Universidade Federal Fluminense

FACULDADE DE ECONOMIA

EDUARDA OLIVEIRA RODRIGUES

ANÁLISE DO FLUXO ESCOLAR À LUZ DA ANÁLISE DE SOBREVIVÊNCIA: UM ESTUDO PARA A UFF

NITERÓI – RJ

2019

EDUARDA OLIVEIRA RODRIGUES

ANÁLISE DO FLUXO ESCOLAR À LUZ DA ANÁLISE DE SOBREVIVÊNCIA: UM ESTUDO PARA A UFF

Monografia apresentada ao curso de Bacharelado em Ciências Econômicas da Universidade Federal Fluminense como requisito parcial para conclusão do curso.

Orientador:
Prof. Dr. Jesus Alexei Luizar Obregon

Niterói – RJ

2019

EDUARDA OLIVEIRA RODRIGUES

ANÁLISE DO FLUXO ESCOLAR À LUZ DA ANÁLISE DE SOBREVIVÊNCIA: UM ESTUDO PARA A UFF

Monografia apresentada ao curso de Bacharelado em Ciências Econômicas da Universidade Federal Fluminense como requisito parcial para conclusão do curso.

BANCA EXAMINADORA

Prof. Dr. Jesus Alexei Luizar Obregon
Orientador
Universidade Federal Fluminense

Profa. Dra. Danielle Carusi Machado
Universidade Federal Fluminense

Prof. Dr. André Barbosa Oliveira
Universidade Federal Fluminense

RESUMO

O presente estudo tem como objetivo fortalecer atividades de pesquisa sobre a evasão escolar do ensino superior. Utilizando o ferramental da análise de sobrevida, tratamos as informações obtidas através do Censo da Educação Superior afim de acompanhar, de 2010 a 2017, a sobrevivência de diferentes perfis de estudantes matriculados no ano de 2010 na Universidade Federal Fluminense até a formatura, o evento da evasão escolar ou a censura da observação. A partir dos resultados, podemos entender a relação entre as características e a evasão escolar, e como alguns grupos – como estudantes de cursos noturnos, em relação aos de cursos integrais – são mais vulneráveis ao evento de evasão.

Palavras-chave: Fluxo Escolar; Análise de Sobrevivência; Ensino Superior.

Área: Métodos Quantitativos em Economia

ABSTRACT

The present study aims to strengthen research activities on school dropout from higher education. Using survival analysis, we treat information obtained through the national Higher Education Census to monitor, from 2010 to 2017, the survival of different profiles of students enrolled in the year 2010 at Universidade Federal Fluminense until graduation, the evasion event school or censorship of the observation. From the results, we can understand the relationship between characteristics and school dropout, and how some groups – such as students enrolled in evening courses, compared to those in full-time courses – are more vulnerable to the dropout event

Key-words: School Flow; Survival Analysis; University Education.

Area: Quantitative Methods in Economics

AGRADECIMENTOS

Agradeço primeiramente aos meus pais, Tatiana e Leandro, que me proporcionaram uma excelente educação, dentro e fora de casa, que me permitiu chegar onde estou hoje.

Ainda, a todos os professores e professoras por cada ensinamento, e em especial os professores Jesus Alexei e Danielle Carusi, que tanto me guiaram neste projeto.

Por fim, agradeço ao meu querido e incansável companheiro Hugo, que me ajudou a manter a sanidade nesses períodos tão turbulentos.

LISTA DE FIGURAS

LISTA DE TABELAS

DICIONÁRIO DE SIGLAS

AFT: *Accelerated Failure Time*, modelo de tempo de vida acelerado

ANDIFES: Associação Nacional dos Dirigentes das Instituções Federais de Ensino Superior

EAD: Educação à distância

EE: Educação especial

EI: Educação inclusiva

EKM: Estimador de Kaplan-Meier

ENEM: Exame Nacional do Ensino Médio

FORPLAD: Fórum Nacional de Pró-Reitores de Planejamento e Administração

ID: Código identificador único

INEP: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

OCDE: Organização para a Cooperação e Desenvolvimento Econômico

PROUNI: Programa Universidade Para Todos

REUNI: Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais

UFF: Universidade Federal Fluminense

UFPB: Universidade Federal da Paraíba

URGS: Univerdade Federal do Rio Grande do Sul

SUMÁRIO

1 INTRODUÇÃO

Ao longo das últimas duas décadas, as mudanças socioeconômicas observadas nos países em desenvolvimento têm sido, de modo geral, acompanhadas de políticas governamentais de cunho social. No Brasil, programas que seguem a evolução das demandas de estratos sociais adquirem maior relevância e são complementados com ações que promovem mudanças favoráveis na redução das disparidades sociais.

A ampliação do ensino superior federal foi notável, principalmente entre 2003 e 2016, através de diversas medidas. No ano de 2007, o governo federal lançou o Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais, o REUNI (MEC, 2007). Em termos gerais, o objetivo anunciado do programa seria reduzir as diferenças sociais advindas do componente educacional.

O REUNI, em seu documento de Diretrizes Gerais, elaborado pelo Ministério da Educação em 2007, apresentou como meta global do programa, que deveria ser alcançada ao final de cinco anos, que *a taxa de conclusão média dos cursos de graduação presenciais deveria atingir 90%* (MEC, 2007). Hoje ainda, entretanto, a evasão escolar em níveis alarmantes perdura. De acordo com o Mapa do Ensino Superior no Brasil 2019, realizado pela Samesp, ainda que tenha ocorrido queda de evasão entre 2016 e 2017 (de 27,2% para 25,9% em cursos presenciais e de 36,1% para 34,3% em cursos à distância), os níveis ainda continuam longe do alvo das Diretrizes Gerais (DESAFIOS DA EDUCACÃO, 2019).

Em 2016, o FORPLAD (Fórum Nacional de Pró-Reitores de Planejamento e Administração), entidade interna da ANDIFES (Associação Nacional dos Dirigentes das Instituições Federais de Ensino Superior)[1], apresentou resultados de estudos que abordam o fluxo escolar no ensino superior para o período 2010-2014 (FORPLAD, 2016). O fluxo escolar é analisado nas dimensões de retenção, evasão e conclusão universitária, as quais implicitamente estão atreladas com o tempo de escolaridade no nível superior e por sua vez com a desigualdade educacional.

As informações contidas no relatório do grupo de trabalho do FORPLAD (2016) mostram, que em geral, a taxa de evasão no ensino superior brasileiro seguiu uma tendência crescente no período 2010-2014, registrando valores inicial e final respectivamente, de 12,9% e de 15,8%. No que se refere as áreas de conhecimento classificadas segundo a Organização para a Cooperação e Desenvolvimento Econômico (OCDE), os autores do mesmo relatório indicam que no ano de 2014 os cursos da área de Saúde e Bem Estar Social apresentam menores taxas de evasão e de retenção, respectivamente 7,8% e 9,1% enquanto os indicadores com maiores valores de evasão e retenção, respectivamente 20,0% e 27,1% são atribuídos aos cursos agrupados nas áreas de

[1] A Associação Nacional dos Dirigentes das Instituições Federais de Ensino Superior, ANDIFES, é a representante das universidades federais na interlocução com o governo federal, com as associações de professores, de técnico-administrativos, de estudantes e com a sociedade brasileira.

Ciências, Matemática e Computação.

O evento de evasão escolar causa, naturalmente, uma série de prejuízos à todos os agentes envolvidos. Em termos de perdas econômicas, por exemplo, é possível citar as perdas que afetam o aluno evadido, dado que não obterá o prêmio relativo ao título de ensino superior, e as perdas que afetam à sociedade (e o Estado), pelos investimentos públicos que não proporcionaram retorno adequado ao que foi esperado; justamente por isso, é fundamental, conforme reconhecido, a compreensão do problema da evasão. De acordo com Oscar Hipólito em entrevista para o G1 (2011), em 2009 o país perdeu mais de R$ 9 bilhões de reais por causa da evasão no ensino superior – o que traz uma dimensão, ao menos monetária, do problema.

O tempo de escolaridade junto com a evasão no ensino superior têm recebido atenção por parte da comunidade acedêmica dedicada ao estudo do fluxo escolar no ensino superior desde a década de 1990. Mais especificamente, a partir de 1995, com o Seminário Sobre Evasão nas Universidades Brasileiras, organizado pela Secretaria de Educação Superior do Ministério da Educação, diversos autores se propuseram à discutir o tema, entrando inclusive em aplicações específicas, como Lima Junior, Silveira e Ostermann (2012), que aplicaram à faculdade de física da UFRGS, Franca e Saccaro (2018), aos cursos de ensino à distância das universidades federais, Hoed (2016), à área da computação, e até a própria Universidade Federal Fluminense (UFF), que conduziu estudos, em 2012 e 2013, com o objetivo de analisar as características dos estudantes retidos (que não integralizaram o curso à época que deveria já ter feito), considerando tamanho amostral de 290 alunos (PONTES, 2015).

1.1 Objeto de pesquisa

O complexo problema da evasão escolar traz, intrinsecamente, uma série de custos. Portanto, sob uma perspectiva econômica e social, a importância da compreensão das variáveis que influenciam o evento é evidente. Identificar e compreender os efeitos de covariação entre características socio-econômicas e demográficas com a evasão no ensino superior se mostra essencial, portanto, para observar o grau de propensão à evasão em diversos grupos.

A Universidade Federal Fluminense atualmente oferece 131 cursos de graduação, em 32 municípios do Estado do Rio de Janeiro, apresentando variabilidade das características e condições socioeconômicas dos alunos desta universidade. Em 2017, segundo o Censo da Educação Superior, UFF ofereceu maior quantidade de vagas nos cursos de graduação presencial dentre as universidades federais. O índice de evasão escolar na UFF para intervalo 2016 a 2017 foi de 19,5%, abaixo da média geral no Brasil de 26,4%, porém acima da média das universidades federais de 16,1%[2].

Por apresentar tal riqueza de características, é possível, através de microdados obtidos dos Censos Escolares de Ensino Superior, replicar metodologias de análise de evasão escolar

[2] Disponível em: <http://www.uff.br/?q=uff-em-numeros-0>. Acesso em: 4 de novembro de 2019

para o caso específico da Universidade Federal Fluminense, delimitando, portanto, como objeto de estudo, a evasão escolar nos cursos de graduação da UFF.

1.2 Justificativas

Conforme abordado, os alunos matriculados contribuem significativamente com os custos de uma instituição de ensino superior; a não realização da conclusão gera, consequentemente, recursos desperdiçados, dado que o estudante não obtém a devida sinalização de capacitação, e a instituição, seja pública ou privada, lida com o custo de oportunidade de não despender orçamento em alguém que poderia vir a concluir a gradução. No caso público, tal orçamento advém de recursos públicos, afetando diretamente a sociedade[3], e no caso de instituições privadas, ocasiona perdas claras de receita. Para além das perdas orçamentárias, a ocorrência da evasão priva a sociedade das externalidades positivas ocasionadas por indivíduos qualificados, incluindo aumento de produtividade, inovação, dentre outros.

Apesar da discussão sobre fluxo escolar, e mais especificamente sobre retenção e evasão, estar em pauta há mais de duas décadas, resultados razoáveis não foram alcançados (como por exemplo, a meta global de combate à evasão do REUNI), o que justifica o esforço em desenvolver contribuições para o debate.

1.3 Objetivos

O objetivo do presente trabalho é, através do arcabouço teórico apropriado, utilizado em estudos de evasão aplicados à casos específicos, identificar e compreender os efeitos de covariação entre características socio-econômicas e demográficas com a evasão no ensino superior. Isso permite observar o grau de propensão à evasão em diversos grupos discentes da Universidade Federal Fluminense, e verificar a hipótese de que certos perfis - como por exemplo, se pessoas negras e pardas, ou pessoas que cursam o turno noturno - tendem a um grau de evasão mais elevado, sendo, portanto, alvos de possíveis políticas de permanência mais rigorosas, para evitar a ocorrência de tal evento.

