ou o estudante se forma do programa. Para fazer a análise se utiliza a metodologia descrita no capítulo 2 o modelo de riscos competitivos e no 3 modelo de riscos competitivos conjunto.

O conjunto de dados se compõe de todos os alunos ingressantes ao curso de graduação em estatística (439) entre o ano 2012 até 2017, obtidos a partir do questionário sociodemográfico aplicado pela Comissão Permanente de Vestibulares (COMVEST) na inscrição do vestibular, as notas da prova do vestibular nas áreas de português, matemática, geografia, historia, física, biologia e química aplicado também pela COMVEST e as informações acadêmicas fornecidas pela Diretoria Acadêmica (DAC) da UNICAMP. Foram considerador apenas o primeiro ingresso na universidade, via vestibular, de cada estudantes. Registros de re-ingressos foram desconsiderados. Estudantes que mudaram de curso via remanejamento interno foram tratados como evadidos.

Outra característica avaliada foi o coeficiente de rendimento (CR) semestre a semestre do aluno que é dado por uma média ponderada das notas notas das disciplinas cursadas em um dados semestre (incluindo aquelas em que foram reprovados). Por meio da modelagem conjunta queremos avaliar, como o CR influência no tempo de permanência dos alunos no programa de graduação da Estatística.

As variáveis preditoras consideradas nas análises são apresentadas na Tabela 5. As observações com dados faltantes em qualquer variável preditora foram excluídos da análise (o tamanho de mostra resultante é de $n = 422$). Também, tempos de eventos maiores que 12 semestres foram considerados censurados (resultando em uma taxa de censura igual a 1.2%).

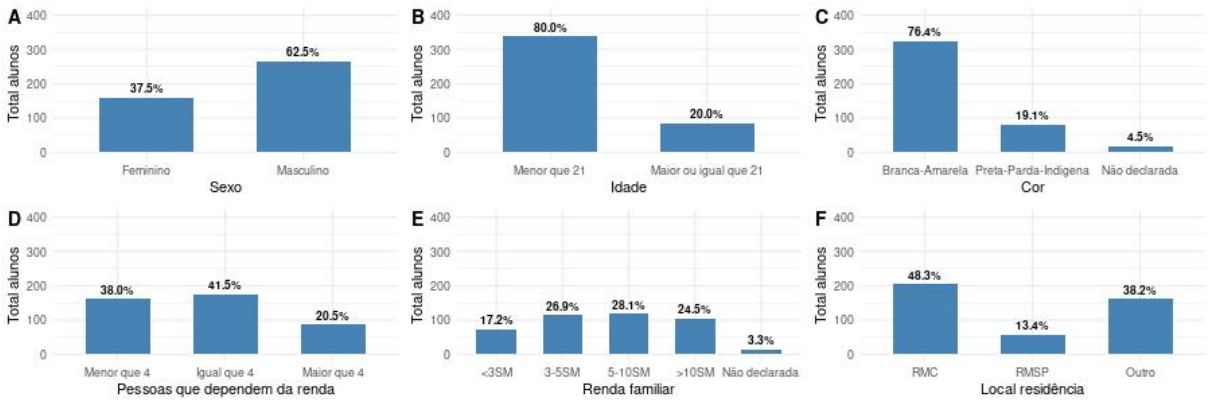
Nas Figuras 3, se apresentam os gráficos de barras a distribuição dos registros segundo cada uma das variáveis preditoras categóricas, onde se observa que entre os anos 2012 e 2017 a maioria dos alunos é do sexo masculino (62.5%), menores de 21 anos (80%), com cor de pele branca ou amarela (76.4%). A renda familiar mensal dos alunos está entre três salários mínimos até mais de dez salários mínimos (79.5%) e dessa renda dependem 4 ou menos pessoas (79.5%). O (86.5%) dos alunos provem da Região Metropolitana de Campinas (RMC) e da Região Metropolitana de São Paulo.

Na Figura 4 se evidencia que a maioria dos alunos que ingressam no programa de

Tabela 5 – Variáveis explanatórias para o tempo de sobrevivência

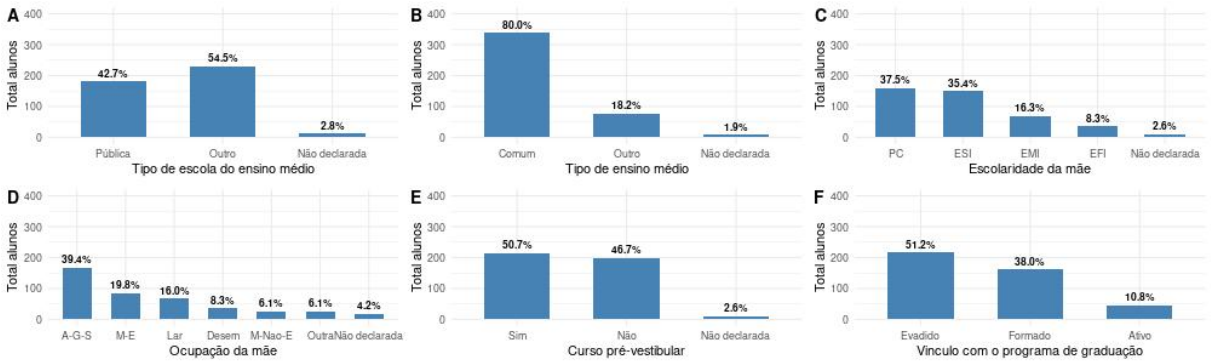
Nome da variável	**Categorias**
Sexo	Feminino
	Masculino
Idade	Menor que 21
	Maior ou igual 21
Cor de pele	Branca-Amarela
	Preta-Parda-Indigena
Pessoãs que dependem da renda familiar	Menos que 4
	4
	Maior que 4
Renda familiar	Menor que 3SM
	3-5 SM
	5-10SM
	Maior que 10SM
Local de residência	Região metropolitana de Campinas (RMC)
	Região metropolina de São Paulo (RMSP)
	Outro locais do pais (Outro)
Ensino Médio	Ensino médio comum (comum)
	curso técnico, para magistério, educação jovem e adultos, certificado de ensino médio pelo ENEM, outros (Outro)
Tipo de escola onde cursou os estudos	Todo em pública (pública)
	Todo em particular, maior parte em escola pública, maior parte em escola particular, no exterior, outra situação (Outro)
Escolaridade da mãe	não estudou, ensino fundamental incompleto (EFI)
	ensino fundamental completo, ensino médio incompleto (EMI)
	ensino médio completo, ensino superior incompleto (ESI)
	ensino superior completo, pós-graduação incompleta, pós-graduação completa (PC)
Ocupação da mãe	proprietárias e altos cargos políticos e/ou administrativos, profissionais liberais, cargos médios de gerência e direção, supervisão de ocupações técnicas ou assemelhadas (A-G-S)
	ocupações não manuais de rotina, supervisão de trabalho manual e ocupações assemelhadas, ocupações manuais especializadas e assemelhadas (M-E)
	ocupações manuais não especializadas (M-Não-E)
	Do lar (LAR)
	Desempregado (Desem)
	Outro (Outro)
Curso pré-vestibular	Sim
	Não
Situação do aluno	Ativo
	Formado
	Evadido
PAAIS programa de ação afirmativa baseado em um acréscimo de nota	Sim
	Não
isenção da taxa do vestibular	Sim
	Não
Ano de ingresso	2012, 2013, 2014, 2015, 2016, 2017, 2018
Notas vestibular Matemáticas, Português, Biologia, Fisica, geografía, historia, quimica	notas padronizadas, sem bônus PAAIS
CR	Coeficiente de rendimento semestral

Figura 3 – Variáveis sociodemográficas dos alunos ingressantes do programa de graduação da estatística nos anos de 2012 a 2017.



Fonte: Produzido pelo autor

Figura 4 – Variáveis sociodemográficas dos alunos ingressantes do programa de graduação da estatística nos anos de 2012 a 2017.

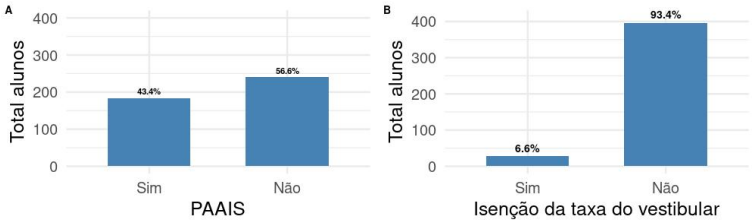


Fonte: Produzido pelo autor

graduação em estatística cursaram o tipo de ensino médio comum (80%) e suas mães tem estudos de pós-graduação e ensino superior incompleto (72.9%), com ocupações de proprietárias, altos cargos políticos ou administrativos, profissionais liberais, cargos médios de gerência e direção, supervisão de ocupações técnicas ou assemelhadas, ocupações não manuais de rotina, supervisão de trabalho manual e ocupações assemelhadas, ocupações manuais especializadas e assemelhadas (65.3%).

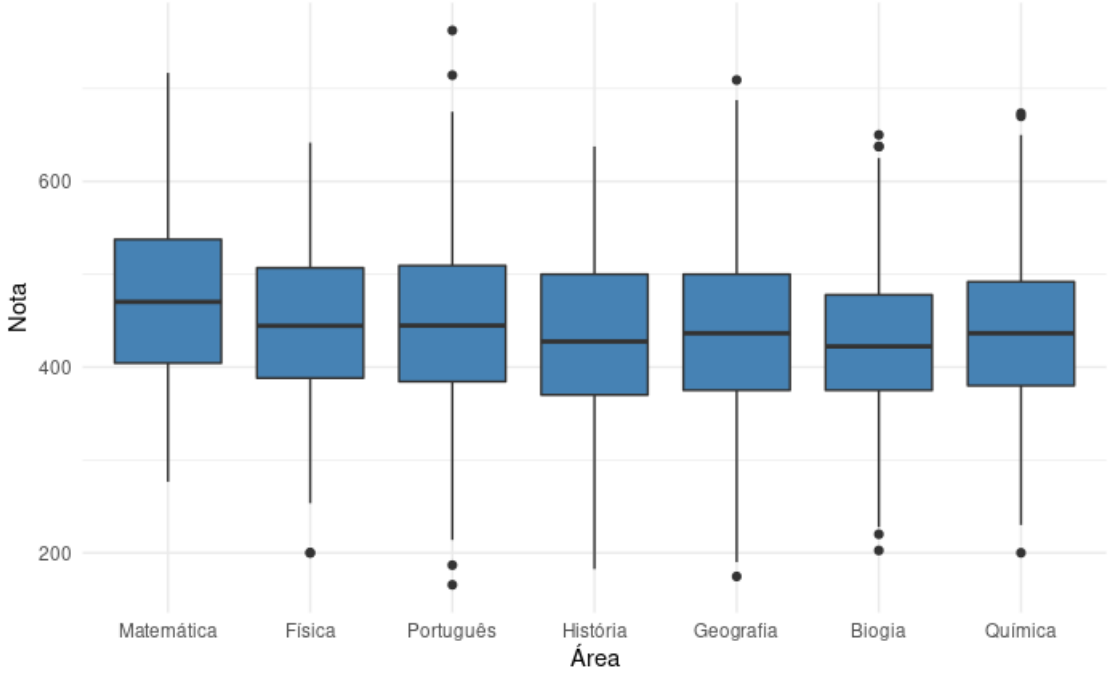
A variável PAAIS (programa de ação afirmativa baseado em um acréscimo de nota) é definida em (KLEINKE, 2006) como "o programa que confere pontos adicionais à nota final do vestibular para candidatos egressos da rede pública e optem por receber esse bônus. Caso esse mesmo candidato se autodeclare preto, pardo ou índio ele pode ainda escolher receber uma segunda pontuação extra em sua nota, relativa ao recorte étnico do programa. Um dos principais objetivos do PAAIS é buscar as excelências escondidas entre os candidatos da escola pública, além de ampliar a diversidade cultural, étnica e de classes sociais entre nossos estudantes". Na Figura 5 mostra-se que o (56.6%) dos alunos receberam o acréscimo de nota e o (6.6%) dos alunos teve isenção da taxa do vestibular.

Figura 5 – Variáveis sociodemográficas dos alunos ingressantes do programa de graduação da estatística nos anos de 2012 a 2017.



Fonte: Produzido pelo autor

Figura 6 – Diagrama de caixa dos resultados da prova do vestibular aplicado pela COMVEST (2012-2017).



Fonte: Produzido pelo autor

Por fim, na Figura 6 se expõem os resultados da prova do vestibular em sete áreas Matemáticas, Física, Português, História, Geografia, Biologia e Química. As notas são padronizadas entre todos os candidatos do vestibular para terem média igual a 500 pontos e desvio padrão igual a 100 pontos. É requisito que todas as áreas do vestibular tenham nota positiva se não, o aluno é eliminado do concurso automaticamente. Para nosso estudo, os dados com nota zero o menor serão tomados como dados faltantes.