Ainda, tem como objetivo explorar a literatura brasileira existente sobre o tema, tanto em termos téoricos quanto metodológicos, e principalmente, colaborar com exemplos de pseudo-códigos que explorem o ferramental de análise de sobrevida e seus respectivos testes em pacotes de análise estatística para posteriores avaliações similares.

[3] De acordo com relatório do FORPLAD (2018), nas universidades federais somente a despesa com pessoal ativo foi de R$18.037,1 por aluno ao ano. Considerando também outras despesas com pessoal, incluindo pessoal inativo, precatórios pessoais, etc, esse valor chegou a R$30.812,8. Incluindo demais gastos, o cálculo alcançou os R$37.551,2 por aluno ao ano.

2 REVISÃO BIBLIOGRÁFICA: EVASÃO NO ENSINO SUPERIOR

Neste capítulo apresentamos de forma sucinta os conceitos básicos da evasão escolar, percorrendo brevemente resultados obtidos em pesquisas prévias, abordando tanto técnicas mais gerais e indicadores desenvolvidos para a avaliação da evasão, quanto a análise de sobrevida especificamente, que será o arcabouço utilizado neste trabalho.

2.1 Técnicas Gerais

Indicadores gerais são amplamente utilizados por instituições e veículos de mídia para compreensão do fluxo estudantil – mais especificamente, a evasão – e divulgação de resultados, em contraste com técnicas mais específicas e complexas, como a própria análise de sobrevida. Na seção a seguir, tais conceitos, bem como resultados obtidos em estudos que se utilizaram de tais indicadores, serão abordados.

2.1.1 Indicadores

Na literatura mais geral, existem diversos indicadores que permitem fazer o acompanhamento da evasão estudantil ao longo do tempo, que são respectivamente aplicados uma vez definidos o conceito de evasão a ser considerado.

A definição do conceito de evasão é bastante relevante, dado que diz respeito ao que efetivamente é considerado evasão - a partir de qual momento, ou em qual nível de desligamento se pode considerar o estudante como evadido. Sobre isso, Vitelli e Fritsch (2016) apontam os conceitos de granularidade - podemos considerar evasão como desconexão com o próprio curso, com a instituição de educação ou com o sistema de educação como um todo - e de temporalidade - onde podemos considerar evadido o aluno assim que ele se desconecta do curso, instituição ou sistema, após um determinado período, ou apenas quando ele não retorna.

Definindo seus próprios conceitos de evasão, diversas instituições, como a já mencionada ANDIFES, observando FORPLAD (2015), utilizam amplamente de tais indicadores. Em seu trabalho, Cândido (2019) compila diferentes tipos de indicadores frequentemente utilizados para avaliar evasão escolar, se utilizando da literatura específica. O autor classifica tais indicadores em três categorias:

1. Os que mensuram a variação anual do abandono universitário;
2. Os que realizam acompanhamento longitudinal;
3. Os que acompanham individualizadamente um grupo ao longo do tempo, o chamado acompanhamento de coortes.

O principal indicador de mensuração anual é o chamado Índice de Evasão Anual (IEA), representado por:

$$IEA_n = 1 - P_n = 1 - \frac{M_n - I_n}{C_n} \quad , \tag{2.1}$$

o subíndice n representa o ano no qual mensurada a evasão, M_n e I_n denotam respectivamente o número de matriculados e o número de ingressantes. O termo do denominador C_n corresponde ao número de estudantes que estão cursando; isto é, matriculados no ano $n-1$ e ainda não formandos no ano n.

A Taxa de Titulação (TT), principal indicador de acompanhamento longitudinal apontado por Cândido (2019), é representado por:

$$TT_{n,x} = \frac{F_n}{I_{n-x}} \quad , \tag{2.2}$$

onde o denominador I_{n-x} representa o número de estudantes com status de ingressante de um curso no ano $n-x$, sendo x o período para integralização curricular, e o numerador F_n corresponde ao total de concluintes após o período de integralização. Finalmente, a Evasão Total média (ET) é calculada a partir da Taxa de Titulação, e expressa por:

$$ET_{n,x} = 1 - TT_{n,x} \quad . \tag{2.3}$$

Outra técnica apontada pelo autor, o **acompanhamento de coortes**, é menos presente na literatura nacional do que as demais técnicas, mas possibilita o entendimento do comportamento dos sujeitos ao longo do tempo, coisa não obtida em métricas de variação anual ou acompanhamento longitudinal. Cândido (2019) define a análise por coorte como:

> Um estudo em que se acompanha ao longo do tempo e considera-se como unidade básica de análise um determinado grupo cuja marca em comum seja a experiência de um mesmo evento em um mesmo intervalo de tempo (como a matrícula em um determinado semestre). (CÂNDIDO, 2019, p.28)

Explorando tal técnica, o autor apresenta três principais indicadores, um que capta o número de cursandos, um que capta o número de formandos, e, a partir dos dois anteriores, a evasão da coorte, apontados na Tabela 1.

2.1.2 Exemplos de resultados obtidos na literatura

Como exemplos de aplicação dos indicadores anteriormente mencionados, podemos observar diversos trabalhos voltados à avaliação do estado da evasão na educação brasileira, como a pesquisa de Oscar Hipólito noticiada no G1 (2011), que se utilizando do Índice de Evasão Anual (equação (2.1)), observou, no Censo da Educação Superior (2008-2009), uma evasão geral

Tabela 1 – Principais indicadores de coorte

Indicador	Fórmula	Descrição
Cursandos da Coorte (TCC)	$\text{TCC}_{a,n} = \frac{C_{a,n}}{I_a}$	C representa o número de cursandos e I o número de ingressantes. n se trata do ano para qual a avaliação está sendo realizada, e a o ano de ingresso, que define a coorte. Portanto, I_a representa o tamanho da coorte - o número de ingressantes no ano definido -, e $C_{a,n}$, o número de sujeitos ingressantes em a que estam cursando o ano n.
Formandos da Coorte (TFC)	$\text{TFC}_{a,n} = \frac{F_{a,n}}{I_a}$	A interpretação da relação acima é bastante similar à interpretação da equação de Taxa de Cursandos da Coorte (TCC), mas ao invés de cursandos ($C_{a,n}$), temos formandos ($F_{a,n}$) - ou seja, o número de sujeitos ingressantes no ano a que tem status de formando no ano n.
Evasão da Coorte (TEC)	$\text{TEC}_{a,n} = 1 - TFC_{a,n} - TCC_{a,n}$	Por fim, temos a métrica de evasão, que é tão simplesmente o complementar da soma das métricas de taxa de cursandos e formandos - ou seja: todos que não estão cursando ou se formando, são entendidos como evadidos.

de 20,9%, sendo 10,5% dentre as instituições públicas, e 24,5% dentre as privadas. Esse mesmo estudo, já mencionado anteriormente, chama atenção para a perda econômica da evasão: quase R$9 bilhões à época, se baseando nos custos médios por estudante em cada tipo de instituição.

Se utilizando da mesma metodologia, o próprio texto de Cândido (2019), aponta seus resultados, indicando que durante todo o período de 2010 a 2018, a evasão dentre as instituições privadas esteve acima do patamar de 25%, enquanto, dentre as públicas federais, para o mesmo período, há uma relativa estabilidade no nível de 15%, e para as públicas estaduais, níveis menores do que a das federais.

Também se utilizando de outra técnica mencionada – a de indicadores de coorte –, Cândido (2019), considerando dados do Censo da Educação Superior para a coorte de 2010, concluiu, a respeito do ritmo de evasão para diferentes esferas de ensino superior (federal, estadual e privada) que a concentração da evasão nos primeiros anos de curso é bastante intensa nas instituições privadas, intermediária em federais e razoável em instituições federais. Os valores da Taxa de Evasão da Coorte (TEC) foram de 20,83%, 12,89% e 7,01% ao final do primeiro ano para privadas, federais e estaduais, respectivamente; 38,34%, 25,56% e 18,64% ao final do segundo ano, e 54,30%, 47,29% e 40,82% ao final do oitavo ano.

2.2 Análise de sobrevida na literatura

Apesar de muito úteis, análises de correlação, covariância, diferença de médias e outros indicadores mais simples não capturam, no entanto, a evolução temporal de uma certa observação. Através da análise de sobrevivência, essa dimensão passa a ser devidamente considerada, e se torna possível entender a trajetória comportamental de um grupo observado, bem como individualmente. Conforme definem Lima Junior, Silveira e Ostermann (2012), a respeito da análise de sobrevida:

> O objetivo de estudo da análise de sobrevivência é o tempo entre eventos, por exemplo: o tempo do diagnóstico à morte de um paciência, o tempo da remissão à recidiva de uma doença, o tempo de venda de um automóvel até seu primeiro defeito mecânico, o tempo de soltura de um preso à sua re-incidência no crime, o tempo de ingresso em um curso de graduação ao desligamento, evasão ou diplomação. Do ponto de vista estatístico, todas essas situações podem ser abordadas com as mesmas ferramentas. (LIMA JUNIOR; SILVEIRA; OSTERMANN, 2012)

Uma característica importante e específica de tal metodologia de análise é a presença do que foi cunhado como **observações censuradas** - ou seja, intervalos incompletos (observações que não atingiram um evento de interesse), mas que não devem ser descartadas. É comum, de acordo com os autores, que um estudo seja encerrado antes de todos as observações atingirem o evento de interesse. Por exemplo, “o acompanhamento de um grupo de pacientes com doença grave, pode ser necessário encerrar a pesquisa antes que todos tenham falecido”; os pacientes

ainda vivos ao final do estudo seriam considerados dados censurados, mas perderíamos informação muito relevante descartando-os da análise e traria *viés* nas funções de sobrevivência. Para evitar tal viés e perda de informação, as técnicas de sobrevida passaram por décadas de aperfeiçoamento para incorporação de tais dados censurados.

Conforme a citação anterior, de Lima Junior, Silveira e Ostermann (2012), ao longo do tempo a análise de sobrevida deixou de ser utilizada somente em testes clínicos – motivo pelo qual foi inicialmente desenvolvida –, e passou a figurar como importante ferramenta nos mais diversos campos. Sua aplicação especificamente no problema de evasão no ensino superior brasileiro pode ser observada em diversos artigos, como o já citado Lima Junior, Silveira e Ostermann (2012), Franca e Saccaro (2018), Silva et al. (2018), Mello et al. (2015), dentre outros.

Lima Junior, Silveira e Ostermann (2012) avaliaram em seu texto o fluxo no cursos de graduação em na área de exatas da Universidade Federal do Rio Grande do Sul (UFRGS). Como principais problemas a serem estudados, os autores apontam a evasão e a retenção - ou seja, a desistência e a permanência prolongada, respectivamente. Observando o registro acadêmico dos estudantes, os autores se utilizaram da análise de sobrevida para avaliar e modelar os dados.

Os achados apontaram que os níveis de evasão nos cursos de física, química e matemática são similares entre si e mais elevados do que a evasão nos cursos de engenharia. Dentro dos cursos de física, as estatísticas descritivas não mostraram diferença significativa nas distribuições de evadidos e diplomados considerando diferentes habilitações (bacharelado e licenciatura) e ou o sexo dos estudantes (homens e mulheres).

Os dados coletados diziam respeito a 1447 observações (estudantes) e cinco variáveis: tempo de permanência no curso, situação do registro (ativo, afastado, diplomado, transferido, evadido, etc), habilitação (bacharelado ou licenciatura), pontuação obtida no vestibular e sexo. As observações censuradas, nesse caso, são aqueles estudantes cuja matrícula estava ativa ou trancada (ou seja, aqueles que futuramente incorrerão em eventos terminais - seja se diplomar ou evadir[1]).

Utilizando as técnicas da análise de sobrevida, os autores percebem o crescimento das ocorrências de diplomação após 3,5 anos de permanência. A proporção de graduados por semestre é máxima entre 3,5 e 4,5 anos, e após isso a probabilidade do estudante diplomar é reduzida; no oitavo ano, já é praticamente nula. Isso significa dizer, em outras palavras, que, caso o estudante não se diplome em até sete a nove semestres, as chances de diplomação se reduzem, e a cada semestre se tornam mais baixas.

Ainda, em termos de evasão, os autores observam sua ocorrência desde o primeiro semestre, só passando a ser nula no décimo ano. Apesar da evasão ocorrer de modo mais intenso nos primeiros períodos, ao fim do quarto ano, quase metade dos eventos ainda não

[1] Seja por transferência ou abandono

ocorreram. Como citado nas linhas acima, o trabalho de Lima et al ((LIMA JUNIOR; SILVEIRA; OSTERMANN, 2012) aborda a evasão na UFRGS, dando lugar a estudos mais abragentes no sentido populacional; isto se aplica para as universidades brasileiras, bem como no sentido metodológico, que em outras palavras sejam incorporadas outras técnicas de análise de sobrevida.

Seguindo o caminho de estudos mais focalizados, Silva et al. (2018) avaliaram a evasão especificamente dentre alunos de estatística da Universidade Federal da Paraíba (UFPB). A amostra utilizada, derivada de dados do Núcleo de Tecnologia da Informação da UFPB, no entanto, conta com apenas 132 observações.

Preliminarmente, observaram que a maioria dos evadidos são homens (69,5%), e acrescentaram ainda que esse número é proporcional à demanda de alunos pelo bacharelado em estatística da UFPB (onde a maioria dos estudantes é do sexo masculino). Além disso, não houve diferença significativa entre os alunos que cursaram o ensino fundamental em escolas públicas e particulares. Por fim, a idade média dos evadidos foi de 28 anos, com desvio padrão de 8 anos.

De acordo com os autores, se utilizando de estimadores de sobrevivência,

> É possível verificar que a chance dos alunos que ingressam no bacharelado desistirem após o primeiro ano de curso é 45,85%, o que indica que mais de 50% dos alunos não conseguem sequer concluir o primeiro ano de curso, período referente às disciplinas básicas, onde estão inseridos os cálculos, que são responsáveis por grande parte das desistências. Em contrapartida, a probabilidade de um aluno desistir após o terceiro ano é de 15,07%, período relacionado às disciplinas especificas do curso de estatística. Ou seja, a partir do terceiro ano de curso a chance de desistência do aluno é bastante pequena. (SILVA et al., 2018, p.140)

Através de um modelo log-normal, observaram que as variáveis cor, naturalidade e forma de ingresso foram significativas para a explicação da evasão da amostra obtida.

Mello et al (2015), outro exemplo citado, avaliaram, com dados do Programa Rede São Paulo de Formação de Docentes, 76 alunos do programa, entre desistentes ou reprovados, sendo 60,5% da amostra composta por estudantes reprovados, e 39,5% de desistentes.

Como resultado, após utilizarem técnicas semi-paramétricos – por sua característica de adaptabilidade a diversos tipos de conjuntos de dados –, os autores obtiveram que o tempo de desistência ou abandono é muito curto – 40% das desistências ou reprovações já tinham ocorrido no segundo mês, com ápice entre 4 a 6 meses. Ainda, através do modelo de regressão estimado, observaram que a desistência ocorre antes do que a reprovação, em média.

Visando avaliar a evasão e suas determinantes a nível nacional, ao contrário dos trabalhos anteriores, que tinham um objeto de estudo mais específico, Franca e Saccaro (2018) se utilizaram dos mesmos métodos observados em Lima Junior, Silveira e Ostermann (2012) para avaliar dados dos Censos de Educação Superior entre 2009 e 2014 de estudantes de bacharelados presenciais das áreas de Ciência, Matemática e Computação e Engenharia, Produção e Construção.

As autoras observaram que "a maior evasão ocorreu em 2009 para os cursos de quatro [ou seja, no primeiro ano] e em 2010 para os cursos de cinco anos", e isso é válido para instituições públicas e privadas, sendo o nível de evasão das privadas maior que o das públicas - o abandono por período decresce mais rapidamente nas instituições públicas.

No contexto da análise de sobrevida, Franca e Saccaro observaram uma taxa de evasão média elevada; no primeiro ano, a taxa de sobrevivência gira em torno de 75% (ou seja, 25% de evasão), chegando a menos de 50% no final do período de seis anos. Ainda, "para avaliar se as curvas de sobrevivência apresentadas [...] são diferentes entre si de forma estatisticamente significante", realizaram testes de hipótese, e obtiveram como resultado nenhum indicativo de diferença significativa entre as curvas de evasão dos cursos públicos de quatro anos e privados de cinco anos para o começo. Para as demais combinações (de anos de curso e tipo de instituição), os p-valores são zero, o que leva a aceitação da hipótese alternativa de que as curvas são diferentes à 1% de significância.

Os resultados apontaram que indivíduos do sexo feminino tem um tempo de sobrevivência 5,5% maior - porém, para os cursos de cinco anos nas universidades privadas, tal taxa é negativa (portanto, o tempo de sobrevivência delas é 5,5% menor que o dos homens). Quanto à idade, quanto mais velho é o estudante, menor é sua taxa de sobrevivência.

Para todos os modelos estimados, o tempo de vida daqueles nas áreas de Ciências, Matemática e Computação é menor do que nas áreas de Engenharia, Produção e Construção – nos primeiros cursos, a evasão é maior. Em termos de turno, os alunos de cursos noturnos apresentam taxa de sobrevivência negativa, enquanto dos cursos integrais é positiva; as autoras ainda alertaram que tais fatores podem ser relacionados com o fato de que o número de cursos de cinco anos que são ofertados no turno integral é maior em relação aos demais, e com o fato de que alunos de cursos noturnos possuem média de idade maior.

De modo geral, Franca e Saccaro concluiram que "a sobrevivência dos estudantes das instituições públicas é 18,4% maior do que dos estudantes das instituições privadas", e que cursos que apresentam maiores níveis de sobrevivência são aqueles cujos alunos possuem melhores condições financeiras (e consequentemente mais oportunidade de se dedicar de forma integral à formação).

3 METODOLOGIA: ANÁLISE DE SOBREVIDA

Neste capítulo apresentamos os conceitos, definições e resultados principais no contexto matemático da teoria de análise de sobrevida. Ademais, optamos por apresentar primeiramente uma breve introdução a respeito da análise de sobrevida, e conceitos iniciais, seguido de métodos não-paramétricos (estimador de Kaplan-Meier e seus testes de hipótese), exemplos paramétricos, e por fim, modelos de regressão semi-paramétricos e paramétricos.

3.1 Introdução

A análise de sobrevivência é o resultado da modelagem estocástica de fenômenos epidemiológicos e tem como foco de estudo o tempo percorrido até a ocorrência de um evento previamente determinado ou prescrito, por exemplo, a morte de um paciente observado. Inicialmente, foi desenvolvida no contexto de estudos clínicos, mas rapidamente se tornou utilizada em diversas outras áreas. No contexto deste trabalho, o evento previamente determinado pode ser, por exemplo, o evento de diplomação, ou a evasão do aluno. O tempo percorrido até o evento recebe o nome de tempo de vida e adquire o papel de variável aleatória, no sentido probabilístico.

A variabilidade do tempo de vida pode ser explicada, naturalmente, pelas características próprias dos indivíduos em estudo. Estes, por sua vez, também estão sujeitos a outros estímulos alheios às condições do estudo e portanto alguns deles podem não estar presentes ao longo do estudo. As informações de tais sujeitos, no entanto, são consideravelmente importantes para o estudo enquanto participaram dele, então adquirem um papel especial e recebem o nome de dados censurados (COLOSIMO; GIOLO, 2006). Conforme já mencionado anteriormente, os dados censurados podem ser definidos como observações incompletas, e podem ser de três tipos:

1. À direita, quando o evento de interesse não é observado até o fim do estudo, ou em casos de remoção do indivíduo observado do estudo;
2. À esquerda, que ocorre quando não conhecemos o momento da ocorrência do evento, mas sabemos que ele ocorreu antes do tempo registrado;
3. Intervalar, onde se sabe que o evento de interesse ocorreu, mas não se sabe quando exatamente.

Em geral, literatura lida com casos de censura à direita, sendo casos de censura à esquerda bem menos comuns. Ainda que certas observações num estudo sejam censuradas, elas revelam informação sobre o tempo de vida dos sujeitos, e sua omissão pode causar conclusões viesadas.

Entre as vantagens da análise de sobrevida, além da já anteriormente mencionada capacidade de lidar com dados censurados, estão as diversas possibilidades de técnicas de estimação

e regressão, tanto paramétricas, quanto semi-paramétricas e não-paramétricas. Dobson e Barnett (2018), em seu capítulo sobre análise de sobrevida, abordam com maior foco os modelos paramétricos, argumentando serem mais amplos e precisos. Porém, considerando os modelos paramétricos, é necessário postular a lei probabilística que governa a variável em estudo, a qual por sua vez não necessariamente é conehcida ou tem expressão matemática manipulável para análises posteriores. Nesse contexto, são propostos também métodos não-paramétricos e semi-paramétricos, os mais utilizados na literatura em geral por sua flexibilidade e simplicidade.

Justamente por sua popularidade, procedimentos semi-paramétricos e não-paramétricos serão também abordados neste trabalho. O modelo de Cox e o estimador de Kaplan-Meier, mencionados em trabalhos como Lima Junior, Silveira e Ostermann (2012) e Franca e Saccaro (2018), são os procedimento mais utilizados dentre os trabalhos de análise de sobrevivência, e se tratam de métodos semi-paramétrico e não-paramétrico, respectivamente, onde nenhuma distribuição de probabilidade específica é assumida para os tempos de sobrevivência.

A principal vantagem dos métodos não-paramétricos consiste na abrangência ao não restringir alguma propriedade ou característica na variável tempo de vida. Este método vai permitir realizar análise preliminares de comparação entre subgrupos amostrais regidos por variáveis do tipo discreto, como sexo do estudante, campus, etc.

Os efeitos das variáveis condicionantes – ou seja, idade, campus da universidade, turno do curso ou outras, na evasão ou retenção de estudantes – devem ser obtidos com os métodos semi-paramétricos, mais explícitamente um modelo de Cox, que segue a estrutura padrão de decompor a variável dependente, tempo de vida, em um termo governado pelo conjunto de variáveis condicionantes e outro que representa a componente não explicada pelo modelo, parte que recebe o nome tradicional de erro.