Segundo o regimento geral de graduação da UNICAMP no capítulo V Da avaliação do aluno na disciplina, define no artículo 55 o seguinte: "O resultado da avaliação do rendimento escolar é expresso por:

I notas de 0,0 (zero vírgula zero) a 10,0 (dez vírgula zero), computadas até a primeira casa

decimal;

II aprovado por frequência (A) ou reprovado por frequência (R); e

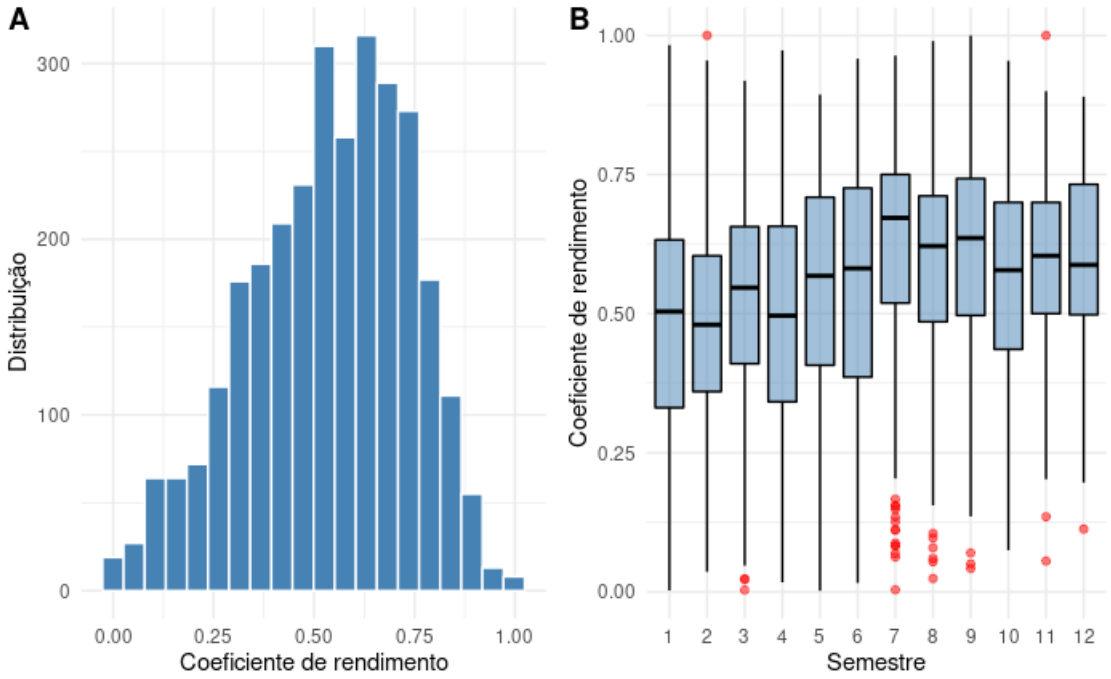
III aprovado por suficiência (S) ou reprovado por insuficiência (I)."

Logo na seção VI do coeficiente de rendimento (CR). Articulo 67. O Coeficiente de Rendimento (CR) é o índice que mede o desempenho acadêmico do aluno ao longo de seu curso e é assim calculado:

$$CR = \frac{\sum_{i=1}^{n} N_i C_i}{10 \sum_{i=1}^{n} C_i} \tag{4.1}$$

em que e N_i = nota relativa a $i-$ésima disciplina dentre as n disciplinas cursadas nesta Universidade; C_i = número de créditos correspondentes a $i-$ésima disciplina (DAC, 2021).

Figura 7 – Distribuição do Coeficiente de rendimento (CR) geral e por semestre cursado.



Fonte: Produzido pelo autor

Na Figura 7 observamos que o CR padronizado entre 0 e 1, com uma distribuição com uma leve assimetria à esquerda. Adicionalmente, o CR apresenta dados atípicos por abaixo da média nos semestres 3, 7, 8, 9, 10 e 11. A tendência média do CR observado se acrescenta entre os semestres 4–7 e depois começa uma queda até o semestre 12.

4.2 Modelos

O objetivo da análise de sobrevivência de riscos competitivos é estudar a ocorrência de duas causas competindo simultaneamente, em nossa pesquisa analisamos as seguintes causas: o aluno se forma do curso de graduação em Estatística ou o aluno evade o curso de graduação em Estatística, o evento sair do curso foi medido ao longo de 12 semestres onde o evento censura ou categoria de referência é o aluno ativo no programa de graduação em Estatística.

4.2.1 Modelo 1: Modelo de sobrevivência para risco competitivos

Na Tabela 6 apresenta as estimativas dos parâmetros do modelo de risco competitivo com tempo discreto ajustado sem o efeito da covariável longitudinal (CR). A variável Evento divide-se em três categorias, o aluno se formou, o aluno evadiu ou o aluno segue ativo no curso de graduação em estatística. A categoria de referência considerada é o evento o aluno segue ativo (censura à direita). Após de todas as etapas terem sido avaliadas, o modelo escolhido é o modelo que inclui as seguintes covariáveis: Ano de Ingresso, Ensino Médio, Sexo, Idade, Cursinho, Nota de Matemáticas, Nota de Biologia. Observamos na Tabela 6 que não é possível estimar as categorias 1 até 7 para a covariável baseline devido a que nesses semestres não temos estudantes formados. As estimativas em geral para os parâmetros do risco se formar são positivas em comparação com o risco de evadir.

Tabela 6 – Estimativa dos parâmetros obtidas do ajuste do modelo de riscos competitivos com tempo discreto para os riscos de se formar e de evadir.

		Risco Formado				Risco Evadido			
Variável	Nível	Estimativa	EP	Z	p-valor	Estimativa	EP	Z	p-valor
baseline	1	-28.13	1620	-	-	-0.21	0.71	-0.29	0.77
	2	-28.08	1647	-	-	0.06	0.71	0.08	0.93
	3	-28.29	1868	-	-	-1.20	0.76	-1.58	0.11
	4	-28.30	1922	-	-	-1.38	0.77	-1.78	0.07
	5	-28.15	1846	-	-	-0.33	0.72	-0.46	0.64
	6	-28.16	1942	-	-	-0.36	0.73	-0.49	0.62
	7	-28.13	1983	-	-	-0.22	0.73	-0.30	0.76
	8	-7.79	1.04	-7.50	0.00	0.46	0.72	0.64	0.52
	9	-7.90	1.04	-7.63	0.00	0.51	0.72	0.70	0.48
	10	-6.76	1.00	-6.79	0.00	0.14	0.75	0.19	0.85
	11	-7.49	1.03	-7.29	0.00	-0.26	0.82	-0.32	0.75
	12	-5.25	0.97	-5.40	0.00	-0.20	0.99	-0.21	0.84
Ano ingresso	2012 (Ref)								
	2013	0.48	0.38	1.24	0.22	-0.45	0.23	-1.99	0.05
	2014	0.99	0.39	2.54	0.01	-0.59	0.24	-2.46	0.01
	2015	2.05	0.44	4.70	0.00	-0.26	0.24	-1.08	0.28
	2016	0.98	0.38	2.61	0.01	-0.91	0.28	-3.27	0.00
	2017	0.69	0.40	1.73	0.08	-0.74	0.26	-2.82	0.00
Ensino médio	Otro (Ref)								
	Publico	0.60	0.26	2.34	0.02	-0.39	0.19	-2.03	0.04
Sexo	Feminino (Ref)								
	Masculino	-0.11	0.21	-0.53	0.59	0.65	0.16	3.97	0.00
Idade	Menores 20 (Ref)								
	Maiores 20	0.50	0.32	1.55	0.12	0.60	0.18	3.38	0.00
Cursinho	Não (Ref)								
	Sim	-0.45	0.21	-2.18	0.03	-0.13	0.15	-0.86	0.39
Nota matemática		0.007	0.00	4.37	0.00	-0.004	0.00	-3.38	0.00
Nota biologia		0.004	0.00	2.93	0.00	-0.001	0.00	-0.81	0.42

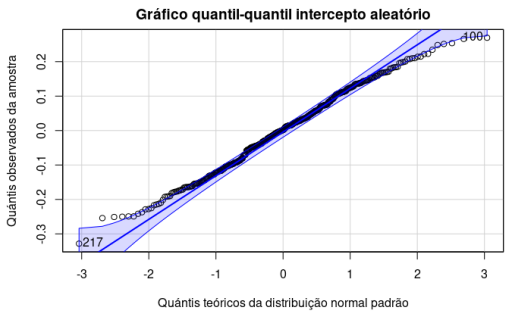
4.2.2 Modelo 2: Modelo de risco competitivo conjunto em dois estágios

Na Tabela 7 mostramos as estimativas dos parâmetros para o modelo linear misto onde a variável resposta longitudinal é o coeficiente de rendimento mensurado ao longo dos semestres e na Tabela 8 apresentamos as estimativas dos parâmetros do modelo conjunto usando a metodologia de dois estágios, onde usamos o efeito do indivíduo do modelo longitudinal 7 no modelo de sobrevivência 6. Para a escolha das variáveis do modelo longitudinal utilizamos o método de Seleção de variáveis passo-a-passo (*stepwise*).

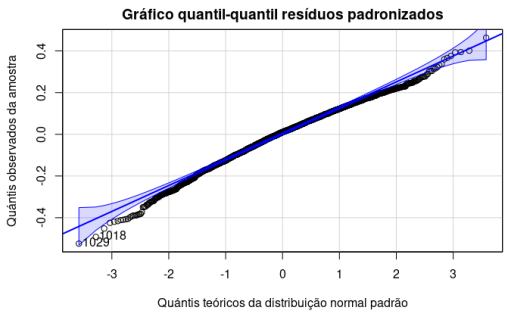
Segundo o capítulo 2 para avaliar o ajuste do modelo longitudinal, temos três tipos de resíduos, os resíduos marginais, os resíduos condicionais e os resíduos de efeitos aleatórios. Na Figura 8 encontramos os seguintes gráficos, na sub-figura a) o gráfico quantil-quantil para os resíduos do efeito aleatório, na sub-figura b) o gráfico quantil-quantil para os resíduos condicionais padronizados e na sub-figura c) o gráfico dos valores ajustados contra os resíduos marginais. Os três gráficos apresentados indicam um modelo razoável, no entanto, a estatística de normalidade não se cumpriu para os resíduos padronizados condicionais.

Figura 8 – Análises dos Resíduos do Modelo Longitudinal ajustado. (a) Gráfico quantil-quantil do erro do efeito (b) Gráfico quantil-quantil dos resíduos (c) Gráfico dos resíduos em relação aos valores ajustados

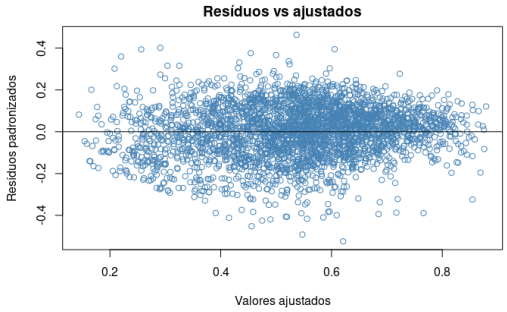
(a) Efeito de erro



(b) Resíduos



(c) Resíduos vs Ajustados



Fonte: Produzido pelo autor

Na Figura 9 apresentamos os resíduos marginais contra as covariáveis continuas, nela não observamos padrão algum que indique uma falta de ajuste no modelo.

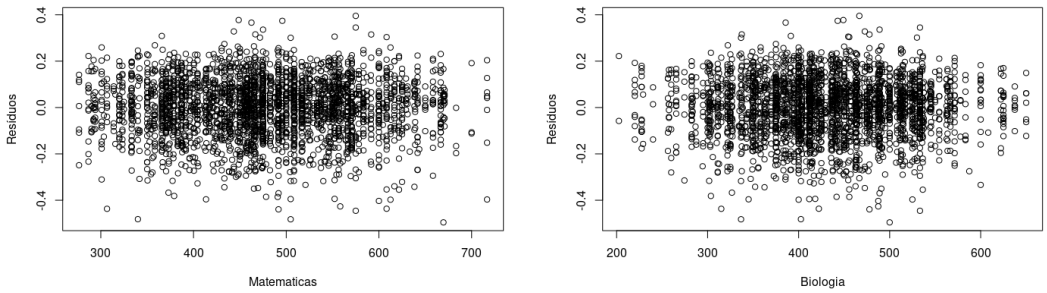
Tabela 7 – Estimativas dos parâmetros obtidos do ajuste do modelo linear misto para o coeficiente de rendimento padronizado.

Variável	Nível	Estimativa	EP	GL	t	p-valor
Intercepto		-0.25	0.07	-3.376	-3.30	0.00
baseline	1 (Ref)					
	2	-0.01	0.01	2,779	-1.36	0.17
	3	0.02	0.01	2,807	2.24	0.03
	4	-0.01	0.01	2,811	-1.50	0.13
	5	0.01	0.01	2,813	1.49	0.14
	6	0.01	0.01	2,813	0.80	0.42
	7	0.05	0.01	2,812	4.96	0.00
	8	0.04	0.01	2,810	3.67	0.00
	9	0.07	0.01	2,803	5.64	0.00
	10	0.04	0.01	2,796	2.74	0.01
	11	0.06	0.02	2,785	3.75	0.00
	12	0.08	0.02	2,773	4.52	0.00
Ano ingresso	2012 (Ref)					
	2013	0.05	0.03	346	1.84	0.07
	2014	0.09	0.03	339	3.36	0.00
	2015	0.10	0.03	352	3.51	0.00
	2016	0.06	0.03	340	2.17	0.03
	2017	0.05	0.03	349	1.87	0.06
Ensino médio	Outro (Ref)					
	Publico	0.11	0.02	333	6.09	0.00
Sexo	Feminino (Ref)					
	Masculino	-0.09	0.02	340	-5.80	0.00
Idade	Menores 20 (Ref)					
	Maiores 20	-0.05	0.02	359	-2.26	0.02
Cursinho	Não (Ref)					
	Sim	-0.03	0.02	347	-1.67	0.10
Nota matemática		0.0005	0.00	344	3.16	0.00
Nota biologia		0.0002	0.00	346	1.93	0.05
Nota física		0.0008	0.00	351	5.30	0.00

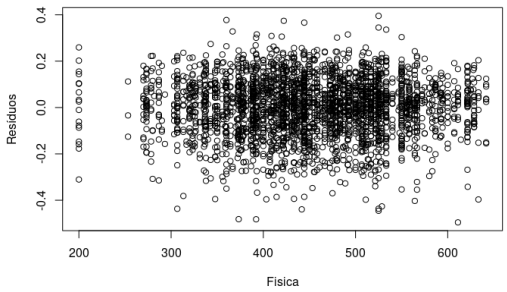
Figura 9 – Gráficos de dispersão, Resíduos Marginais contra as notas nas disciplinas (a) matemática, (b) biologia e (c) física.

(a) Matemáticas

(b) Biologia



(c) Física



Fonte: Produzido pelo autor

Tabela 8 – Estimativa dos parâmetros obtidas do ajuste do modelo conjunto de risco competitivo com tempo discreto em dois estágios.

		Modelo Formado				Modelo Evadido			
Covariável	Níveis	Estimativa	EP	Z	p-valor	Estimativa	EP	Z	p-valor
baseline	1	-38.88	2,141	-	-	0.31	0.75	0.41	0.68
	2	-38.93	2,211	-	-	0.72	0.75	0.96	0.34
	3	-39.10	2,552	-	-	-0.55	0.80	-0.69	0.49
	4	-39.14	2,613	-	-	-0.73	0.81	-0.90	0.37
	5	-39.05	2,514	-	-	0.39	0.76	0.51	0.61
	6	-39.06	2,639	-	-	0.36	0.77	0.47	0.64
	7	-39.10	2,705	-	-	0.56	0.77	0.73	0.47
	8	-17.15	1.72	-9.99	0.00	1.23	0.77	1.60	0.11
	9	-16.93	1.68	-10.10	0.00	1.41	0.78	1.80	0.07
	10	-15.12	1.57	-9.61	0.00	1.03	0.81	1.27	0.21
	11	-15.41	1.57	-9.82	0.00	0.58	0.88	0.67	0.50
	12	-12.82	1.49	-8.62	0.00	0.75	1.04	0.72	0.47
Ano ingresso	2012 (Ref)								
	2013	0.96	0.46	2.10	0.04	-0.36	0.23	-1.56	0.12
	2014	2.56	0.49	5.28	0.00	-0.58	0.25	-2.35	0.02
	2015	3.51	0.54	6.50	0.00	-0.23	0.25	-0.94	0.35
	2016	2.12	0.45	4.67	0.00	-0.90	0.29	-3.15	0.00
	2017	2.09	0.49	4.26	0.00	-0.89	0.28	-3.20	0.00
Sexo	Feminino (Ref)								
	Masculino	-1.13	0.27	-4.26	0.00	0.68	0.17	3.98	0.00
Idade	Menores 20 (Ref)								
	Maiores 20	0.37	0.39	0.97	0.33	0.73	0.19	3.94	0.00
Ensino médio	Outro (Ref)								
	Publica	1.55	0.32	4.84	0.00	-0.57	0.20	-2.84	0.00
Cursinho	Não (Ref)								
	Sim	-0.25	0.24	-1.05	0.29	-0.11	0.16	-0.70	0.48
Nota matemáticas		0.02	0.00	8.08	0.00	-0.006	0.00	-4.33	0.00
Nota biologia		0.01	0.00	4.24	0.00	-0.001	0.00	-0.90	0.37
Efeito aleatório		17.08	1.75	9.77	0.00	-6.72	0.64	-10.49	0.00

4.2.3 Modelo 3: Modelo de risco competitivo com parâmetro compartilhado

Na Tabela 9 apresentamos as estimativas dos parâmetros do modelo longitudinal e na Tabela 10 as estimativas dos parâmetros do modelo conjunto usando a metodologia de parâmetro compartilhado, nesta metodologia se estima simultaneamente o efeito da covariável longitudinal e as covariáveis fixas.

Uma das diferenças entre o modelo de risco competitivo em dois estágios e bivariado apresenta-se na hora de comparar os parâmetros estimados do modelo longitudinal. No modelo de dois estágios o modelo longitudinal pode ter covariáveis diferentes ao modelo de riscos competitivos sem embargo no modelo bivariado se compartilham as covariáveis. Por exemplo na Tabela 7 encontramos o parâmetro estimado da covariável Nota de física e ela não se encontra presente na Tabela 9.

Tabela 9 – Estimativas dos parâmetros obtidas do ajuste do modelo conjunto de risco competitivo com parâmetro compartilhado para o coeficiente de rendimento padronizado.

Covariável	Níveis	Estimativa	DP	t	p-valor
baseline	1	-0,16	0,07	-2,25	0,02
	2	-0,18	0,07	-2,41	0,02
	3	-0,14	0,07	-1,92	0,05
	4	-0,18	0,07	-2,42	0,02
	5	-0,15	0,07	-2,02	0,04
	6	-0,15	0,07	-2,10	0,04
	7	-0,11	0,07	-1,50	0,13
	8	-0,12	0,07	-1,67	0,09
	9	-0,09	0,07	-1,29	0,20
	10	-0,13	0,07	-1,71	0,09
	11	-0,10	0,07	-1,40	0,16
	12	-0,08	0,07	-1,03	0,31
Ano ingresso	2012 (Ref)				
	2013	0,06	0,03	2,06	0,04
	2014	0,10	0,03	3,49	0,00
	2015	0,12	0,03	4,32	0,00
	2016	0,09	0,03	3,25	0,00
	2017	0,08	0,03	2,80	0,01
Ensino médio	Otro (Ref)				
	Publico	0,10	0,02	4,96	0,00
Sexo	Feminino (Ref)				
	Masculino	-0,09	0,02	-5,57	0,00
Idade	Menores 20 (Ref)				
	Maiores 20	-0,05	0,02	-2,23	0,03
Cursinho	Não (Ref)				
	Sim	-0,02	0,02	-1,46	0,14
Nota matemática		0,0009	0,00	7,85	0,00
Nota biologia		0,0004	0,00	3,05	0,00

Tabela 10 – Estimativa dos parâmetros obtidas do ajuste do modelo conjunto de risco competitivo com tempo discreto com parâmetro compartilhado.

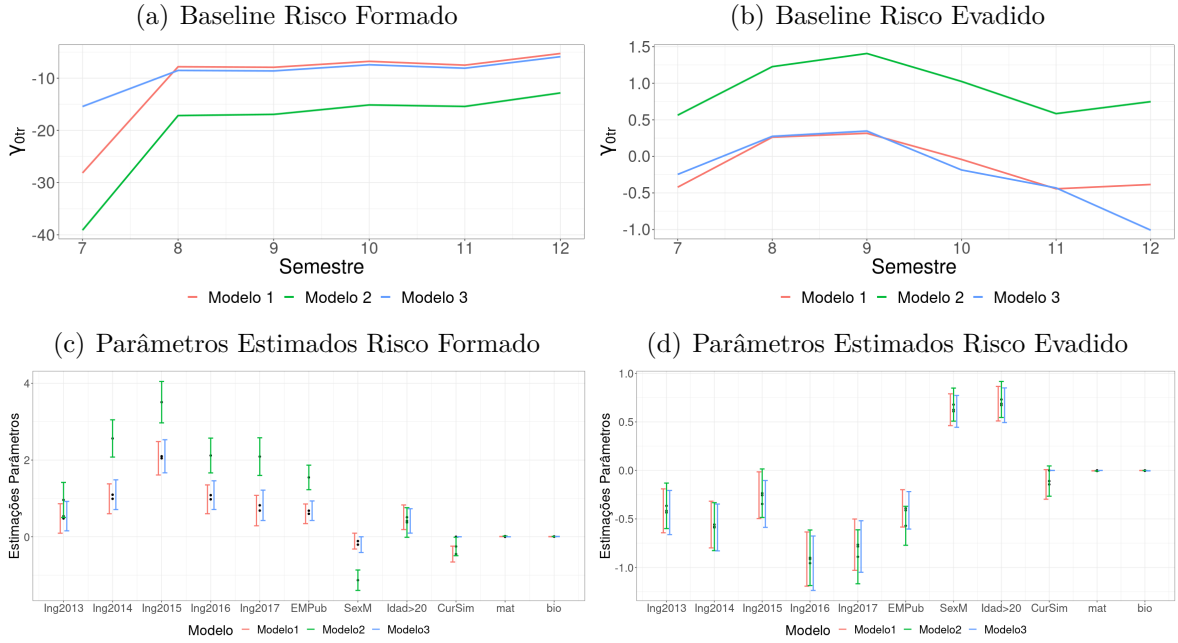
		Modelo Formado				Modelo Evadido			
Covariável	Níveis	Estimativa	EP	t	p-valor	Estimativa	EP	t	p-valor
baseline	1	-15,36	2,31	-6,65	0,00	-0,15	0,72	-0,20	0,84
	2	-15,38	2,38	-6,46	0,00	0,11	0,71	0,16	0,88
	3	-15,39	2,48	-6,20	0,00	-1,16	0,76	-1,52	0,13
	4	-15,39	2,51	-6,13	0,00	-1,36	0,78	-1,75	0,08
	5	-15,39	2,54	-6,07	0,00	-0,31	0,73	-0,43	0,67
	6	-15,41	2,61	-5,91	0,00	-0,37	0,73	-0,51	0,61
	7	-15,42	2,67	-5,77	0,00	-0,25	0,73	-0,34	0,74
	8	-8,51	1,03	-8,27	0,00	0,27	0,73	0,38	0,71
	9	-8,61	1,02	-8,41	0,00	0,35	0,73	0,48	0,63
	10	-7,43	0,99	-7,54	0,00	-0,19	0,76	-0,25	0,81
	11	-8,09	1,02	-7,97	0,00	-0,43	0,82	-0,52	0,60
	12	-5,88	0,96	-6,10	0,00	-1,01	0,99	-1,02	0,31
Ano ingresso	2012 (Ref)								
	2013	0,54	0,38	1,41	0,16	-0,43	0,23	-1,91	0,06
	2014	1,10	0,39	2,83	0,00	-0,59	0,24	-2,44	0,01
	2015	2,10	0,43	4,87	0,00	-0,35	0,24	-1,43	0,15
	2016	1,09	0,37	2,90	0,00	-0,96	0,28	-3,41	0,00
	2017	0,82	0,40	2,08	0,04	-0,78	0,27	-2,95	0,00
Ensino médio	Otro (Ref)								
	Publico	0,68	0,25	2,67	0,01	-0,41	0,19	-2,13	0,03
Sexo	Feminino (Ref)								
	Masculino	-0,20	0,21	-1,00	0,32	0,61	0,16	3,70	0,00
Idade	Menores 20 (Ref)								
	Maiores 20	0,41	0,32	1,30	0,19	0,67	0,18	3,76	0,00
Cursinho	Não (Ref)								
	Sim	-0,43	0,21	-2,11	0,04	-0,13	0,15	-0,82	0,41
Nota matemática		0,007	0,00	4,97	0,00	-0,004	0,00	-3,25	0,00
Nota biologia		0,005	0,00	3,11	0,00	-0,001	0,00	-0,93	0,35

4.3 Comparação dos Modelos

Na seção anterior estimamos três modelos: um modelo sem a variável longitudinal e dois modelos interatuando com o covariável longitudinal, os dois modelos são chamados de modelos conjuntos para risco competitivo. Nesta seção mostramos as diferenças dos trés modelos, segundo a estimativa dos parâmetros, as curvas de sobrevivência e as funções de risco, logo avaliamos o ajuste dos trés modelo usando a estatística deviance e os resíduos Martingale, Cox-Snell e deviance ajustados. Nesta seção chamaremos de Modelo 1 ao modelo de risco competitivo sem a covariável longitudinal, Modelo 2 ao modelo conjunto de risco competitivo em dois estágios e Modelo 3 ao modelo de risco competitivo de parâmetro compartilhado.

Nas estimativas da baseline, como mostra a Figura 10, observamos que as estimativas dos parâmetros para o modelo 1 e o modelo 3 são similares, no entanto, a estimativas do modelo 2 produziu resultados longe do modelo 3, devido à metodologia de estimação. Nas estimativas dos parâmetros das covariáveis, obtemos que para o risco evadido às três abordagens geram estimativas similares, enquanto para o risco formado as estimativas dos parâmetros para o modelo 1 e 3 são similares e as estimativas dos parâmetros do modelo 2 ficam distantes dos dois modelos. Outro resultado que observamos é que o intervalo de confiança para os parâmetros fica menor para os modelos 1 e 3 comparado com o modelo 2.

Figura 10 – Comparação das estimativas de baseline e dos parâmetros dos três modelos ajustados.



Fonte: Produzido pelo autor

Um resultado interessante apresentado na Tabela 11 diz que os dois modelos conjuntos geram estimativas das variâncias e do desvio padrão do efeito aleatório e do erro do modelo semelhantes.

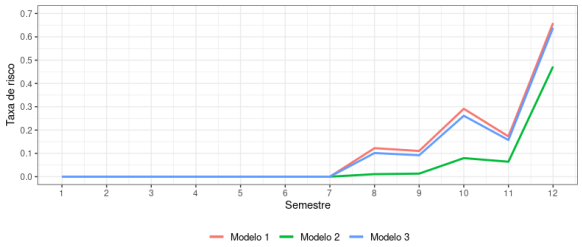
Tabela 11 – Comparação das estimativas das variâncias do efeito aleatório.

	Modelo dois estágios		**Modelo compartilhado**	
Parâmetro	Estimativa	EP	Estimativa	EP
Efeito-Aleatório	0.021	0.0017	0.022	0.0018
Erro-modelo-CR	0.017	0.0004	0.017	0.0004

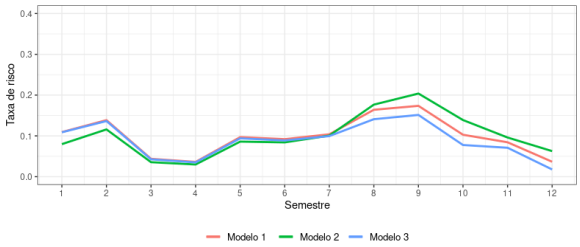
Conforme mostra a Figura 11, as funções de risco e as curvas de sobrevivência são similares para os modelos 1 e 3, como visto na estimação dos parâmetros dos modelos, contudo, os três modelos não sejam semelhantes, eles se apresentam o mesmo padrão para as funções de risco e as curvas de sobrevivência.