3.2 Conceitos iniciais: Funções de sobrevivência e de risco

A primeira definição importante a ser feita é a respeito da variável aleatória Y, o objeto de estudo, que admitamos contínua e que representa o tempo de vida (DOBSON; BARNETT, 2018). A partir dela, a função de sobrevivência, probabilística ou estocástica, apresenta a forma de uma função de distribuição acumulada, podendo ser representada pela equação abaixo:

$$F(y) = P(Y \leq y) \quad . \tag{3.1}$$

Como dito, $F(y)$ é a função de distribuição acumulada de Y; sendo Y o tempo de vida, podemos entender que $P(Y \leq y)$ é a probabilidade de sobreviver até y unidades de tempo. Assim sendo, o complementar, a **função de sobrevivência** $S(y)$, corresponde à probabilidade de um indivíduo sobreviver (continuar) no seu estado inicial . Portanto, a probabilidade de um

indivíduo permanecer no curso superior até o tempo y é igual a:

$$S(y) = P(Y > y) = 1 - F(y) \quad . \tag{3.2}$$

A **função de risco** representa a chance da ocorrência do evento de falha num tempo entre y e $y + \delta y$ (um intervalo infinitesimalmente pequeno), dado que o sujeito sobreviveu até o tempo y. Simbolicamente,

$$h(y) = \lim_{\delta y \to 0} \frac{Pr(y \leq Y < y + \delta y | Y > y)}{\delta y} = \lim_{\delta y \to 0} \frac{F(y + \delta y) - F(y)}{\delta y} \times \frac{1}{S(y)} \quad .$$

Dado que a função densidade de probabilidade $f(y)$, ou seja, a taxa instantânea de morte no instante y, é descrita como:

$$f(y) = \lim_{\delta y \to 0} \frac{F(y + \delta y) - F(y)}{\delta y} \quad ,$$

obtemos:

$$h(y) = \frac{f(y)}{S(y)} \quad . \tag{3.3}$$

A distribuição acumulada, simbolicamente denotada como $H(y)$, ou função de risco integrada, é expressa por:

$$H(y) = -\log S(y) \quad , \tag{3.4}$$

e, igualmente, $S(y)$ pode também ser representada como $\exp[-H(y)]$.

3.3 Técnicas não-paramétricas: Kaplan-Meier e testes de hipótese

Não ocorrendo nenhuma censura de dado, a função de sobrevivência empírica pode ser simplesmente calculada pela razão entre o número de sujeitos que não foram acometidos por evento de falha até certo tempo e o número total de sujeitos observados. No entanto, havendo censura, o cálculo é prejudicado. **O estimador de Kaplan-Meier**, ou estimador produto-limite, foi desenvolvido com o objetivo de sanar tal problema no cálculo da função de sobrevivência empírica, permitindo a incorporação da censura.

De acordo com Lima Junior, Silveira e Ostermann, podemos entender o estimador de Kaplan-Meier "intuitivamente, [...] se levarmos em consideração que, para sobreviver a M intervalos de tempo, um indivíduo precisa ter sobrevivido a cada intervalo de tempo anterior" (LIMA JUNIOR; SILVEIRA; OSTERMANN, 2012). Através dele, é possível realizar comparações

entre funções de sobrevivência de dois grupos. Após a obtenção das funções de sobrevivência dos dois grupos, podemos ainda testar se a diferença é significativa, com testes de hipótese como o teste de Logrank, Tarone-Ware e Gehan, cuja hipótese nula é de igualdade das distribuições (COLOSIMO; GIOLO, 2006).

A curva de sobrevivência de Kaplan-Meier é definida como a probabilidade de sobrevivência em um determinado período de tempo. Tal probabilidade de sobrevivência, já definida anteriormente, é representada pela equação (3.2), onde $S(y)$ é a probabilidade de se observar um tempo de vida igual ou maior que y. No caso específico do estimador de Kaplan-Meier com dados censurados, podemos representar a função de probabilidade de sobrevivência como:

$$\hat{S}(y) = \prod_{j-1}^{k} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j-1}^{k} \left(1 - \frac{d_j}{n_j} \right) \quad , \tag{3.5}$$

onde n_j é o número de indivíduos sobreviventes até imediatamente antes do tempo y_j, e d_j é o número de falhas ocorridas no tempo infinitesimalmente pequeno entre $y_t - \delta$ e y_j, já que o tempo é contínuo. A probabilidade de sobreviver ao tempo y_j se apresenta como $(n_j - d_j)/n_j$ - ou, em outras palavras, a razão entre os sobreviventes ao tempo y_j e o total de indivíduos observados imediatamente antes do tempo y_j (DOBSON; BARNETT, 2018).

3.3.1 Comparação de curvas de sobrevivência

Uma das principais utilidades da análise de sobrevida é a possibilidade de comparação da distribuição de sobrevivência de dois grupos. Para isso, existem vários testes não-paramétricos que nos permitem averiguar as diferenças de distribuição (testar hipótese nula de que as amostras seguem a mesma distribuição). Neste trabalho, serão utilizados três dos testes mais utilizados: o teste de **Logrank**, o de **Gehan** e a classe de testes **Tarone-Ware**.

De acordo com Klein e Moeschberger (2003), a diferença entre os testes está na ponderação dos pesos ao longo do período observado: o teste de Logrank tem pesos iguais para qualquer momento; o teste de Gehan coloca mais peso em mortes mais prematuras, enquanto Tarone-Ware também coloca mais peso em mortes mais prematuras, porém um peso menor do que Gehan.

Colosimo e Giolo (2006) dão ênfase ao teste de Logrank, o mais utilizado na literatura, e seguem para a generalização dos testes de hipótese. Tomando o mesmo raciocínio, podemos partir pressupondo as curvas $S_1(y)$ e $S_2(y)$, onde $y_1 < y_2 < y_3 ... < y_n$ são os tempos de falha. A cada tempo y_j são observadas d_j falhas, e n_j representa o número de sujeitos em risco em um tempo imediatamente anterior a y_j na amostra combinada. Para uma amostra i, onde $i = 1, 2$ e $j = 1, ..., n$, temos d_{ij} e n_{ij}. Para cada tempo de falha, os resultados daquele tempo podem ser na Tabela 2 a seguir:

Tabela 2 – Tabela de contingência no tempo y_j

	Grupos		
	1	2	
Falha	d_{1j}	d_{2j}	d_j
Não Falha	n_{1j} - d_{1j}	n_{2j} - d_{2j}	n_j - d_j
	n_{1j}	n_{2j}	n_j

A distribuição de d_{2j} é uma hipergeométrica, $d2_j \sim \mathcal{H}$:

$$\frac{\binom{n_{1j}}{d_{1j}}\binom{n_{2j}}{d_{2j}}}{\binom{n_j}{d_j}} \quad ,$$

que tem como média $n_{2j}d_jn_j^{-1}$, simbolizado por ρ_{2j}, e significa dizer que:

> Se não houver diferença entre as duas populações no tempo y_j, o número total de falhas (d_j) pode ser dividido entre as duas amostras de acordo com a razão entre o número de indivíduos sob risco em cada amostra e o número total sob risco. (COLOSIMO; GIOLO, 2006, p.43)

A variância por sua vez - obtida da própria distribuição hipergeométrica - , é expressa por $(V_j)_2 = n_{2j}(n_j - n_{2j})d_j(n_j - d_j)n_j^{-2}(n_j - 1)^{-1}$.

Considerando que d_{2j} - ρ_{2j} apresenta média zero e variância $(V_j)_2$, se as n tabelas de contingência (dos n tempos y) foram independentes, Colosimo e Giolo (2006) desenvolvem como a estatística de teste aproximado para a hipótese nula de igualdade das duas funções de sobrevivência $S_1(y)$ e $S_2(y)$:

$$T = \frac{\left[\sum_{j=1}^{n}(d_{2j} - \rho_{2j})\right]^2}{\sum_{j=1}^{n}(V_j)_2} \quad . \tag{3.6}$$

Uma generalização da equação (3.6) pode ser obtida introduzindo efeitos de amplificação ou contração governada por uma distrbuição de pesos multiplicando a ordem de grandeza dos desvios $d - 2j - \rho - 2j$ na seguinte forma:

$$S = \frac{\left[\sum_{j=1}^{n}u_j(d_{2j} - \rho_{2j})\right]^2}{\sum_{j=1}^{n}u_j^2(V_j)_2} \quad , \tag{3.7}$$

onde u_j representa os pesos do teste. Como foi dito, a diferença entre os testes está justamente na ponderação dos pesos – aqui, u_j. Para o caso de Logrank, os pesos que são iguais para todos

os momentos da curva significam que u_j = 1, conforme visto na equação (3.6). No caso do teste de Gehan, $u_j = n_j$, colocando mais peso em mortes mais prematuras. Para a classe de testes de Tarone Ware, por fim, $u_j = \sqrt{n_j}$, que também coloca mais peso em mortes prematuras, mas não tanto quanto o teste de Gehan. A Tabela 3 sumariza as formas dos testes de hipótese, que apresentam distribuição Qui-quadrada com 1 grau de liberdade para amostras suficientemente grandes.

Tabela 3 – Formas dos testes de hipótese

Teste	Forma
Logrank	$S = \dfrac{\left[\sum_{j=1}^{n}(d_{2j-\rho_{2j}})\right]^2}{\sum_{j=1}^{n}(V_j)_2}$
Gehan	$S = \dfrac{\left[\sum_{j=1}^{n}n_j(d_{2j-\rho_{2j}})\right]^2}{\sum_{j=1}^{n}n_j^2(V_j)_2}$
Tarone-Ware	$S = \dfrac{\left[\sum_{j=1}^{n}\sqrt{n_j}(d_{2j-\rho_{2j}})\right]^2}{\sum_{j=1}^{n}\sqrt{n_j^2}(V_j)_2}$

3.4 Exemplos paramétricos

A literatura especializada apresenta exemplos de variáveis aleatórias utilizadas na modelagem de tempo de sobrevida. A mais simples delas é a **distribuição exponencial** definida a seguir (DOBSON; BARNETT, 2018):

$$f(y;\theta) = \theta e^{\theta y},\ para\ y \geq 0,\ \theta > 0 \quad , \tag{3.8}$$

onde θ representa parâmetro de decaimento ou taxa de decaimento. Assim, a função de distribuição acumulada, o complementar da função de sobrevivência, é:

$$F(y;\theta) = \int_0^y \theta e^{-\theta t} dt = 1 - e^{-\theta y} \quad . \tag{3.9}$$

Portanto, a função de sobrevivência é:

$$S(y;\theta) = e^{-\theta y} \quad , \tag{3.10}$$

cujas função densidade e acumulada de risco são respectivamente,

$$\begin{aligned} h(y;\theta) &= \theta \quad , \\ H(y;\theta) &= \theta y \quad . \end{aligned} \tag{3.11}$$

Como podemos observar na equação acima, a função de densidade de risco (3.11) é representada por uma constante; isso dizer significa que o risco de falha não depende do quanto o sujeito já sobreviveu, o que é geralmente limitante, já que na maioria dos casos, a chance de falha aumenta conforme o passar do tempo.

Considerando isso, outras distribuições, que apresentam função de risco com memória temporal podem ser mais adequadas para a maioria dos casos. A **distribuição de Weibull**, que figura como a mais utilizada na análise paramétrica de sobrevida, apresenta tal característica, cuja função de densidade é expressa como:

$$f(y;\lambda,\theta) = \frac{\lambda y^{-1}}{\theta^{\lambda}} \exp\left[-\left(\frac{y}{\theta}\right)^{\lambda}\right]\ para\ y \geq 0 \quad , \tag{3.12}$$

onde λ e θ são os parâmetros de forma e escala da distribuição. Outra forma de escrever a distribuição de Weibull é aplicando a reparametrização $\varphi = \theta^{-\lambda}$, de forma que adquire a forma:

$$f(y;\lambda,\varphi) = \lambda\varphi y^{\lambda-1} \exp(-\varphi y^{\lambda}) \quad , \tag{3.13}$$

com isso, a função de sobrevivência é dada por:

$$S(y;\lambda,\varphi) = \exp(-\varphi y^{\lambda}) \quad , \tag{3.14}$$

e as funções de risco e risco acumulada são dadas por, respectivamente:

$$\begin{aligned} h(y;\lambda,\varphi) &= \varphi y^{\lambda-1} \quad , \\ H(y;\lambda,\varphi) &= \varphi y^{\lambda} \quad . \end{aligned} \tag{3.15}$$

Uma propriedade importante da distribuição de Weibull consiste em ser a base para a construção tanto de modelos de riscos proporcionais quanto de modelos de tempo acelerado. Os primeiros, e mais populares na análise de sobrevivência, implicam que as funções de sobrevivência para indivíduos de dois grupos são paralelas ao longo de todos os períodos, e um cruzamento nas curvas de sobrevivência ou variância relevante nas distâncias significa ausência de proporcionalidade. Já modelos de tempo de vida acelerado, uma alternativa aos riscos proporcionais, não assumem riscos constantes, e podem ser usados como saídas em casos de ser observada a violação dos riscos proporcionais. É válido mencionar que existem outras distribuições razoavelmente utilizadas pela literatura, como a Distribuição Gama (DOBSON; BARNETT, 2018), mas, por serem menos comuns do que a distribuição de Weibull, não serão abordadas.