Figura 11 – Comparação das taxas de risco e curvas de sobrevivência dos três modelos estimados baseados no perfil médio *(Ano Ingresso:2012, Tipo Ensino Médio: Publica, Sexo: Masculino, Idade: >20,Cursinho: Sim, Nota matemáticas = 473, Nota biologia = 426).*

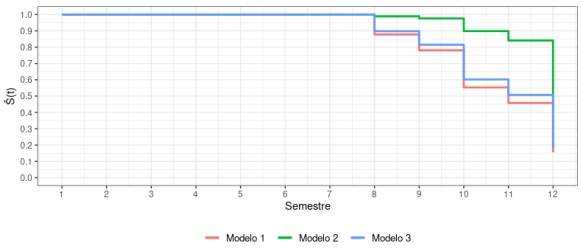
(a) Taxa de Risco Formado



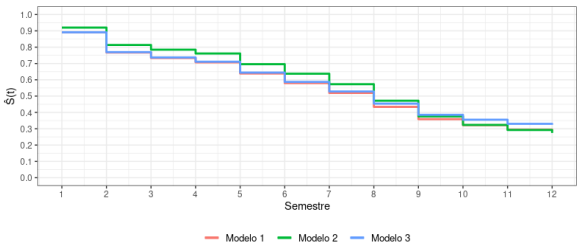
(b) Taxa de Risco Evadido



(c) Curva de sobrevivência Formado



(d) Curva de sobrevivência Evadido



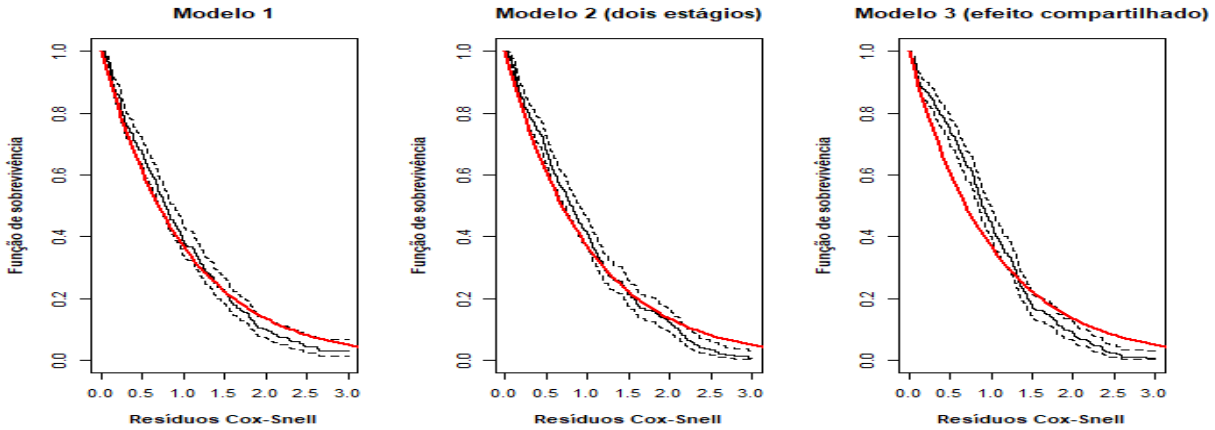
Fonte: Produzido pelo autor

4.3.1 Análise dos resíduos dos três modelos ajustados

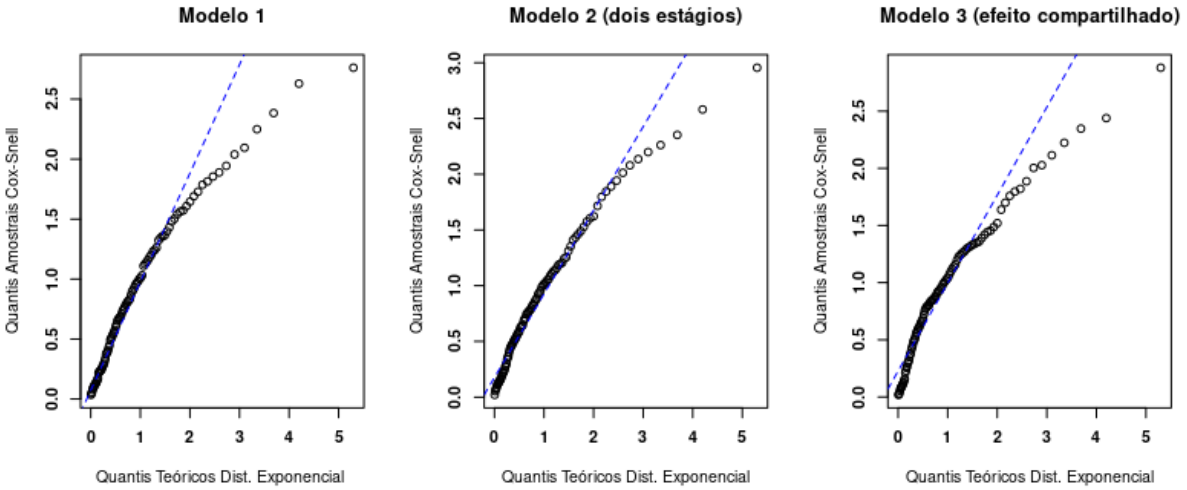
Por último, exibimos a análise gráfica para avaliar o ajuste dos três modelos de risco competitivo com tempo discreto, na Figura 12 exibimos os resíduos Cox-Snell, eles indicam que os modelos 1 e 2 tem um ajuste aceitável, em quanto o modelo 3 apresenta mais discrepância entre a curva de Kaplan-Meier dos resíduos de Cox-Snell e a distribuição exponencial padrão, na Figura 13 mostramos os resíduos deviance padrão e deviance ajustado, nos gráficos observamos que os resíduos deviance padrão apresentam um melhor comportamento que os resíduos deviance ajustados, os resíduos ajustados para os três modelos não apresentam uma distribuição próxima da distribuição normal, mas o gráfico quantil-quantil apresenta um comportamento aceitável.

Figura 12 – Comparação entre os modelos segundo a análise dos resíduos Cox-Snell.

(a) Comparação do estimador Kaplan-Meier dos resíduos Cox-Snell com a distribuição exponencial padrão



(b) Comparação dos quantis dos resíduos Cox-Snell com os quantis da distribuição exponencial padrão



Fonte: Produzido pelo autor

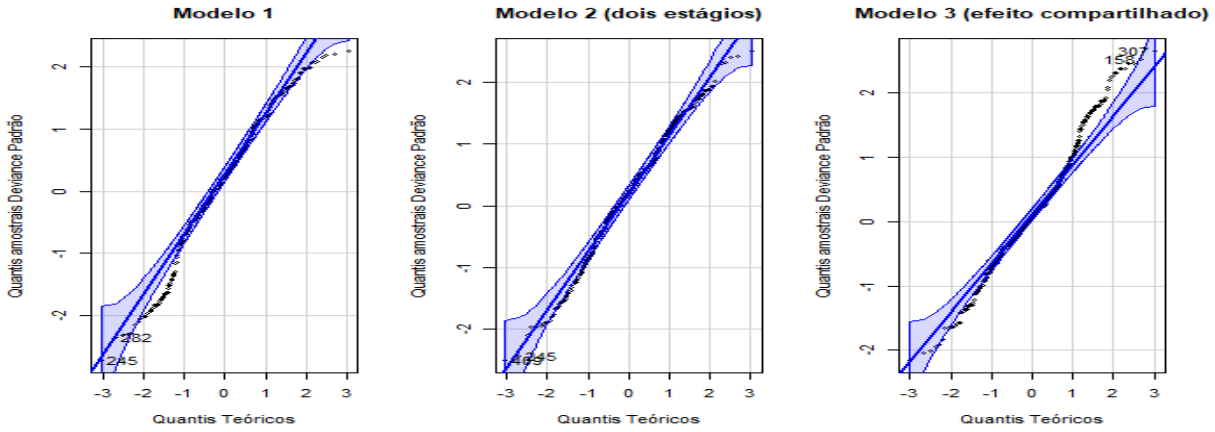
Analisando os gráficos não dispomos de um diagnóstico que permita a seleção do modelo com melhor ajuste, pois os três podem ser aceitáveis, contudo existe outra ferramenta a estatística dos resíduos deviance ao quadrado para tal fim. Segundo a estatística deviance ao quadrado apresentada na Tabela 12 o modelo de parâmetro compartilhado é o modelo que melhor se ajusta aos dados educacionais.

Tabela 12 – Estatística deviance ao quadrado

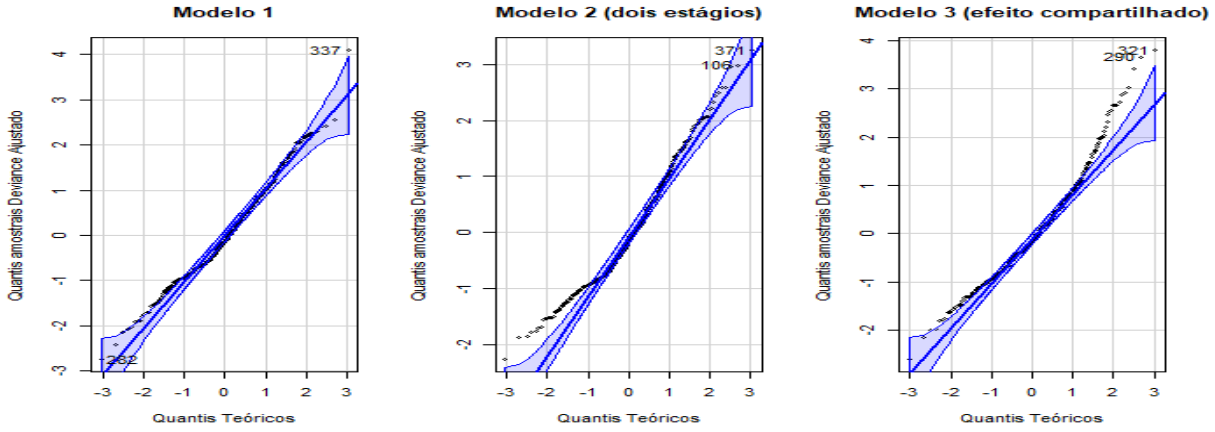
Modelo	Estatística
Risco competitivo (Modelo 1)	14.485
Dois estágios (Modelo 2)	16.199
Parâmetro compartilhado (Modelo 3)	14.149

Figura 13 – Comparação entre os modelos segundo análise dos resíduos deviance.

(a) Deviance Padrão



(b) Deviance Ajustado



Fonte: Produzido pelo autor

4.4 Interpretação dos resultados a partir do modelo conjunto com parâmetro compartilhado

Na Figura 14 são apresentadas algumas comparações de perfis de alunos a partir de um perfil médio dado por alunos ingressantes em 2012, oriundo de escolas públicas do ensino médio, do sexo masculino, com idade superior a 20 anos, que fez cursinho e que obtiveram notas em matemática e biologia iguais a 473 e 426 pontos, respectivamente.

Segundo a Figura 14 observamos que os alunos de sexo feminino, que não fizeram cursinho, alunos que cursaram o ensino médio em escolas públicas apresentam maior chance de se formar em algum semestre quando comparados aos respectivos perfis de referência. Alunos de cursaram integralmente o ensino médio em escolas públicas apresentaram 97% mais chances de se formar em algum semestre em comparação aqueles que não cursaram todo o ensino médio em escolas públicas. As mulheres apresentaram 22% mais chances de se formarem em comparação as homens. A nota em matemática foi outra variável que apresentou efeito positivo, para cada ponto adicionado a nota o aluno aumenta suas chances em se formar em 0,7% (lembrando que as noas são padronizadas para terem média igual a 500 pontos e desvio padrão igual a 100 pontos). Em consequência, quando olhamos para as chances de evadir, Agora na Figura 14(b). observamos que o tipo de ensino outros têm maior chance de evadir o programa que o tipo de ensino comum, pelo contrário, que o aluno tenha menos de 20 anos e seja do sexo feminino diminui a chance de sair do programa pela causa evadir.

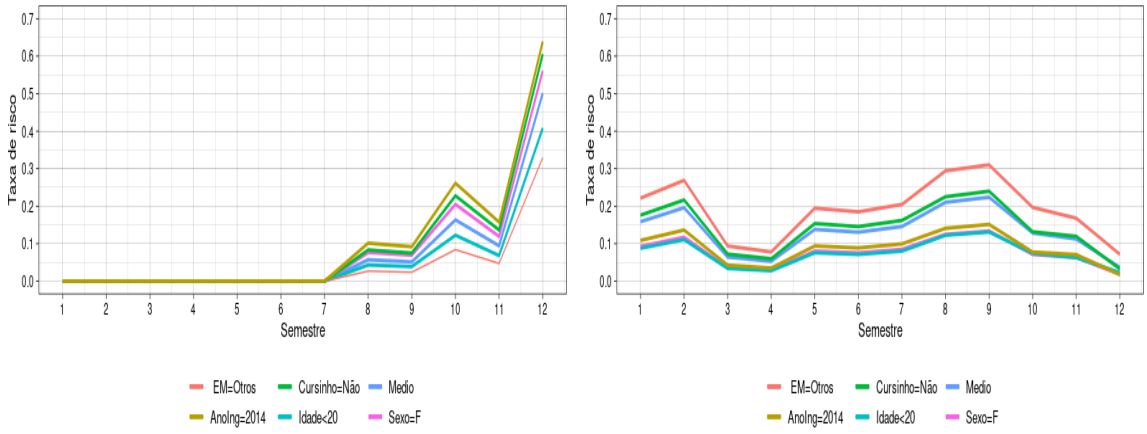
Na sub-Figura 14(d) observamos que probabilidade para que um aluno se forme é maior para os alunos ingressantes no ano 2014, que não fizeram curso pré-vestibular e alunos do sexo femininos, pelo contrário, a probabilidade diminui para os alunos de Idade menor que 20 anos e de ensino médio outros em comparação ao perfil médio. Imediatamente temos também na sub-Figura 14(d) que a probabilidade de que um aluno evada o programa é maior para os alunos de ensino médio outros, que não fizeram curso pré-vestibular em comparação com o perfil médio e os alunos do sexo feminino e de idade menor do que 20 anos tem menor probabilidade de evadir o programa que os alunos de perfil médio.

Por fim, as covariáveis que maior efeito tiveram sob a curva de sobrevivência para o risco de formado são: o ano de ingresso, o tipo de ensino médio, o cursinho e as notas de matemática e biologia, para o risco de evadir, as covariáveis significativas são: o tipo de ensino médio, o Sexo, a idade e a nota de matemáticas.

Figura 14 – Taxa de risco e curva de sobrevivência estimada do modelo de risco competitivo. Nas sub-figuras a) e b) apresentamos a taxa de risco e nas sub-figuras c) e d) apresentamos a curva de sobrevivência para as causas evadido e formado. baseados no perfil médio *(Ano Ingresso: 2012, Tipo Ensino Médio: Publica, Sexo: Masculino, Idade: >20,Cursinho: Sim, Nota matemáticas = 473. Nota biologia = 426).*

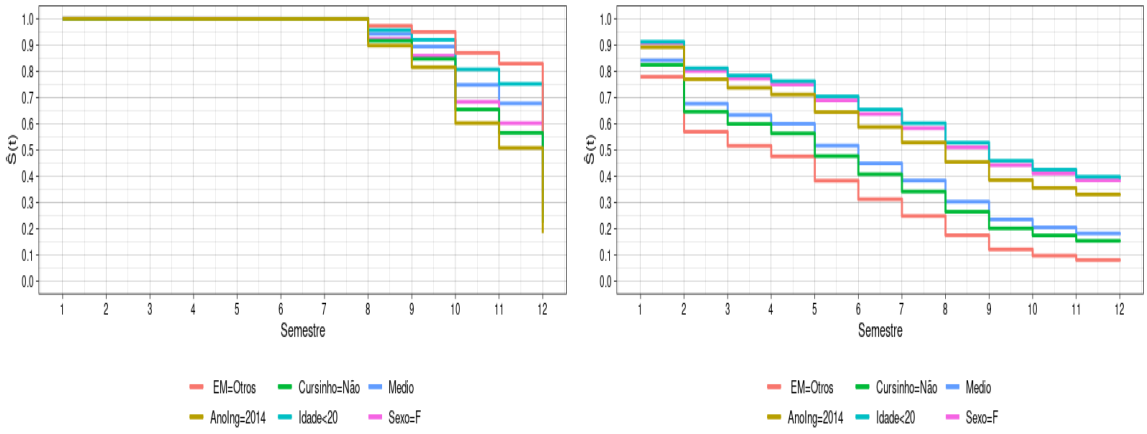
(a) Taxa de risco: Formado

(b) Taxa de risco: Evadido



(c) Curva de sobrevivência: Formado

(d) Curva de sobrevivência: Evadido



Fonte: Produzido pelo autor

5 Considerações Finais

Análogo aos resultados observados na pesquisa de (SACCARO; FRANÇA; JACINTO, 2019) para as áreas das ciências naturais e engenharia, concluímos que a taxa de evasão para o curso de graduação em estatística é elevada Neste estudo se fornecem algumas informações sob as características dos estudantes com mais chance de concluir o curso. Por exemplo, as mulheres e alunos que cursaram todo o ensino médio em escolas públicas.

A partir das análises feitas dos dados de evasão estudantil, concluímos que os três modelos de riscos competitivos para estimar a curva de sobrevivência são aceitáveis. Porém, é importante ressaltar que os modelos conjuntos tem a vantagem de incluir a variável longitudinal, o que permite estimar o efeito desta na curva de sobrevivência. Nos dois modelos conjuntos observamos que a variável longitudinal teve um efeito significativo na modelagem da curva de sobrevivência. Para a escolha entre os dois modelos conjuntos usamos a análise dos resíduos e a estatística dos resíduos deviance ao quadrado.

Na tese de doutorado de (QIU, 2012) se apresentam os modelos conjuntos para a previsão, onde o modelo conjunto de parâmetros compartilhados tem melhor desempenho geral e capacidade de discriminação do que o modelo em dois estágios para os dados de tuberculose. Nesta dissertação apresentamos os modelos conjuntos para inferência em dados educacionais, obtendo que o modelo de parâmetro compartilhado tem melhor ajuste segundo a estatística deviance ao quadrado. Nos modelos conjunto supomos que a variável longitudinal tem distribuição normal, mas nem sempre é verdade, no artigo de (VIVIANI; ALFÓ; RIZOPOULOS, 2014) se propõem os modelos lineais mistos generalizados para resolver o problema de não normalidade, e a estrutura de um modelo conjunto de parâmetro compartilhado para estimar a função de sobrevivência, baseado no método de máxima verossimilhança e apoiados no algoritmo EM para estimar os parâmetros do modelo.

A ligação entre a resposta longitudinal e a resposta de sobrevivência tem sido muito estudada quando a variável tempo de sobrevivência é contínuo, no artigo (PAPAGEORGIOU et al., 2019) se faz uma revisão dos modelos conjuntos e se apresenta uma extensão para a estrutura de ligação, técnicas de estimação e previsão dinâmica que poderia ser explorada quando o tempo de sobrevivência é discreto.

Referências

AALEN, O. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, JSTOR, p. 701–726, 1978. Citado na página 20.

COLOSIMO, E.; GIOLO, S. *Análise de Sobrevivência Aplicada*. [S.l.]: Blucher, 2006. ISBN 9788521215042. Citado 3 vezes nas páginas 17, 18 e 20.

CROWTHER, M. J. *Development and application of methodology for the parametric analysis of complex survival and joint longitudinal-survival data in biomedical research*. Tese (Doutorado) — University of Leicester, 2014. Citado 2 vezes nas páginas 16 e 34.

CZEPIEL, S. A. Maximum likelihood estimation of logistic regression models: theory and implementation. *Available at czep. net/stat/mlelr*, Citeseer, v. 83, 2002. Citado na página 25.

DAC. *Regimento Geral de Graduação*. 2021. Disponível em: <https://www.dac.unicamp.br/portal/graduacao/regimento-geral>. Acesso em: 2021-04-27. Citado na página 41.

DEMIDENKO, E. *Mixed Models: Theory and Applications with R*. [S.l.]: Wiley, 2013. (Wiley Series in Probability and Statistics). ISBN 9781118091579. Citado na página 30.

DIGGLE, P. J.; SOUSA, I.; CHETWYND, A. G. Joint modelling of repeated measurements and time-to-event outcomes: the fourth armitage lecture. *Statistics in Medicine*, Wiley Online Library, v. 27, n. 16, p. 2981–2998, 2008. Citado 2 vezes nas páginas 15 e 33.

HARVILLE, D. A. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American statistical association*, Taylor & Francis, v. 72, n. 358, p. 320–338, 1977. Citado na página 29.

HENDERSON, R.; DIGGLE, P.; DOBSON, A. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, Oxford University Press, v. 1, n. 4, p. 465–480, 2000. Citado 2 vezes nas páginas 15 e 33.

IBRAHIM, J. G.; CHU, H.; CHEN, L. M. Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*, American Society of Clinical Oncology, v. 28, n. 16, p. 2796, 2010. Citado 2 vezes nas páginas 15 e 33.

INEP. *Resumo técnico do Censo da Educação Superior 2019*. [S.l.: s.n.], 2021. ISBN 978-65-5801-023-4. Citado na página 13.

KALBFLEISCH, J.; PRENTICE, R. *The Statistical Analysis of Failure Time Data*. [S.l.]: Wiley, 2011. (Wiley Series in Probability and Statistics). ISBN 9781118031230. Citado na página 20.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, Taylor & Francis, v. 53, n. 282, p. 457–481, 1958. Citado na página 20.

KLEINKE, M. U. O vestibular unicamp e a inclusão social: Experiências e perspectivas. *Workshop de Cursinhos Pré-Vestibulares da UNESP*, 2006. Citado na página 39.

LAIRD, N. M.; WARE, J. H. Random-effects models for longitudinal data. *Biometrics*, JSTOR, p. 963–974, 1982. Citado 2 vezes nas páginas 29 e 30.

LIMA, A. C. L. Modelagem conjunta de dados longitudinais e de sobrevivência. Universidade Federal de Minas Gerais, 2007. Citado 2 vezes nas páginas 15 e 33.

MAIORANO, A. C. Modelagem conjunta de dados longitudinais e de sobrevivência para avaliação de desfechos clínicos do parto. Universidade Federal de São Carlos, 2018. Citado 2 vezes nas páginas 15 e 33.

MONDAL, P. K. *Joint modeling of longitudinal measurements and survival data with competing risks: application to HIV/AIDS study*. Tese (Doutorado) — University of Saskatchewan, 2017. Citado 2 vezes nas páginas 16 e 34.

PAPAGEORGIOU, G.; MAUFF, K.; TOMER, A.; RIZOPOULOS, D. An overview of joint modeling of time-to-event and longitudinal outcomes. *Annual review of statistics and its application*, Annual Reviews, v. 6, p. 223–240, 2019. Citado 3 vezes nas páginas 16, 34 e 57.

QIU, F. *Joint modeling of longitudinal data and discrete-time survival outcome with application to studying tuberculosis immunology data*. [S.l.]: Case Western Reserve University, 2012. Citado 4 vezes nas páginas 15, 34, 36 e 57.

QIU, F.; STEIN, C. M.; ELSTON, R. C.; (TBRU), T. R. U. Joint modeling of longitudinal data and discrete-time survival outcome. *Statistical methods in medical research*, SAGE Publications Sage UK: London, England, v. 25, n. 4, p. 1512–1526, 2016. Citado 2 vezes nas páginas 15 e 34.

RIZOPOULOS, D. *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. [S.l.]: Taylor & Francis, 2012. (Chapman & Hall/CRC Biostatistics Series). ISBN 9781439872864. Citado 2 vezes nas páginas 14 e 36.

SACCARO, A.; FRANÇA, M. T. A.; JACINTO, P. d. A. Fatores associados à evasão no ensino superior brasileiro: um estudo de análise de sobrevivência para os cursos das áreas de ciência, matemática e computação e de engenharia, produção e construção em instituições públicas e privadas. *Estudos Econômicos (São Paulo)*, SciELO Brasil, v. 49, p. 337–373, 2019. Citado 2 vezes nas páginas 13 e 57.

SALAZAR, J. C.; CORREA, J. C. *Introducción a los modelos mixtos*. [S.l.]: Universidad Nacional de Colombia, 2016. ISBN 9789587759532. Citado na página 30.

SCHMID, M.; BERGER, M. Competing risks analysis for discrete time-to-event data. *Wiley Interdisciplinary Reviews: Computational Statistics*, Wiley Online Library, v. 13, n. 5, p. e1529, 2021. Citado 3 vezes nas páginas 14, 25 e 34.

SEMESP. *Mapa do ensino superior*. 2021. Disponível em: <https://www.semesp.org.br/mapa-do-ensino-superior/edicao-11/dados-estados-e-regioes/sudeste/sao-paulo/>. Acesso em: 2021-04-27. Citado 2 vezes nas páginas 13 e 37.

SINGER, J. M.; NOBRE, J. S.; ROCHA, F. M. *Análise de dados longitudinais Versão parcial preliminar*. [S.l.: s.n.], 2018. (Departamento de Estatística Universidade de São Paulo). Citado 2 vezes nas páginas 28 e 31.

TUTZ, G.; SCHMID, M. *Modeling Discrete Time-to-Event Data*. [S.l.]: Springer International Publishing, 2016. (Springer Series in Statistics). ISBN 9783319281582. Citado 8 vezes nas páginas 14, 21, 23, 24, 25, 27, 34 e 36.

VIVIANI, S.; ALFÓ, M.; RIZOPOULOS, D. Generalized linear mixed joint model for longitudinal and survival outcomes. *Statistics and Computing*, Springer, v. 24, n. 3, p. 417–427, 2014. Citado na página 57.

VONESH, E. F.; GREENE, T.; SCHLUCHTER, M. D. Shared parameter models for the joint analysis of longitudinal data and event times. *Statistics in medicine*, Wiley Online Library, v. 25, n. 1, p. 143–163, 2006. Citado 2 vezes nas páginas 15 e 33.

WEN, C.-C.; CHEN, Y.-H. Discrete-time survival data with longitudinal covariates. *Statistics in Medicine*, Wiley Online Library, v. 39, n. 29, p. 4372–4385, 2020. Citado na página 34.

WULFSOHN, M. S.; TSIATIS, A. A. A joint model for survival and longitudinal data measured with error. *Biometrics*, JSTOR, p. 330–339, 1997. Citado 2 vezes nas páginas 15 e 33.