3.4.1 Estimação dos parâmetros

Como apresentado na seção anterior, a lei probabilística do tempo de sobrevida é governado por um conjunto de parâmetros, que devem ser estimados. Para a estimação de tais parâmetros das funções de sobrevivência, o Método dos Quadrados Ordinários não é apropriado, dado que não é possível considerar censura na estimação. Desse modo, é necessário o uso do Método de Máxima Verossimilhança, cujo objetivo é encontrar os valores que maximizam a probabilidade ou densidade dos dados observados.

Segundo Dobson e Barnett (2018), para um j_n sujeito, temos o tempo de sobrevivência y_j, uma variável δ_j que indica se o tempo de sobrevivência é cesurado (0) ou não censurado (1), e um vetor x_j de regressores. Por existirem dados censurados e não censurados, a função de verossimilhança abaixo é composta por duas partes, uma que é contribuição das variáveis não censuradas $\left(\prod_{j=1}^{r} f(y_j)\right)$ – a função de densidade de cada observação não censurada –, e outra, das variáveis censuradas $\left(\prod_{j=r+1}^{n} S(y_j)\right)$ – a função de sobrevivência de cada observação censurada. Juntando as duas partes, temos:

$$L = z\prod_{j=1}^{n} f(y_j)^{\delta_j} S(y_j)^{1-\delta_j} \quad . \tag{3.16}$$

A função de densidade (contribuição dos não censurados) aponta um produtório de 1 a r, e a função de sobrevivência (contribuição dos censurados), apontam um produtório de $r+1$ a n, indicando a ordenação das observações. Assim sendo, uma vez aplicado a função logarítmica e propriedades algébricas, obtemos a função de log-verossimilhança:

$$l = \sum_{j=1}^{n} [\delta_j \log h(y_i) + \log S(y_j)] \quad . \tag{3.17}$$

3.5 Modelos de regressão semi-paramétricos e paramétricos

Modelos de regressão são fundamentais na análise de sobrevivência para permitir a incorporação de diferentes características que afetam fatores de risco numa população. Aqui serão apontados dois tipos de modelos, os mais utilizados, e já mencionados anteriormente: modelos com funções de riscos proporcionais e modelos de tempo de vida acelerado, com ênfase no primeiro, que será aplicado empíricamente neste trabalho. Ambos podem ser tanto paramétricos quanto semi-paramétricos, fazendo as devidas modificações em termos de parâmetros, e ambos são modelos lineares.

3.5.1 Modelos de funções com riscos proporcionais

Partindo das funções de risco para sujeitos em um suposto grupo de tratamento e outro suposto grupo de controle – ou quaisquer comparação entre dois grupos que possa ser realizada –, respectivamente $h_1(y)$ e $h_0(y)$, no tempo y, e respeitando o pressuposto de riscos proporcionais em todo tempo, a função de risco do modelo e de sobrevivência são, respectivamente:

$$\begin{aligned} h_1(y) &= \phi h_0(y) \quad , \\ S(y) &= [S_0(y)]^{\phi} \quad , \end{aligned} \tag{3.18}$$

onde ϕ é a taxa de risco, um valor constante. Se $\phi < 1$, o risco de ocorrência do evento de falha no momento y é menor para o sujeito em grupo de tratamento do que comparado ao grupo de controle, o resultado contrátio obtemos se $\phi > 1$. Por exemplo, se $\phi = 2$, quer dizer que o risco de falha (em y) é dobro para o grupo de tratamento em relação ao grupo de controle em todos os momentos observados. Além disso, dado que ϕ não pode ser negativo, $\phi = e^{\beta}$, onde β é log da taxa de risco – e qualquer valor de β acarretará num ϕ positivo. Assim:

$$h_i(y) = h_0(y) e^{\beta x_i} = h_0(y) e^{\beta_1 x_{1i} + \ldots + \beta_n x_{ni}} \quad , \tag{3.19}$$

onde x_i representa os componentes de um vetor de condicionantes simbolicamente denotado com x.

Naturalmente, o pressuposto de paralelidade das funções de sobrevivência (riscos propocionais) deve ser cumprido, podendo acarretar em vícios na estimação dos coeficientes do

modelo caso não seja cumprido. Para modelos com poucos regressores, é razoável observar se as linhas dos gráficos do log da função de risco acumulada são paralelos. Para situações mais complexas, com mais regressores, testes de resíduos e testes específicos de checagem de riscos proporcionais são válidos.

Dentre a classe de modelos de riscos proporcionais, há a possibilidade de definição ou não da distribuição do tempo de sobrevivência - ou seja, podem ser tanto paramétricos quanto semi-paramétricos; porém, por sua versatilidade, o exemplo mais popular de modelo de riscos proporcionais é o já mencionado **modelo de Cox**, semi-paramétrico, abordado a seguir.

3.5.1.1 Modelo de Cox

A regressão de Cox é um método semi-paramétrico publicado em 1972 por D. R. Cox, do Imperial College de Londres. A interpretação dos coeficientes se dá pela razão de taxas de falha, ou risco relativo. Por ser um modelo mais flexível , se tornou extremamente popular, sendo hoje o modelo mais utilizado na área de análise de sobrevivência.

Como abordado anteriormente, o modelo de riscos proporcionais, em sua forma mais simples, tem como pressuposto taxas de falhas proporcionais - ou seja, o risco de falha das variáveis é constante ao longo do tempo. Suas funções de risco (e generalização para n covariantes) e de sobrevivência podem ser observadas nas equações (3.18) e (3.19).

É válido mencionar também a existência de extensões do modelo de Cox, como o modelo de taxas de falhas proporcionais estratificado, que foi proposto para contornar situações onde a condição de taxas proporcionais não é cumprida para uma covariável categórica do modelo. Nesse caso, o modelo de riscos proporcionais estratificado supõe que os riscos devem ser paralelos em cada estrato mas não entre estratos (COLOSIMO; GIOLO, 2006).

3.5.2 Modelos de tempo de vida acelerado (AFT)

Em modelos de tempo de vida acelerado, a variável aleatória Y é dada por $Y = Y_0/\varphi$, e as funções de risco para sujeitos no grupo de tratamento e de controle $h_1(y)$ e $h_0(y)$, respectivamente, no tempo y, as funções de risco é (BORGES, 2014):

$$h_1(y) = \varphi h_0(y\varphi) \quad , \tag{3.20}$$

onde φ é a taxa de risco. A função de sobrevivência é expressa por:

$$S(y) = S_0(y\varphi) \quad . \tag{3.21}$$

Como é possível observar nas equações, há impacto multiplicativo (acelerado ou desacelerado) no tempo.

3.5.3 Adequação do modelo

É amplamente conhecido que na modelagem econométrica a avaliação dos residuo é utilizada na availação dos modelos postulados, e a análise de sobrevida não é uma exceção. Dobson e Barnett (2018) discorrem sobre diferentes formas de avaliação, e de acordo com as autoras, os resíduos mais simples para a análise de sobrevivência são os **resíduos de Cox-Snell** (r_C), que auxiliam na avaliação global, e são definidos como:

$$r_{Cj} = \hat{H}_j(y_j) = -\log[\hat{S}_j(y_j)] \quad , \tag{3.22}$$

onde j representa um indívíduo não censurado e $\hat{H}_j$ e $\hat{S}_j$ são as funções de risco acumulada e sobrevivência acumulada estimadas para o indivíduo j no tempo y_j. Para sujeitos censurados, r_{Cj} é naturalmente muito pequeno; por essa razão, é proposto um ajuste dos resíduos dos dados censurados: $r'_{Cj} = r_{Cj} + \Delta$, onde $\Delta = 1$ ou $\Delta = \log 2$ (DOBSON; BARNETT, 2018).

Outros testes relevantes são os **resíduos de Martingale** (r_M), que é basicamente uma transformação linear dos resíduos de Cox-Snell, e os **resíduos deviance** (r_D), que, por sua vez, é uma transformação dos resíduos de Martingale. Eles são, respectivamente, para o sujeito j:

$$r_{Mj} = \delta_j - r_{Cj}, \quad \text{com} \quad \begin{cases} \delta_j = 1 & \text{para dados não censurados} \\ \delta_j = 0 & \text{para dados censurados} \end{cases} \quad , \tag{3.23}$$

e

$$r_{Dj} = sgn(r_{Mj})\{-2[r_{Mj} + \delta_j \log(r_{Cj})]\}^{1/2} \quad . \tag{3.24}$$

Para os resíduos de Martingale, os valores esperados são zero, com uma distribuição com inclinação negativa (*negatively skewed*, ou *left-skewed*). Para os resíduos deviance, é esperado que os resíduos tenham um comportamento aleatório, distribuídos ao redor de zero, com valores grandes provavelmente indicando outliers.

4 ANÁLISE DE DADOS

Considerando o desenvolvimente teórico apresentado nos capítulos anteriores, neste capítulo apresentamos os resultados obtidos na representação do evasão no ensino superior da Universidade Federal Fluminense. Nesse sentido, na primeira parte explicamos a base de dados utilizada e principais variáveis, e seguidamente mostramos as estatísticas descritivas mais importantes. Na terceira seção dedicamos espaço aos resultados da análise de sobrevida no contexto não-paramétrico e na ultima seção relatamos as estimativas no caso da representação semi-parametrica.

4.1 Base de dados

Para este trabalho, a base de dados teve como fonte os microdados do Censo da Educação Superior, de 2010 a 2017, disponibilizados publicamente pelo INEP, entidade do Ministério da Educação (INEP, 2018). O censo conta com informações sobre instituição, curso, pólo do curso, turno, modalidade, sexo, cor ou raça, idade, esfera de escola de origem (pública ou privada) e situação de matrícula, dentre outros, à nível de alunos, que são passíveis de rastreamento através de um código identificador único (ID).

Com tal base de dados, podemos compreender o perfil dos estudantes do ano de 2010, a distribuição de cada característica, e por fim, utilizá-los como base para a estimação de métodos não-paramétricos e semi-paramétricos posteriormente.

Os dados foram tratados afim de se restringirem somente aos alunos da Universidade Federal Fluminense, conforme já dito, por sua diversidade discente e por ser a instituição federal com maior número de matrículas anuais, atualmente. Dado que o objetivo é realizar o acompanhamento dos estudantes (através do ID único), foram escolhidos somente os alunos ingressantes em 2010 (10.106 indivíduos), e filtrados nas bases de 2011 a 2017 seus respectivos IDS para o acompanhamento dos respectivos status.

Após o filtro dos alunos ingressantes em 2010, as bases foram empilhadas ano após ano, e observamos a quantidade de anos (períodos) que se pode acompanhar cada um dos IDS. Tal quantidade de anos foi transformada numa variável, que contabiliza o tempo de permanência do estudante na instituição. Ao fim do tratamento dos dados, obtemos uma base com 10.106 registros, variáveis invariantes no tempo (*time-invariant*), a variável de tempo acompanhado e uma variável com a última situação observada daquele estudante (se o estudante foi observado por 5 anos e a última situação observada dele foi matrícula trancada, na variável constará a situação *matrícula trancada*).

Como apresentado anteriormente, a definição do evento de falha é um componente extremamente relevante na análise de sobrevida. Para nosso estudo, tal evento corresponde

à evasão escolar. No Censo da Educação Superior, base de dados fonte do projeto, existem múltiplas categorias de status do estudante, e após avalição de tais categorias, determinamos que o evento de evasão seria observado para os casos de estudantes nas condições de "desvinculado ao curso" ou "transferido para outro curso da mesma IES". Os casos censurados, por sua vez são aqueles nas condições de "cursando" ou "matrícula trancada", dado que esses indivíduos eventualmente poderão incorrer no evento terminal ou não.

4.2 Estatísticas descritivas

Para a compreensão das estatísticas, usamos o dataset tratado que apresenta os 10.106 registros e as variáveis criadas de tempo observado e situação do aluno, mencionadas anteriormente. Sugerimos a leitura do Apêndice A, que contém detalhes de como foi realizado tanto o processamento dos microdados do Censo, quanto os comandos utilizados para geração das estatísticas descritivas e seus respectivos gráficos, além dos procedimentos de análise de sobrevida, cuja aplicação será apresentada posteriormente.

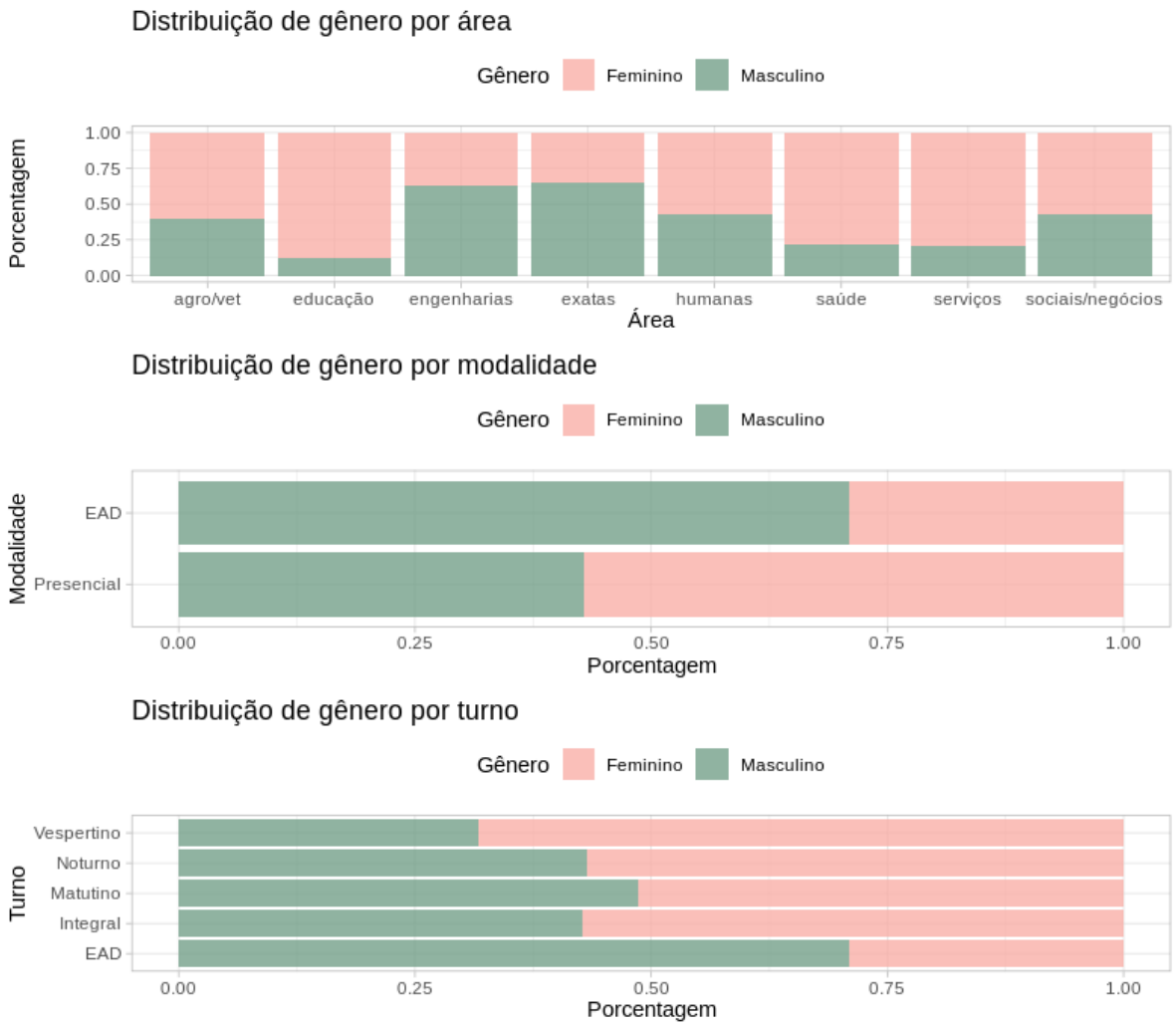
Em termos descritivos, a base tratada conta com 5.168 mulheres (51,12%) e 4.938 homens (48,88%). A Figura 1 apresenta essa distribuição área de conhecimento[1], modalidade e turno. A partir dela, podemos observar que mulheres são ligeiramente mais presentes em cursos presenciais, enquanto homens compõe quase 3/4 dos estudantes dos cursos à distância. Essa disparidade nos cursos de EAD está relacionada ao fato de que, à época, a maioria dos cursos à distância oferecidos eram da área de exatas, que, nos dados observados, apresenta cerca de 70% dos estudantes homens.

Dos 10.106 estudantes, 39,72% não possui cor ou raça declarada, 39,55% são brancos, 5,20% pretos, 13,98% pardos, 1,06% amarelos e 0,48% indígenas. Discriminando por modalidade (presencial ou à distância), pessoas sem cor ou raça declarada representam mais de 95% dos estudantes de cursos à distância e apenas 24% dos estudantes de cursos presenciais. Por esse motivo, as estatísticas de cor ou raça serão observadas na modalidade presencial.

Através da Figura 2, podemos observar que os cursos das áreas de agronomia e veterinária, engenharias e serviços (basicamente, o curso de turismo) possuem as maiores taxas de alunos brancos (cerca de 60% nas áreas mencionadas). As áreas com maior número de pardos e pretos, em termos de proporção, são as áreas de educação e saúde. No último gráfico da Figura 2, é trazida a relação entre cor ou raça e evasão, sob a luz das diferentes áreas: nos cursos de engenharia, por exemplo, percebemos (vide linha vermelha auxiliar) que a proporção de brancos que evadem é menor do que a proporção geral de brancos que entram para os cursos de engenharias (9 pontos percentuais menor), enquanto a proporção de brancos que não evade é ligeiramente maior do que a geral (2 pontos percentuais maior).

[1] A agregação por área do conhecimento foi feita a partir da classificação internacional da OCDE, disponível em <http://download.inep.gov.br/download/superior/2009/Tabela_OCDE_2009.pdf>.

Figura 1 – Distribuição de gênero por área, modalidade e turno

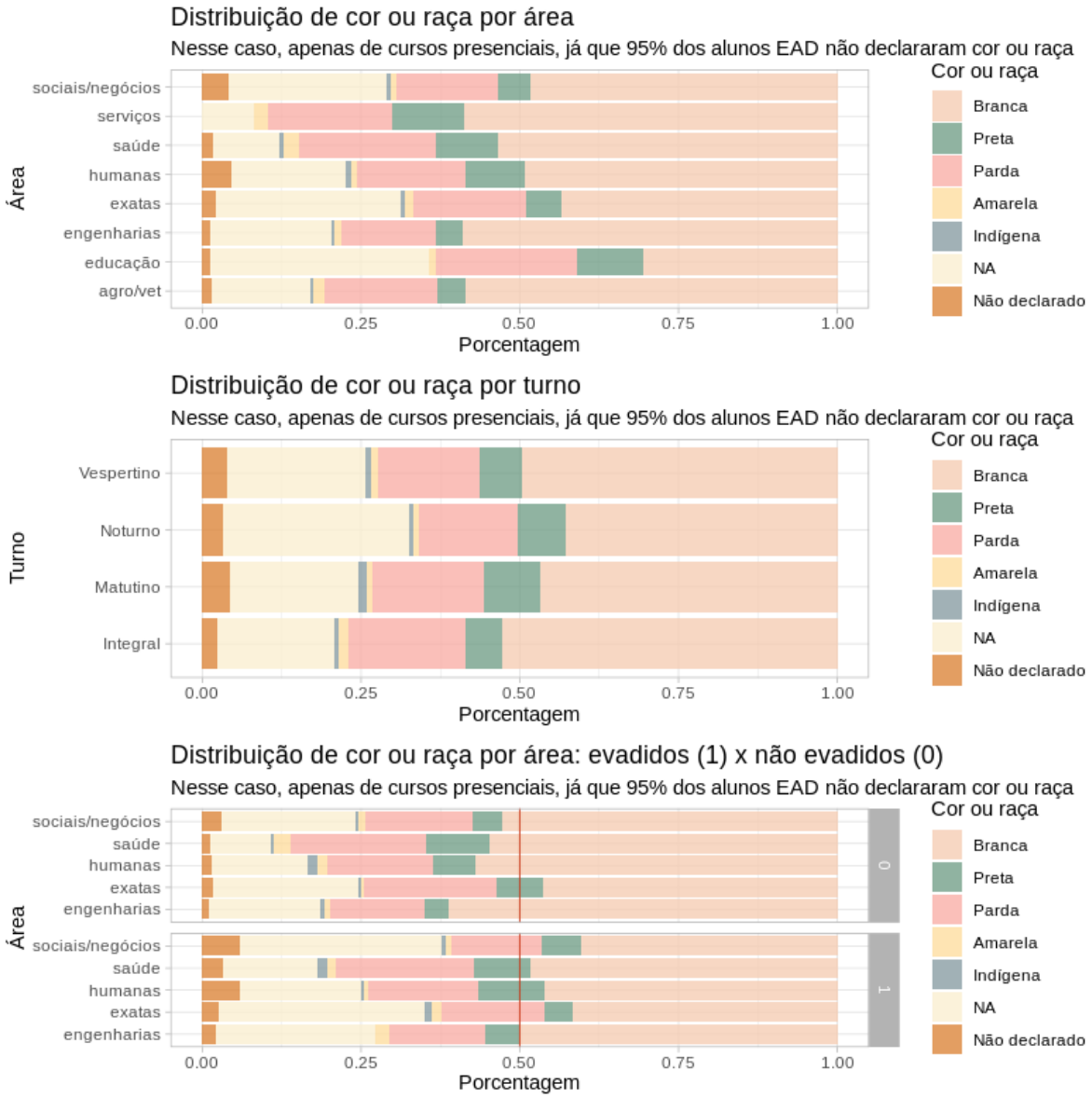


Os alunos brancos que evadem as áreas de educação, serviços e agronomia e veterinária são pouquíssimo numerosos para serem tiradas maiores conclusões, por isso não foram incluídos no gráfico que considera evasão.