A Anexo: Código do R

```
options(scipen = 100)
# bibliotecas
require(Rdmu) #Modelo conjunto bivariado
require(car)  #
require(survival) #Fun  es de sobreviv ncia
require(VGAM)   #Modelo multinomial
require(lme4)
library(nlme)
library(lmerTest)
require(hnp)
require(ggplot2) #Gr ficos
require(Hmisc)

inv.logit <- function(x) exp(x)/(1+exp(x))

#lendo os dados
setwd("C:Projeto Mestrado/ModeloMisto")
dados <- read.csv("Dados_longo_novo.csv")


dados$Cen3Ativo <- ifelse(dados$SITUACAO2=="Ativo",1,0)

dados <- subset(dados, !is.na(Cursinho))
dados <- subset(dados, !is.na(EM))

#Alunos formados em menos de 6 semestres
alum <- c(2915,2998,5960,6213,6462,6684,6874,
7846,7913,8423,10094,10120,13849,16581,19085,22334)
#dados <- dados[!(dados$SITUACAO2=="Formado" & dados$baseline %in
   % c(1,2,3,4,5,6) ) ,]
dados <- dados[!(dados$id %in% alum ) ,]

dados <- dados[!(dados$ANOING %in% c(2018) ) ,]

table(dados$ANOING,dados$SITUACAO2)

#imputar dados longitudinais
dados <- tidyr::fill(dados, CR_POR_PERIODO)
```

```
38
39 #write.csv(dados, "dados_imputar.csv")
40
41 #dados <- subset(dados, !is.na(CR_POR_PERIODO))
42
43 dados$ANOING <- as.character(dados$ANOING)
44 dados$baseline <- factor(dados$baseline,levels = c(1:12))
45
46 table(dados$baseline)
47 ##########################################################
48 ## modelo sem efeito aleat rio
49 ##########################################################
50
51 dados$y <- dados$Cen1Formado + dados$Cen2Evadido
52 dados$Y <- dados$y
53 dados$Y <- with(dados, ave(Y, id, FUN = sum))
54
55 Y <- dados$Y[dados$baseline==1]
56 table(Y)
57
58 Modelo1 <- vglm(cbind(Cen1Formado, Cen2Evadido, Cen3Ativo) ~
59                   baseline +ANOING+EM +Sexo+ Idade + Cursinho+mat
                      + bio-1,
60                   family = multinomial(refLevel = "Cen3Ativo"),
61                  data = dados)
62 summary(Modelo1)
63
64 hazard_m1 <- predictvglm(Modelo1,dados,type="response")
65 dados$lambda_1 <- hazard_m1[,1]+hazard_m1[,2]
66 dados$CS_1 <- with(dados, ave(lambda_1, id, FUN = sum))
67 CS_1 <- dados$CS_1[dados$baseline==1]
68 km.cs1 <- survfit(Surv(CS_1, Y)~1)
69 #Martingale
70 M_1 <- Y - CS_1
71 #Deviance (padr o)
72 D_1 <- sign(M_1)*((-2)*(M_1+Y*log(Y-M_1)))^(0.5)
73
74 ## Res duos deviance ajustados (Tutz and Schmid, 2015 p gina
   80)
75 d1 <- with(dados, sign(y - lambda_1) * sqrt(
76                         ifelse(y == 1, log(1/lambda_1),
77                                log(1/(1-lambda_1)))))
```

```
78 d2 <- with(dados, (1 - 2*lambda_1)/sqrt(lambda_1*(1-lambda_1)*36)
   )
79 d1 <- ave(d1, dados$id, FUN = sum)
80 d2 <- ave(d2, dados$id, FUN = sum)
81 dados$dev_1 <- d1 + d2
82
83 dev_1 <- dados$dev_1[dados$baseline==1]
84
85 #Resumo
86 hist(dev_1)
87 summary(dev_1)
88 boxplot(dev_1)
89 qqPlot(dev_1)
90
91 ######################################################
92 ## modelo em dois estagios
93 ######################################################
94 Modelo_CR <- lmer(CR_POR_PERIODO ~ baseline +ANOING+EM +Sexo+
   Idade + Cursinho+mat + bio + fis + (1 | id), data=dados)
95 summary(Modelo_CR)
96
97 vcov(Modelo_CR)
98
99 qqPlot(residuals(Modelo_CR))
100 # Crear una sucesion de numeros entre -0.5y 0.5, aumentando en
   0.01.
101 x <- seq(-0.5, 0.5, by = 0.01)
102
103 # Suponiendo que los par metros son: mu=2 y sigma=1.1.
104
105 hist(residuals(Modelo_CR),freq = F, ylim = c(0,3.5))
106 lines(density(residuals(Modelo_CR)))
107 curve(dnorm(x, 0.00, 0.12), xlim = c(-0.5,0.5), add = T)
108
109 efeito_cr <- data.frame(ranef(Modelo_CR))
110
111 efeito_cr <- data.frame(id=efeito_cr$grp,
112                         b_cr=efeito_cr$condval)
113
114 dados <- merge(dados, efeito_cr, by = "id", all.x = T)
115
116 summary(dados$b_cr)
```

```
117
118
119 Modelo_2 <- vglm(cbind(Cen1Formado, Cen2Evadido, Cen3Ativo) ~
120                     baseline + ANOING+EM + Sexo + Idade + Cursinho
                        +mat + bio+b_cr - 1, family = multinomial(
                        refLevel = "Cen3Ativo"),
121                   data = dados)
122 summary(Modelo_2)
123
124 hazard_m2 <- predictvglm(Modelo_2,dados,type="response")
125
126 dados$lambda_2 <-hazard_m2[,1]+hazard_m2[,2]
127
128 dados$CS_2 <- with(dados, ave(lambda_2, id, FUN = sum))
129 CS_2 <- dados$CS_2[dados$baseline==1]
130 km.cs2 <- survfit(Surv(CS_2, Y)~1)
131
132 # Martingale
133 M_2 <- Y - CS_2
134 # Deviance (padr o)
135 D_2 <- sign(M_2)*((-2)*(M_2+Y*log(Y-M_2)))^(0.5)
136
137 ## adjusted deviance residual (Tutz and Schmid, 2015  pagina 80)
138 d1 <- with(dados,
139            sign(y - lambda_2) * sqrt(
140            ifelse(y == 1, log(1/lambda_2),
141            log(1/(1-lambda_2)))))
142 d2 <- with(dados,
143            (1 - 2*lambda_2)/sqrt(lambda_2*(1-lambda_2)*36))
144 d1 <- ave(d1, dados$id, FUN = sum)
145 d2 <- ave(d2, dados$id, FUN = sum)
146 dados$dev_2 <- d1 + d2
147 dev_2 <- dados$dev_2[dados$baseline==1]
148
149 qqPlot(dev_2)
150
151 ############################################################
152 # ajustando o modelo conjunto com efeito compartilhado
153 ############################################################
154 dados$id <- as.character(dados$id)
155 dados.dmu <- dados
156 table(dados.dmu$ANOING)
```

```
157
158 dados.dmu$ANOING <- factor(dados.dmu$ANOING, levels = c
       (2013:2017,2012))
159 dados.dmu$ANOING <- recode(dados.dmu$ANOING, "c(2012) = 3012")
160 dados.dmu$Sexo <- factor(dados.dmu$Sexo, levels = c("M","F"))
161 dados.dmu$Sexo <- recode(dados.dmu$Sexo, "c('M') = 20; c('F') =
       21 ")
162 dados.dmu$Idade <- factor(dados.dmu$Idade, levels = c(">20","<21"
       ))
163 dados.dmu$Idade <- recode(dados.dmu$Idade,  "c('>20') = 30; c
       ('<21') = 31 ")
164 dados.dmu$EM <- factor(dados.dmu$EM, levels = c("Publica","Outros
       "))
165 dados.dmu$EM <- recode(dados.dmu$EM, "c('Publica') = 40; c('
       Outros') = 41")
166 dados.dmu$Cursinho <- factor(dados.dmu$Cursinho, levels = c("Sim"
       ,"Nao"))
167 dados.dmu$Cursinho <- recode(dados.dmu$Cursinho, "c('Sim') = 50;
       c('Nao') = 51")
168
169 #Modelo efeito bivariado
170 mod_comp <- Rdmuai(c(
171           Cen1Formado ~ -1 +baseline +ANOING+Sexo+ Idade+EM+
                 Cursinho+mat + bio + (1|id),
172           Cen2Evadido ~ -1 +baseline +ANOING++Sexo+ Idade+EM+
                 Cursinho+mat + bio + (1|id),
173           CR_POR_PERIODO ~ -1 +baseline +ANOING+EM +Sexo+ Idade +
                  Cursinho+mat + bio + (1|id)
174         ),
175 data = dados.dmu,
176 dir.file = "mod_comp.DIR",
177 NOCOV = "all",
178 VARREST = "VAR 2 E 1 2",
179 start = list(diag(c(1,1,1)), matrix(c(1,NA,NA,NA,1,NA,NA,NA,1)
       ,3,3)),
180 control = list(SOL = T, CONV.NDELTA = 1e-04,
181                CONV.NGRAD = 1e-04, residuals = T),
182 family = binomial(link = "logit"),
183 execute = F)
184 write.table(c("$REDUCE", "1 1", "1 2", "1 3"), file = "mod_comp.
       DIR", append = T,
185             row.names = F, col.names = F, quote = F)
```

```
186
187 #antes de executar o modelo e preciso abrir o arquivio "mod_comp.
       DIR" e
188 #deletar a linha $GLMM 3 VARF=BINOMIAL LINK=LOGIT
189 #salvar e fechar o arquivo
190
191 mod_comp <- execute.dmuai(mod_comp)
192
193 ## Estimativas efeitos fixos (betas)
194 efeitos.fixos_con <- as.data.frame(mod_comp$SOL[mod_comp$SOL[,1]
       == 1,-c(1,3,6,7)])
195 efeitos.fixos_con$V4 <- efeitos.fixos_con$V4 + 6
196
197 efeitos.fixos_cat <- as.data.frame(mod_comp$SOL[mod_comp$SOL[,1]
       == 2,-c(1,3,6,7)])
198
199 efeitos.fixos <- rbind(efeitos.fixos_cat, efeitos.fixos_con)
200 table(dados.dmu$Idade)
201
202 efeitos.fixos <- efeitos.fixos[order(efeitos.fixos$V2, efeitos.
       fixos$V4,
203                                      efeitos.fixos$V5, decreasing
                                           = F),]
204
205 names(efeitos.fixos) <- c("Resposta","Variavel","Nivel","
       Estimativa","EP")
206
207 efeitos.fixos$Resposta <- ifelse(efeitos.fixos$Resposta==1,"
       Conclusao",                          ifelse(efeitos.
       fixos$Resposta==2,"Evasao","CR"))
208
209 efeitos.fixos$Variavel[efeitos.fixos$Variavel==1] <- "Baseline"
210 efeitos.fixos$Variavel[efeitos.fixos$Variavel==2] <- "ANOING"
211 efeitos.fixos$Variavel[efeitos.fixos$Variavel==3] <- "Sexo"
212 efeitos.fixos$Variavel[efeitos.fixos$Variavel==4] <- "Idade"
213 efeitos.fixos$Variavel[efeitos.fixos$Variavel==5] <- "EM"
214 efeitos.fixos$Variavel[efeitos.fixos$Variavel==6] <- "Curisinho"
215 efeitos.fixos$Variavel[efeitos.fixos$Variavel==7] <- "mat"
216 efeitos.fixos$Variavel[efeitos.fixos$Variavel==8] <- "bio"
217
218 efeitos.fixos$Nivel[efeitos.fixos$Nivel==20]  <- "Masculino"
219 efeitos.fixos$Nivel[efeitos.fixos$Nivel==21]  <- "Feminino"
```

```
220
221 efeitos.fixos$Nivel[efeitos.fixos$Variavel=="Idade" & efeitos.
      fixos$Nivel==30]  <-  ">20"
222 efeitos.fixos$Nivel[efeitos.fixos$Variavel=="Idade" & efeitos.
      fixos$Nivel==31]  <- "<=21"
223
224 efeitos.fixos$Nivel[efeitos.fixos$Nivel==40]  <- "Publica"
225 efeitos.fixos$Nivel[efeitos.fixos$Nivel==41]  <- "Outros"
226
227 efeitos.fixos$Nivel[efeitos.fixos$Nivel==50]  <- "Sim"
228 efeitos.fixos$Nivel[efeitos.fixos$Nivel==51]  <- "Nao"
229
230 efeitos.fixos$t <- efeitos.fixos$Estimativa/efeitos.fixos$EP
231 efeitos.fixos$p.value <- 2*(1-pnorm(abs(efeitos.fixos$t)))
232
233 efeitos.fixos$Estimativa <- round(efeitos.fixos$Estimativa,4)
234 efeitos.fixos$EP <- round(efeitos.fixos$EP,4)
235 efeitos.fixos$t <- round(efeitos.fixos$t,4)
236 efeitos.fixos$p.value <- round(efeitos.fixos$p.value,4)
237 efeitos.fixos
238
239 ## Estimativa vari ncia do efeitos aleat rio
240 variancias <- as.data.frame(mod_comp$PAROUT.STD)
241 variancias <- variancias[c(1,4),]
242 names(variancias) <- c("Parametro", "Estimativa", "EP")
243 variancias$Parametro <- ifelse(variancias$Parametro==1,
244                                "Efeito-aleat rio",
245                                "Erro - modelo CR")
246
247 ## Valores preditos dos efeitos aleat rios
248 efeito <- as.data.frame(mod_comp$SOL[mod_comp$SOL[,1] == 3 &
      mod_comp$SOL[,4] == 1,-c(1:4,6:7)])
249 dim(efeito)
250 summary(efeito)
251 head(efeito)
252 plot(efeito$V8)
253 qqPlot(efeito$V8)
254 hist(efeito$V8)
255 shapiro.test(efeito$V8)
256 names(efeito) <- c("id","b","b.se")
257 dados <- merge(dados, efeito, by = "id", all.x= T)
258
```

```
## Analise de residuos

##obtendo a estimativa do preditor linear para cada linha

eta_f <- mod_comp$fitted[,1] + dados$b
eta_e <- mod_comp$fitted[,2] + dados$b

lambda_f <- inv.logit(eta_f)
lambda_e <- inv.logit(eta_e)

dados$lambda_3 <- lambda_f + lambda_e

dados$CS_3 <- ave(dados$lambda_3,
                          dados$id,
                          FUN = sum)

CS_3 <- dados$CS_3[dados$baseline==1]
km.cs3 <- survfit(Surv(CS_3, Y)~1)

# Martingale
M_3 <- Y - CS_3

# Deviance (padr o)
D_3 <- sign(M_3)*((-2)*(M_3+Y*log(Y-M_3)))^(0.5)

## adjusted deviance residual (Tutz and Schmid, 2015  pagina 80)
d1 <- with(dados,
           sign(y - lambda_3) * sqrt(
           ifelse(y == 1, log(1/lambda_3),