Em termos de idade, a média dos ingressantes em 2010 foi de 23,4 anos, com mediana de 20 anos. Para as áreas de saúde e engenharias, observamos um pico mais acentuado entre 19 e 20 anos, enquanto em exatas a distribuição é mais suave, sem picos aparententes, conforme a Figura 3. A distribuição de idade mais achatada na área de exatas se dá pelo fato de que a maioria dos cursos de EAD são nessa área, e conforme o gráfico de baixo da Figura 3, a curva para cursos à distância é significativamente diferente da curva de cursos presenciais, sendo bem mais achatada. Observando apenas os estudantes de exatas que cursam presencialmente, a curva seria similar à das áreas de engenharias e saúde, com um pico por volta dos 19 anos.

Em relação às modalidades de ensino, 78,68% dos estudantes se matricularam em cursos presenciais, e 21,32% em cursos à distância. Dentre os que se matricularam em cursos presenciais, 62,71% foram em cursos integrais, 29,52% em cursos noturnos, 2,50% em cursos

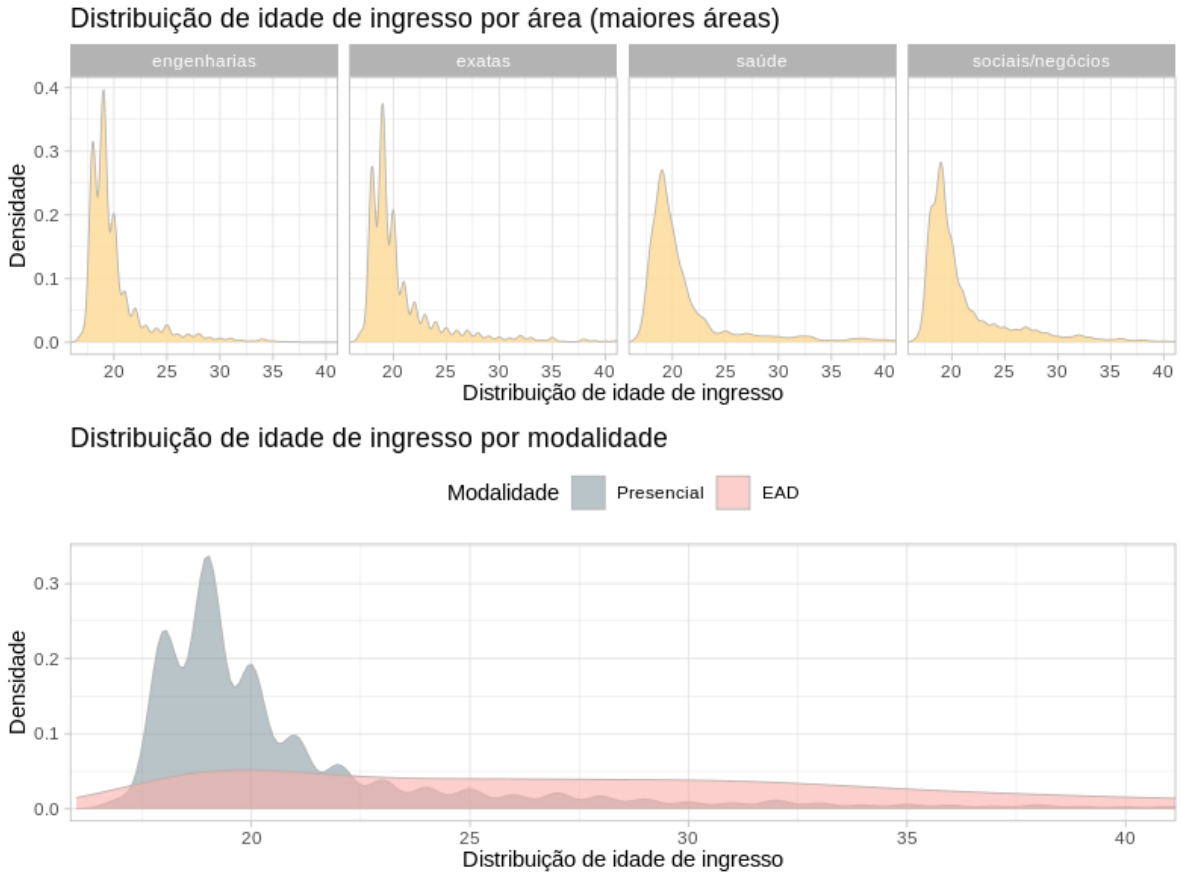
Figura 2 – Distribuição de cor ou raça por área e turno



vespertinos e 5,27% em matutinos. Na Figura 4, observamos que os cursos EAD representam a totalidade dos cursos tecnológicos e mais da metade dos cursos de licenciatura, enquanto, nas graduações com habilitação para bacharelado, a sua presença é ínfima. A diferença de proporção dos grupos evadidos e não evadidos, observando grau *versus* modalidade não é visualmente relevante. Por fim, em termos de perfil, a Figura 1 aponta que os estudantes de cursos à distância são majoritariamente homens, e Figura 3 não indica um pico aparente de idade de ingresso, sendo a curva bem mais achatada, com o primeiro quartil em 21 anos, o segundo quartil (mediana) em 28 anos, e o terceiro quartil em 35 anos; comparativamente, para os alunos de cursos presenciais, esses valores são, respectivamente, 19, 20 e 22 anos.

Quanto à habilitação, 14% da base dos 10.106 alunos não foi corretamente preenchida, e o grau não foi especificado, mas dentre os 85% restantes, 66,69% são matrículas relativas à cursos de bacharelado, 22,36% são cursos com habilitação para licenciatura e 10,95% são tecnólogos.

Figura 3 – Distribuição de idade de ingresso por área e modalidade



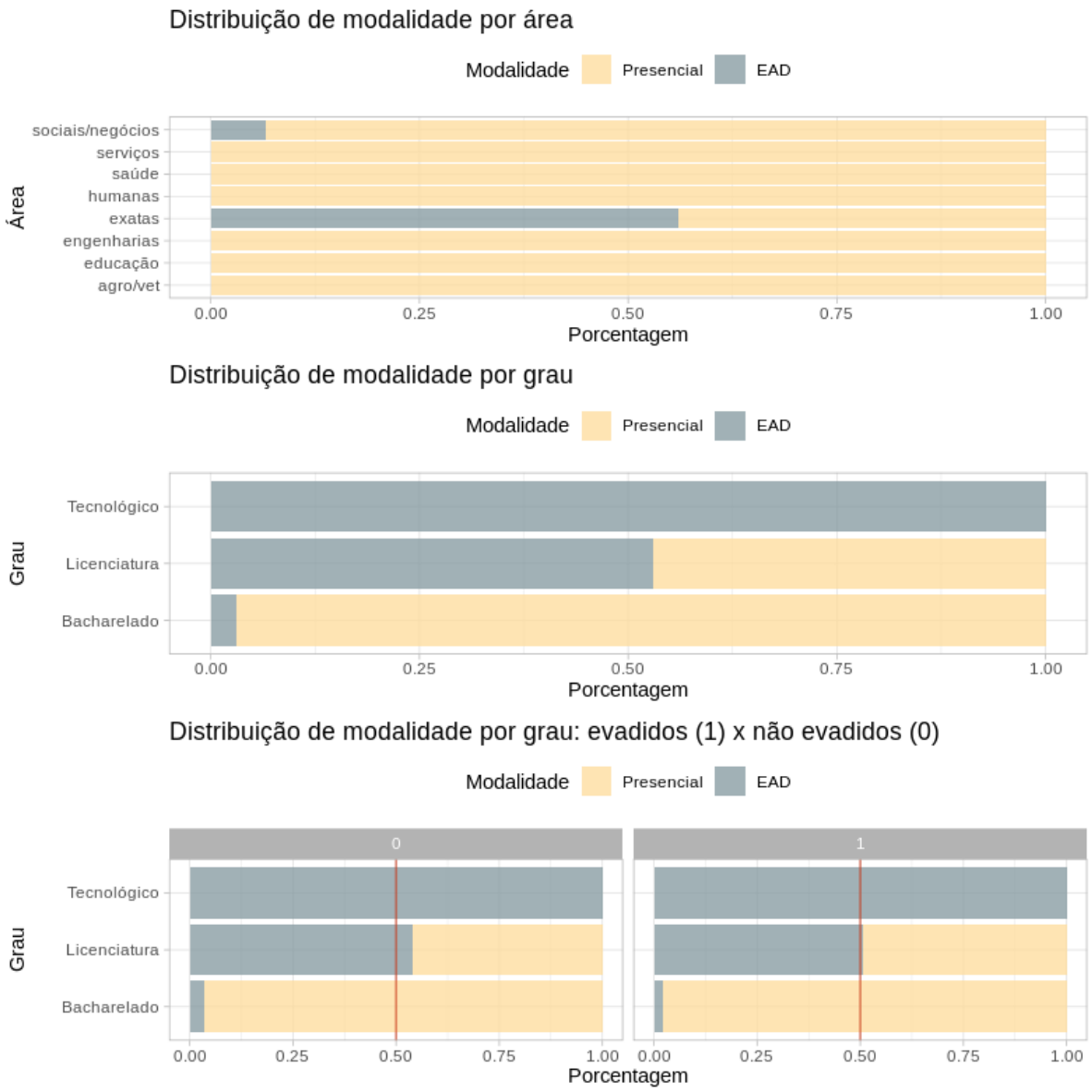
Como mencionado anteriormente, os cursos tecnólogo tem seus alunos completamente EAD, licenciatura tem mais da metade deles, e bacharelado, somente 3% dos alunos em EAD.

Na Figura 5 é possível observar a distribuição dos diferentes tipos de graus nas áreas de conhecimento - por exemplo, as engenharias, serviços, ciências sociais e negócios e agricultura e veterinária são compostos inteiramente por cursos de bacharelado, enquanto as humanas são majoritamente licenciatura, saúde possui algumas poucas observações de licenciatura (do curso de educação física), e exatas apresenta uma diversidade maior.

O segundo gráfico da Figura 5 mostra a distribuição de idade por grau, e é notável como nos cursos de bacharelado há um pico por volta dos 19 anos, enquanto os cursos de licenciatura e tecnólogos tem curvas muito mais achatadas. As medianas são, respectivamente, 19, 23 e 26 anos, com o terceiro quartil em 22, 32 e 34 anos.

Por fim, 63,1% dos ingressantes entraram via vestibular da universidade (de acordo com a variável dummy de ingresso via vestibular), e 8,27% via ENEM (de acordo com a variável dummy de ingresso via ENEM), o que é razoável, já que em 2010 o exame nacional não era obrigatório. No entanto, discriminando a entrada por vestibular próprio contra a modalidade do curso (Figura 6), observamos que os cursos EAD (21% da base de estudantes) tiveram seus alunos ingressados via outros meios que não o vestibular próprio ou o ENEM. No grupo dos

Figura 4 – Distribuição de modalidade por área e grau



estudantes presenciais, 20% não ingressou via vestibular próprio, sendo 10% ingressante via ENEM, e o restante, vindos de outros meios.

Isso justifica, por exemplo, a disparidade da proporção de ingressantes via vestibular próprio na área de exatas em relação às demais áreas - como foi mencionado anteriormente (Figura 4), a área de exatas tem mais da metade dos estudantes em cursos à distância.

É importante mencionar que algumas variáveis, como reserva de vagas[2] (étnicas/raciais, por renda, etc), escola de ensino médio de origem, recebimento de auxílio permanência, realização de iniciação científica ou estágio, dentre outras, não puderam ser aproveitadas pela falta de preenchimento ou mau preenchimento nos censos do Ensino Superior.

[2] A Lei das Cotas só foi aprovada em 2012, de modo que somente a partir do ano de ingresso de 2013 houve a reserva de vagas.