           log(1/(1-lambda_3)))))
d2 <- with(dados,
           (1 - 2*lambda_3)/sqrt(lambda_3*(1-lambda_3)*36))
d1 <- ave(d1, dados$id, FUN = sum)
d2 <- ave(d2, dados$id, FUN = sum)
dados$dev_3 <- d1 + d2

dev_3 <- dados$dev_3[dados$baseline==1]

## graficos

##Cox-Snell
par(mfrow = c(1,3))
```

```
plot(km.cs1,
     xlab = "Res duos Cox-Snell",
     ylab = "Fun  o de sobrevivencia",
     main = "Modelo 1",
     font.lab = 2,
     xlim = c(0,3))
curve(exp(-x), 0, 6, col = "red", lwd = 2, add = T)
plot(km.cs2,
     xlab = "Res duos Cox-Snell",
     ylab = "Fun  o de sobrevivencia",
     main = "Modelo 2 (dois estagios)",
     font.lab = 2,
     xlim = c(0,3))
curve(exp(-x), 0, 6, col = "red", lwd = 2, add = T)
plot(km.cs3,
     xlab = "Res duos Cox-Snell",
     ylab = "Fun  o de sobrevivencia",
     main = "Modelo 3 (efeito compartilhado)",
     font.lab = 2,
     xlim = c(0,3))
curve(exp(-x), 0, 6, col = "red", lwd = 2, add = T)

#Cox-Snell vs exponencial padr o
p <- ppoints(100)    # 100 equally spaced points on (0,1),
   excluding endpoints
q_1 <- quantile(CS_1,p=p) # percentiles of the sample
   distribution
plot(qexp(p) ,q_1, main="Modelo 1",
     xlab="Quantis Teoricos Dist. Exponencial ",
     ylab="Quantis Amostrais Cox-Snell", font = 2)
qqline(q_1, distribution=qexp,col="blue", lty=2)

q_2 <- quantile(CS_2,p=p) # percentiles of the sample
   distribution
plot(qexp(p) ,q_2, main="Modelo 2 (dois estagios)",
     xlab="Quantis Teoricos Dist. Exponencial",
     ylab="Quantis Amostrais Cox-Snell", font = 2)
qqline(q_2, distribution=qexp,col="blue", lty=2)

q_3 <- quantile(CS_3,p=p) # percentiles of the sample
   distribution
plot(qexp(p) ,q_3, main="Modelo 3 (efeito compartilhado)",
```

```
339      xlab="Quantis Teoricos Dist. Exponencial",
340      ylab="Quantis Amostrais Cox-Snell", font = 2)
341 qqline(q_3, distribution=qexp,col="blue", lty=2)
342 
343 # Deviance (padr o)
344 qqPlot(D_1,
345        xlab = "Quantis Teoricos",
346        ylab = "Quantis amostrais Deviance padr o",
347        main = "Modelo 1")
348 qqPlot(D_2,
349        xlab = "Quantis Teoricos",
350        ylab = "Quantis amostrais Deviance padr o",
351        main = "Modelo 2 (dois estagios)")
352 qqPlot(D_3,
353        xlab = "Quantis Teoricos",
354        ylab = "Quantis amostrais Deviance padr o",
355        main = "Modelo 3 (efeito compartilhado)")
356 
357 # Deviance (ajustado)
358 qqPlot(scale(dev_1),
359        xlab = "Quantis Teoricos",
360        ylab = "Quantis amostrais Deviance Ajustado",
361        main = "Modelo 1")
362 qqPlot(scale(dev_2),
363        xlab = "Quantis Teoricos",
364        ylab = "Quantis amostrais Deviance Ajustado",
365        main = "Modelo 2 (dois estagios)")
366 qqPlot(scale(dev_3),
367        xlab = "Quantis Teoricos",
368        ylab = "Quantis amostrais Deviance Ajustado",
369        main = "Modelo 3 (efeito compartilhado)")
370 
371 # Deviance (padr o) - HNP
372 hnp(D_1, resid.type = "deviance",
373     paint.out = T, print.on= T,
374     halfnormal = F,
375     xlab = "Quantis Teoricos",
376     ylab = "Quantis amostrais Deviance padr o -HNP",
377     main = "Modelo 1")
378 hnp(D_2, resid.type = "deviance",
379     paint.out = T, print.on= T,
380     halfnormal = F,
```

```
381     xlab = "Quantis Teoricos",
382     ylab = "Quantis amostrais Deviance padr o-HNP",
383     main = "Modelo 2 (dois estagios)")
384 hnp(D_3, resid.type = "deviance",
385     paint.out = T, print.on= T,
386     halfnormal = F,
387     xlab = "Quantis Teoricos",
388     ylab = "Quantis amostrais Deviance padr o-HNP",
389     main = "Modelo 3 (efeito compartilhado)")
390
391 # Deviance (ajustado) - HNP
392 hnp(scale(dev_1), resid.type = "deviance",
393     paint.out = T, print.on= T,
394     halfnormal = F,
395     xlab = "Quantis Teoricos",
396     ylab = "Quantis amostrais Deviance Ajustado-HNP",
397     main = "Modelo 1")
398 hnp(scale(dev_2), resid.type = "deviance",
399     paint.out = T, print.on= T,
400     halfnormal = F,
401     xlab = "Quantis Teoricos",
402     ylab = "Quantis amostrais Deviance Ajustado-HNP",
403     main = "Modelo 2 (dois estagios)")
404 hnp(scale(dev_3), resid.type = "deviance",
405     paint.out = T, print.on= T,
406     halfnormal = F,
407     xlab = "Quantis Teoricos",
408     ylab = "Quantis amostrais Deviance Ajustado-HNP",
409     main = "Modelo 3 (efeito compartilhado)")
410
411 #Carregar imagem dados ----
412 setwd("Projeto Mestrado/ModeloMisto")
413 load(file="resultados_3Modelos.RData")
414
415 #Residuos de desvio quadrado----
416
417 splitY <- split(dados$Y,dados$id)
418 splitlambda_1 <- split(dados$lambda_1,dados$id)
419 splitlambda_2 <- split(dados$lambda_2,dados$id)
420 splitlambda_3 <- split(dados$lambda_3,dados$id)
421
422 SquDevResid_m1 <- function (x) {-2*sum(splitY [[x]] * log(
```

```
   splitlambda_1 [[x]]) + (1 - splitY [[x]]) * log(1 -
   splitlambda_1 [[x]] ))}
SqResiduals_m1 <- sapply(1:length(splitY), SquDevResid_m1)
sum(SqResiduals_m1)
SquDevResid_m2 <- function (x) {-2*sum(splitY [[x]] * log(
   splitlambda_2 [[x]]) + (1 - splitY [[x]]) * log(1 -
   splitlambda_2 [[x]] ))}
SqResiduals_m2 <- sapply(1:length(splitY), SquDevResid_m2)
sum(SqResiduals_m2)
SquDevResid_m3 <- function (x) {-2*sum(splitY [[x]] * log(
   splitlambda_3 [[x]]) + (1 - splitY [[x]]) * log(1 -
   splitlambda_3 [[x]] ))}
SqResiduals_m3 <- sapply(1:length(splitY), SquDevResid_m3)
sum(SqResiduals_m3)

#Grafico comparacoes tres modelos ----

#compara  o baseline ----

dtfest<-data.frame("Modelo1F"=coef(Modelo1)[seq(13,24,by=2)],"
   Modelo1E" = coef(Modelo1)[seq(14,24, by = 2)],"Modelo2F" = coef
   (Modelo_2)[seq(13,24, by = 2)],"Modelo2E"=coef(Modelo_2)[seq
   (14,24, by = 2)],"Modelo3F" = efeitos.fixos$Estimativa[7:12],"
   Modelo3E" = efeitos.fixos$Estimativa[35:40], "Semestre" = c
   (7:12))
ggplot(dtfest, aes(Semestre)) +
  geom_line(aes(y = Modelo1F, colour = "Modelo 1"), size = 1.2) +
  geom_line(aes(y = Modelo2F, colour = "Modelo 2"), size = 1.2) +
  geom_line(aes(y = Modelo3F, colour = "Modelo 3"), size = 1.2)+
  scale_x_continuous("Semestre", labels = as.character(
     dtfest$Semestre),     breaks =dtfest$Semestre)+
  ylab(expression(gamma[0][t][r]))+
  scale_color_discrete(name = "", )+theme_bw()+theme(legend.
     position = "bottom",text = element_text(size=30))

ggplot(dtfest, aes(Semestre)) +
  geom_line(aes(y = Modelo1E, colour = "Modelo 1"), size = 1.2)+
   geom_line(aes(y = Modelo2E, colour = "Modelo 2"),size = 1.2)+
  geom_line(aes(y = Modelo3E, colour = "Modelo 3"),size = 1.2)+
  scale_x_continuous("Semestre", labels = as.character(
     dtfest$Semestre),     breaks =dtfest$Semestre)+
  ylab(expression(gamma[0][t][r]))+
```

```
451   scale_color_discrete(name = "", )+theme_bw()+theme(legend.
        position = "bottom",text = element_text(size=30))
452
453 #compara  o covari veis ----
454
455 sd1 <- coef(summary(Modelo1))[, "Std. Error"]
456 sd2 <- coef(summary(Modelo_2))[, "Std. Error"]
457 efeitos.fixos$EP[c(13:17,23,19,21,24,26,27)]
458
459 dfModelo1 <- data.frame("coefF"=coef(Modelo1)[seq(25,46,by=2)],"
      sdF"=sd1[seq(25,46,by=2)], "coefE" = coef(Modelo1)[seq(26,46,
      by = 2)],"sdE"=sd1[seq(26,46, by = 2)], "Modelo" = rep("Modelo1
      ",11), "ponto" = c(1:11))
460
461 dfModelo2 <- data.frame("coefF" = coef(Modelo_2)[seq(25,46, by =
      2)],"sdF"=sd2[seq(25,46,by=2)], "coefE"=coef(Modelo_2)[seq
      (26,46, by = 2)], "sdE"=sd2[seq(26,46, by = 2)], "Modelo" = rep
      ("Modelo2",11), "ponto" = c(1:11))
462
463 dfModelo3 <- data.frame("coefF" = efeitos.fixos$Estimativa[c
      (13:17,23,19,21,24,26,27)], "sdF" = efeitos.fixos$EP[c
      (13:17,23,19,21,24,26,27)], "coefE" = efeitos.fixos$Estimativa[
      c(41:45,51,47,49,52,54,55)], "sdE" = efeitos.fixos$EP[c
      (41:45,51,47,49,52,54,55)], "Modelo" = rep("Modelo3",11), "
      ponto" = c(1:11) )
464
465 dfTotal <- rbind(dfModelo1, dfModelo2, dfModelo3)
466
467 #par metros formado
468 qplot(ponto,coefF, data = dfTotal, geom = 'point')+
469   geom_errorbar(aes(x=ponto, ymin=coefF-sdF, ymax=coefF+sdF,
        color = Modelo), width=0.3, size = 1, position =
        position_dodge(0.2))+  scale_x_continuous(c(), breaks=c(1:11)
        , labels = c("Ing2013","Ing2014","Ing2015","Ing2016","Ing2017
        ","EMPub","SexM","Idad>20","CurSim","mat","bio")) +
470   ylab("Estimacoes par metros") + theme_bw()+theme(legend.
        position = "bottom", text = element_text(size=25))
471
472 #par metros evadido
473 qplot(ponto,coefE, data = dfTotal)+
474   geom_errorbar(aes(x=ponto, ymin=coefE-sdE, ymax=coefE+sdE,
        color = Modelo), width=0.3, size=1, position = position_dodge
```

```
      (0.2))+  scale_x_continuous(c(), breaks=c(1:11), labels = c("
      Ing2013","Ing2014","Ing2015","Ing2016","Ing2017","EMPub","
      SexM","Idad>20","CurSim","mat","bio"))+
475   ylab("Estimacoes par metros") +theme_bw()+theme(legend.
        position = "bottom", text = element_text(size=25))
476
477 # Grafico taxa de risco ----
478
479 PerfilRiscos_M1<-function(perfil,Modelo){
480   #Funcao de risco estimada para cada risco
481   #Termos para o risco 1
482   #Baseline risco 1 Formado
483   Exr0t1<-exp(coef(Modelo)[seq(1,24,by=2)])
484   #covari veis risco 1 formado
485   coefModelR1<-c(coef(Modelo)[seq(25,46,by=2)])
486   #baseline Risco 2 Evadido
487   Exr0t2<-exp(coef(Modelo)[seq(2,24,by=2)])
488   #covariavesi risco 2 Evadido
489   coefModelR2<-c(coef(Modelo)[seq(26,46,by=2)])
490   #produto dos coeficientes pelo perfil
491   ExCvt1<-exp(perfil%*%coefModelR1)
492   #Termos para o risco 2
493   #produto das covari veis pelo perfil
494   ExCvt2<-exp(perfil%*%coefModelR2)
495   #Fucoes de risco
496   #Risco Formado
497   lambda_risFC<-Exr0t1*ExCvt1/(1+Exr0t1*ExCvt1+Exr0t2*ExCvt2)
498   #Risco Evadido
499   lambda_risEC<-Exr0t2*ExCvt2/(1+Exr0t1*ExCvt1+Exr0t2*ExCvt2)
500   #Sob Evadido
501   sobr_risEC<-cumprod(1-lambda_risEC[1:12])
502   #Sob Formado
503   sobr_risFC<-cumprod(1-lambda_risFC[1:12])
504
505   salida<-list(lambda_risEC,lambda_risFC,sobr_risEC,sobr_risFC)
506   return(salida)
507 }
508
509 PerfilRiscos_M2<-function(perfil,Modelo){
510   #Funcao de risco estimada para cada risco
511   #Termos para o risco 1
512   #Baseline risco 1 Formado
```

```
513   Exr0t1<-exp(coef(Modelo)[seq(1,24,by=2)])
514   #covari veis risco 1 formado
515   coefModelR1<-c(coef(Modelo)[seq(25,46,by=2)])
516   #baseline Risco 2 Evadido
517   Exr0t2<-exp(coef(Modelo)[seq(2,24,by=2)])
518   #covariavesi risco 2 Evadido
519   coefModelR2<-c(coef(Modelo)[seq(26,46,by=2)])
520   #produto dos coeficientes pelo perfil
521   ExCvt1<-exp(perfil%*%coefModelR1)
522   #Termos para o risco 2
523   #produto das covari veis pelo perfil
524   ExCvt2<-exp(perfil%*%coefModelR2)
525   #Fucoes de risco
526   #Risco Formado
527   lambda_risFC<-Exr0t1*ExCvt1/(1+Exr0t1*ExCvt1+Exr0t2*ExCvt2)
528   #Risco Evadido
529   lambda_risEC<-Exr0t2*ExCvt2/(1+Exr0t1*ExCvt1+Exr0t2*ExCvt2)
530   #Sob Evadido
531   sobr_risEC<-cumprod(1-lambda_risEC[1:12])
532   #Sob Formado
533   sobr_risFC<-cumprod(1-lambda_risFC[1:12])
534
535   salida<-list(lambda_risEC,lambda_risFC,sobr_risEC,sobr_risFC)
536   return(salida)
537 }
538
539 PerfilRiscosM_3<-function(perfil){
540   #Funcao de risco estimada para cada risco
541   #Termos para o risco 1
542   #Baseline risco 1 Formado
543   Exr0t1<-exp(efeitos.fixos$Estimativa[1:12])
544   #covari veis risco 1 formado
545   coefModelR1<-efeitos.fixos$Estimativa[c
        (13:17,23,19,21,25,27,28)]
546   #baseline Risco 2 Evadido
547   Exr0t2<-exp(efeitos.fixos$Estimativa[29:40])
548   #covariavesi risco 2 Evadido
549   coefModelR2<- efeitos.fixos$Estimativa[c
        (41:45,51,47,49,53,55,56)]
550
551   #Termos para o risco 1
552   #produto dos coeficientes pelo perfil
```

```
553   ExCvt1<-exp(perfil%*%coefModelR1)
554   #Termos para o risco 2
555   #produto das covari veis pelo perfil
556   ExCvt2<-exp(perfil%*%coefModelR2)
557
558   #Fucoes de risco
559   #Risco Formado
560   lambda_risFC<-(Exr0t1)*(ExCvt1)/(1+Exr0t1*ExCvt1+Exr0t2*ExCvt2)
561   #Risco Evadido
562   lambda_risEC<-Exr0t2*ExCvt2/(1+Exr0t1*ExCvt1+Exr0t2*ExCvt2)
563   #Sob Evadido
564   sobr_risEC<-cumprod(1-lambda_risEC[1:12])
565   #Sob Formado
566   sobr_risFC<-cumprod(1-lambda_risFC[1:12])
567
568   salida<-list(lambda_risEC,lambda_risFC,sobr_risEC,sobr_risFC)
569   return(salida)
570 }
571
572 #2012,Publica,Masculino,>20,Sim
573 perfil_M1<-PerfilRiscos_M1(c(0,0,0,0,0,1,1,1,1,473,426),Modelo1)
574 perfil_M2<-PerfilRiscos_M2(c(0,0,0,0,0,1,1,1,1,473,426),Modelo_2)
575 perfil_M3<-PerfilRiscosM_3(c(0,0,0,0,0,1,1,1,1,473,426))
576 #2012,Outros,Masculino,<21,Sim
577 perfil_M1<-PerfilRiscos_M1(c(0,0,0,0,0,0,1,0,1,473,426),Modelo1)
578 perfil_M2<-PerfilRiscos_M2(c(0,0,0,0,0,0,1,0,1,473,426),Modelo_2)
579 perfil_M3<-PerfilRiscosM_3(c(0,0,0,0,0,0,1,0,1,473,426))
580
581 dados$baseline <- as.numeric(dados$baseline)
582 dados$Tempo <- 1
583 dados$Tempo <- ave(dados$baseline, dados$id, FUN = max )
584
585 dados.km <- subset(dados, ANOING==2012 & EM=="Outros" &Sexo=="M"
      & Idade=="<21" &baseline==Tempo)
586 dados.km$Tempo
587
588 KaplanM <- survfit(Surv(Tempo, Cen1Formado)~1, dados.km)
589
590 KaplanM$surv
591 plot(KaplanM)
592
593 # sob Formado
```

```
594
595 dfsf<-data.frame(perfil_M1[[4]],perfil_M2[[4]],perfil_M3[[4]],"KM
      "=KaplanM$surv, semestre=c(1:12))
596 ggplot(dfsf, aes(semestre)) +
597   geom_step(aes(y = perfil_M1[[4]], colour = "Modelo 1 "  ),size
        =1) +
598   geom_step(aes(y = perfil_M2[[4]], colour = "Modelo 2 "  ),size
        =1) +
599   geom_step(aes(y = perfil_M3[[4]], colour = "Modelo 3 "  ),size
        =1)+
600 geom_step(aes(y = KM, colour = "K-M "  ),size=1)+
601   scale_x_continuous("Semestre", labels = as.character(
        dfsf$semestre),     breaks =dfsf$semestre)+scale_y_continuous
        (breaks=seq(0, 1, 0.1))+
602   ylab("S(t)")+  scale_color_discrete(name = "")+
603   coord_cartesian(ylim = c(0,1)) +theme_bw()+theme(legend.
        position = "bottom",legend.text = element_text(size=10))
604
605 # risco Formado
606 dfhf<-data.frame(perfil_M1[[2]],perfil_M2[[2]],perfil_M3[[2]],
      semestre=c(1:12))
607 ggplot(dfhf, aes(semestre)) +
608   geom_line(aes(y = perfil_M1[[2]], colour = "Modelo 1"  ),size
        =1) +             geom_line(aes(y = perfil_M2[[2]], colour =
        "Modelo 2"),size=1)+
609   geom_line(aes(y = perfil_M3[[2]], colour = "Modelo 3"  ),size
        =1) +             scale_x_continuous("Semestre", labels = as.
        character(dfhf$semestre),     breaks =dfhf$semestre)+
        scale_y_continuous(breaks=seq(0, 1, 0.1))+
610   ylab("Taxa de risco")+
611   scale_color_discrete(name = "")+
612   coord_cartesian(ylim = c(0,0.7)) +theme_bw()+theme(legend.
        position = "bottom",legend.text = element_text(size=10))
613
614 ###risco Evadido
615 dfhe<-data.frame(perfil_M1[[1]],perfil_M2[[1]],perfil_M3[[1]],
      semestre=c(1:12))
616 ggplot(dfhe, aes(semestre)) +
617   geom_line(aes(y = perfil_M1[[1]], colour = "Modelo 1 "  ),size
        =1) +          geom_line(aes(y = perfil_M2[[1]], colour = "
        Modelo 2"),size=1)+
618   geom_line(aes(y = perfil_M3[[1]], colour = "Modelo 3 "  ),size
```

```
    =1) +           scale_x_continuous("Semestre", labels = as.
    character(dfhe$semestre),     breaks =dfhe$semestre)+
    scale_y_continuous(breaks=seq(0, 1, 0.1))+
619   ylab("Taxa de risco")+
620   scale_color_discrete(name = "")+
621   coord_cartesian(ylim = c(0,0.4)) +theme_bw()+theme(legend.
    position = "bottom",legend.text = element_text(size=10))
622
623 #Sob Formado
624
625 dfsf<-data.frame(perfil_M1[[4]],perfil_M2[[4]],perfil_M3[[4]],
    semestre=c(1:12))
626 ggplot(dfsf, aes(semestre)) +
627   geom_step(aes(y = perfil_M1[[4]], colour = "Modelo 1 "  ),size
    =1) +
628   geom_step(aes(y = perfil_M2[[4]], colour = "Modelo 2 "  ),size
    =1) +
629   geom_step(aes(y = perfil_M3[[4]], colour = "Modelo 3 "  ),size
    =1) +scale_x_continuous("Semestre", labels = as.character(
    dfsf$semestre),     breaks =dfsf$semestre)+scale_y_continuous
    (breaks=seq(0, 1, 0.1))+
630   ylab("S(t)")+  scale_color_discrete(name = "")+
631   coord_cartesian(ylim = c(0,1)) +theme_bw()+theme(legend.
    position = "bottom",legend.text = element_text(size=10))
632
633 #Sob Evadido
634
635 dfse<-data.frame(perfil_M1[[3]],perfil_M2[[3]],perfil_M3[[3]],
    semestre=c(1:12))
636 ggplot(dfse, aes(semestre)) +
637   geom_step(aes(y = perfil_M1[[3]], colour = "Modelo 1 "  ),size
    =1) +
638   geom_step(aes(y = perfil_M2[[3]], colour = "Modelo 2 "  ),size
    =1) +
639   geom_step(aes(y = perfil_M3[[3]], colour = "Modelo 3 "  ),size
    =1) +scale_x_continuous("Semestre", labels = as.character(
    dfse$semestre),     breaks =dfse$semestre)+scale_y_continuous
    (breaks=seq(0, 1, 0.1))+
640   ylab("S(t)")+  scale_color_discrete(name = "")+
641   coord_cartesian(ylim = c(0,1)) +theme_bw()+theme(legend.
    position = "bottom",legend.text = element_text(size=10))
642
```

```
643 #compara  o de perfis do modelo final ----
644 #Mean mat =473
645 #Mean bio=426
646 #perfil
647
648 #2012,Publica,Masculino,>20,Sim
649 perfilM<-PerfilRiscosM_3(c(0,0,0,0,0,1,1,1,1,473,426))
650 perfilMSexoF<-PerfilRiscosM_3(c(0,0,0,0,0,1,0,1,1,473,426))
651 perfilMIdadeMenor20<-PerfilRiscosM_3(c(0,0,0,0,0,1,1,0,1,473,426)
    )
652 perfilMEMOtros<-PerfilRiscosM_3(c(0,0,0,0,0,0,1,1,1,473,426))
653 perfilMCursinhoNao<-PerfilRiscosM_3(c(0,0,0,0,0,1,1,1,0,473,426))
654 perfilMAnoIngr2014<-PerfilRiscosM_3(c(1,0,0,0,0,1,1,1,1,473,426))
655 semestre<-c(1:12)
656
657 #Forma do de risco Formado
658 dfhf<-data.frame(perfilM[[2]],perfilMSexoF[[2]],
    perfilMIdadeMenor20[[2]],perfilMEMOtros[[2]],perfilMCursinhoNao
    [[2]],perfilMAnoIngr2014[[2]],semestre)
659 ggplot(dfhf, aes(semestre)) +
660   geom_line(aes(y = perfilM[[2]], colour = "Medio"  ),size=1) +
                  geom_line(aes(y = perfilMSexoF[[2]], colour = "
      Sexo=F"),size=1)+
661   geom_line(aes(y = perfilMIdadeMenor20[[2]], colour = "Idade<20"
       ),size=1) +  geom_line(aes(y = perfilMEMOtros[[2]], colour
      = " EM=Otros"),size=0.5)+
662   geom_line(aes(y = perfilMCursinhoNao[[2]], colour = "Cursinho=
      Nao"),size=1) + geom_line(aes(y = perfilMAnoIngr2014[[2]],
      colour = "AnoIng=2014"),size=1)+
663   scale_x_continuous("Semestre", labels = as.character(
      dfhf$semestre),     breaks =dfhf$semestre)+scale_y_continuous
      (breaks=seq(0, 1, 0.1))+
664   ylab("Taxa de risco")+
665   scale_color_discrete(name = "")+
666   coord_cartesian(ylim = c(0,0.7)) +theme_linedraw()+theme(legend
      .position = "bottom",legend.text = element_text(size=10))
667
668 #Taxa risco Evadido
669 ###Funcoes de risco Evadido
670 dfhf<-data.frame(perfilM[[1]],perfilMSexoF[[1]],
    perfilMIdadeMenor20[[1]],perfilMEMOtros[[1]],perfilMCursinhoNao
    [[1]],perfilMAnoIngr2014[[1]],semestre)
```

```
671 ggplot(dfhf, aes(semestre)) +
672   geom_line(aes(y = perfilM[[1]], colour = "Medio"  ),size=1) +
                  geom_line(aes(y = perfilMSexoF[[1]], colour = "
      Sexo=F"),size=1)+
673   geom_line(aes(y = perfilMIdadeMenor20[[1]], colour = "Idade<20"
       ),size=1) +  geom_line(aes(y = perfilMEMOtros[[1]], colour
      = " EM=Otros"),size=1)+
674   geom_line(aes(y = perfilMCursinhoNao[[1]], colour = "Cursinho=
      Nao"  ),size=1) + geom_line(aes(y = perfilMAnoIngr2014[[1]],
      colour = "AnoIng=2014"),size=1)+
675   scale_x_continuous("Semestre", labels = as.character(
      dfhf$semestre),     breaks =dfhf$semestre)+scale_y_continuous
      (breaks=seq(0, 1, 0.1))+
676   ylab("Taxa de risco")+
677   scale_color_discrete(name = "")+
678   coord_cartesian(ylim = c(0,0.7)) +theme_linedraw()+theme(legend
      .position = "bottom",legend.text = element_text(size=10))
679
680 #Recuperar a funcao de sobrevivencia ----
681 #Sob Formado
682
683 dfhf<-data.frame(perfilM[[4]],perfilMSexoF[[4]],
     perfilMIdadeMenor20[[4]],perfilMEMOtros[[4]],perfilMCursinhoNao
     [[4]],perfilMAnoIngr2014[[4]],semestre)
684 ggplot(dfhf, aes(semestre)) +
685   geom_step(aes(y = perfilM[[4]], colour = "Medio"  ),size=1) +
                  geom_step(aes(y = perfilMSexoF[[4]], colour = "
      Sexo=F"),size=1)+
686   geom_step(aes(y = perfilMIdadeMenor20[[4]], colour = "Idade<20"
       ),size=1) +  geom_step(aes(y = perfilMEMOtros[[4]], colour
      = " EM=Otros"),size=1)+
687   geom_step(aes(y = perfilMCursinhoNao[[4]], colour = "Cursinho=
      Nao"  ),size=1) + geom_step(aes(y = perfilMAnoIngr2014[[4]],
      colour = "AnoIng=2014"),size=1)+
688   scale_x_continuous("Semestre", labels = as.character(
      dfhf$semestre),     breaks =dfhf$semestre)+scale_y_continuous
      (breaks=seq(0, 1, 0.1))+
689   ylab("S(t)")+
690   scale_color_discrete(name = "")+
691   coord_cartesian(ylim = c(0,1)) +theme_linedraw()+theme(legend.
      position = "bottom",legend.text = element_text(size=10))
692
```

```
693 #Sob Evadido
694
695 dfhf<-data.frame(perfilM[[3]],perfilMSexoF[[3]],
       perfilMIdadeMenor20[[3]],perfilMEMOtros[[3]],perfilMCursinhoNao
       [[3]],perfilMAnoIngr2014[[3]],semestre)
696 ggplot(dfhf, aes(semestre)) +
697   geom_step(aes(y = perfilM[[3]], colour = "Medio"  ),size=1) +
                    geom_step(aes(y = perfilMSexoF[[3]], colour = "
       Sexo=F"),size=1)+
698   geom_step(aes(y = perfilMIdadeMenor20[[3]], colour = "Idade<20"
       ),size=1) +  geom_step(aes(y = perfilMEMOtros[[3]], colour
       = " EM=Otros"),size=1)+
699   geom_step(aes(y = perfilMCursinhoNao[[3]], colour = "Cursinho=
       Nao"  ),size=1) + geom_step(aes(y = perfilMAnoIngr2014[[3]],
       colour = "AnoIng=2014"),size=1)+
700   scale_x_continuous("Semestre", labels = as.character(
       dfhf$semestre),     breaks =dfhf$semestre)+scale_y_continuous
       (breaks=seq(0, 1, 0.1))+
701   ylab("S(t)")+
702   scale_color_discrete(name = "")+
703   coord_cartesian(ylim = c(0,1)) +theme_linedraw()+theme(legend.
       position = "bottom",legend.text = element_text(size=10))
```