Figura 5 – Distribuição de grau por área e distribuição de idade de ingresso por grau

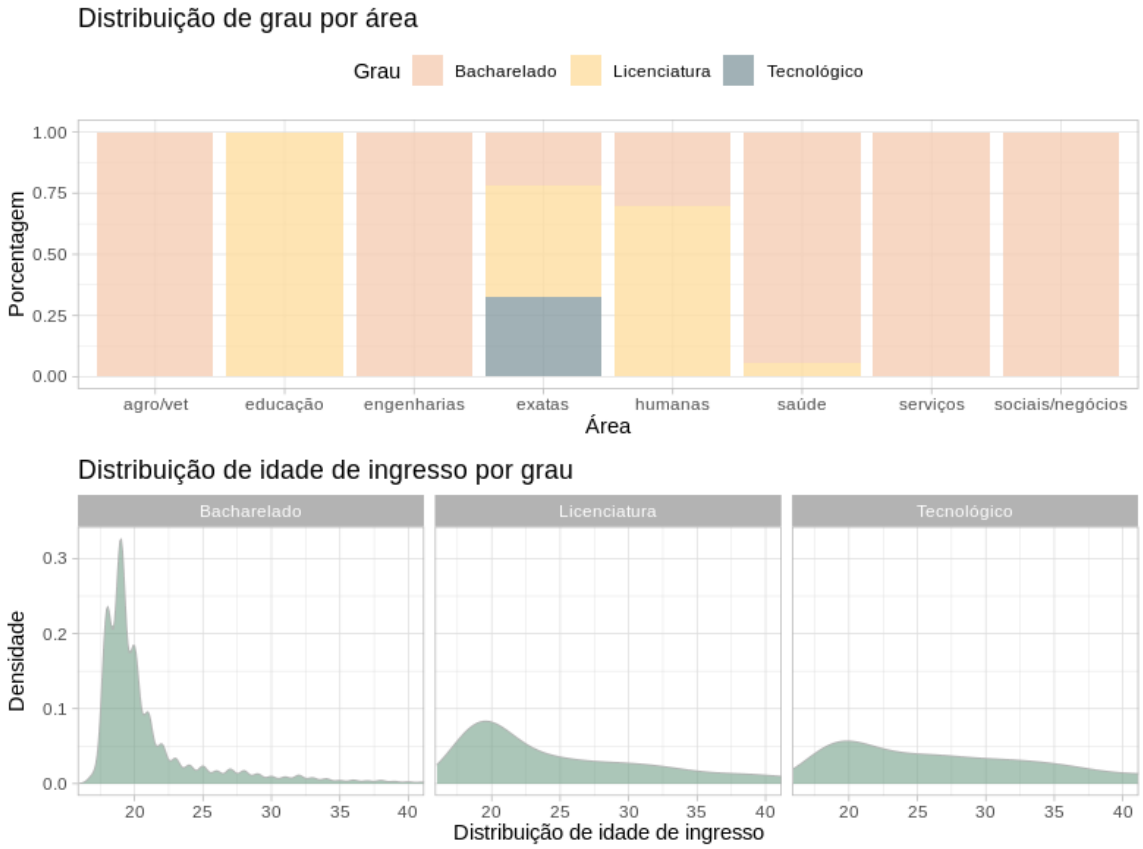
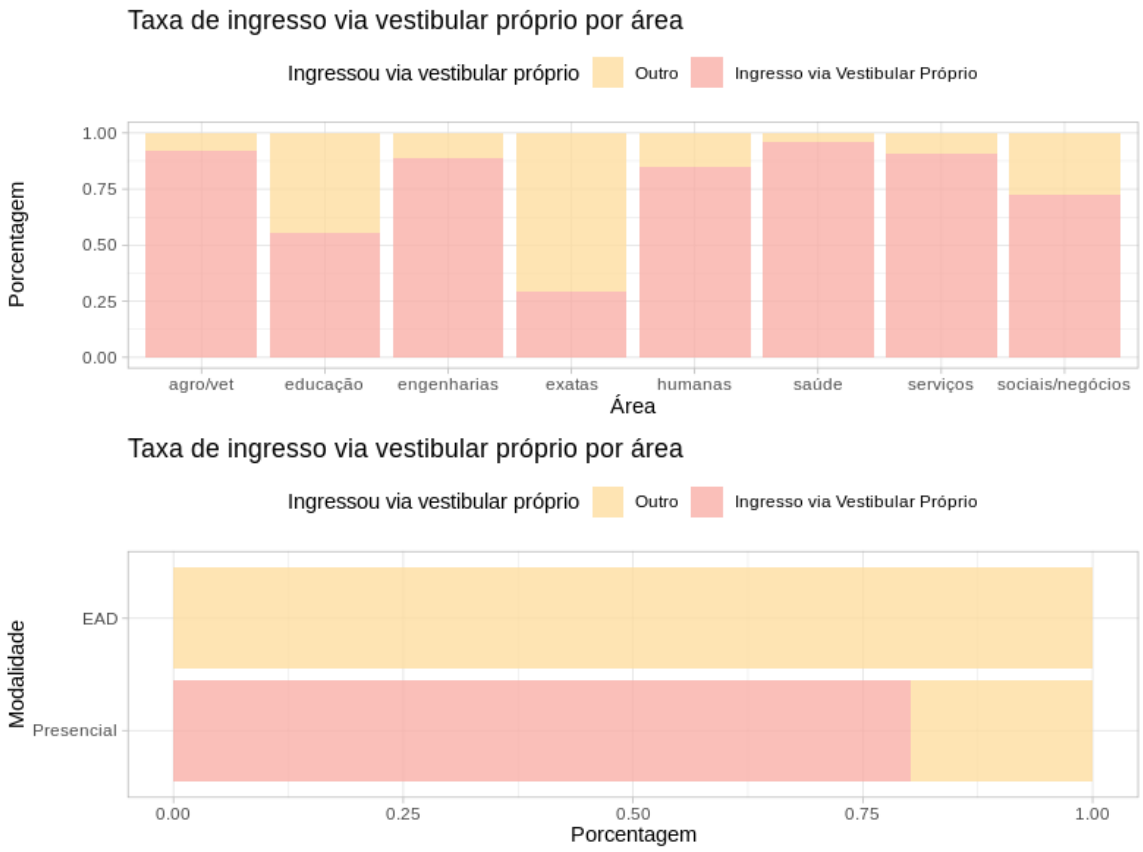


Figura 6 – Distribuição de modalidade por área e modalidade



4.3 Estimador de Kaplan-Meier

Nesta seção, apresentamos os resultados da aplicação do método de Kaplan-Meier e os testes de hipótese utilizados, para os dados relatados na seção anterior. Os comandos necessários para realização das estimações e dos respectivos testes estão especificados no Apêndice A.

Como indicado no Capítulo 3, o método de Kaplan-Meier é útil quando desejamos comparar o comportamento da evasão entre grupos. Assim, com base na análise descritiva optamos por considerar os agrupamentos por sexo, modalidade do curso, cor ou raça declarada pelo aluno e tempo do curso como mostramos na Tabela 4. O valor 0 (zero) representa o grupo de controle.

Tabela 4 – Variáveis utilizadas para estimação de Kaplan-Meier

Variável	Descrição
modalidade	Os valores são Presencial ou Curso à Distância
sexo	Os valores são Feminino e Masculino
bacharelado	Os valores são Bacharelado (1) e Não-Bacharelado (0)
branco	Os valores são Branco (1) e Não-Branco (0)
integral	Os valores são Integral (1) e Não-Integral (0)

4.3.1 Estimação das curvas de sobrevivência empíricas e testes

Como mencionamos na seção 3.3.1, os testes de Tarone-Ware, Logrank e Gehan são comumente utilizados e fornecem informações do comportamento das estimativas ao longo da distribuição. A partir de tais testes, podemos entender se as distribuições das curvas que estão sendo comparadas são estatisticamente distintas – ou seja, se as sobrevivências são realmente diferentes.

A Tabela 5 abaixo apresenta os p-valores dos testes de hipótese realizados para cada uma das variáveis apresentadas na primeira coluna da tabela. É possível observar que a variável "modalidade" é significativa ao nível de 1% de nível de significância para todos os três testes realizados, bem como as variáveis binárias de turno integral, e bacharelado.

A variável sexo é apenas significativa nos testes de Tarone-Ware a 10% de significância e Logrank a 5% de significância – para o teste de Gehan, o p-valor é ligeiramente acima de 15%, de modo que não podemos rejeitar a hipótese nula. Por fim, a variável binária que indica que o estudante se autodeclarou branco é significativa a 5% de nível de significância para os testes de Gehan e Tarone-Ware e a 10% para o teste de Logrank. Todos os testes foram calculados usando a função `coin::logrank_test()`, do software `R` (R CORE TEAM, 2020).

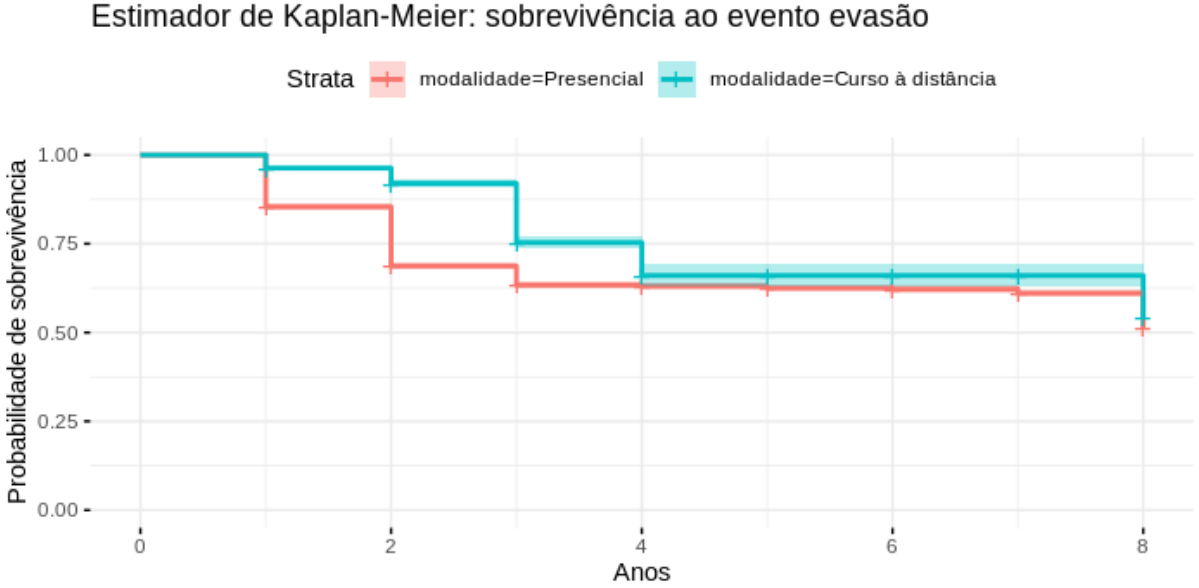
Partindo da variável de modalidade, a Figura 7 – gerada utilizando os pacotes `survival` e `survminer` (uma expansão do pacote `ggplot2`) do software `R`, conforme indicado no Apêndice A – mostra que evasão nos cursos presenciais é muito maior do que dos cursos à

Tabela 5 – Testes de Gehan, Tarone-Ware e Logrank

Variável	Gehan	Tarone-Ware	Logrank
modalidade	0,0000	0,0000	0,0000
sexo	0,1501	0,0850	0,03476
bacharelado	0,0000	0,0000	0,0000
branco	0,0232	0,0276	0,0654
integral	0,0000	0,0000	0,0000

distância nos 4 primeiros anos, chegando a um gap de mais de 20 pontos percentuais; no terceiro ano, a evasão de cursos presenciais praticamente se estabiliza, e fica apenas ligeiramente maior do que a dos cursos de EAD do quarto ano em diante.

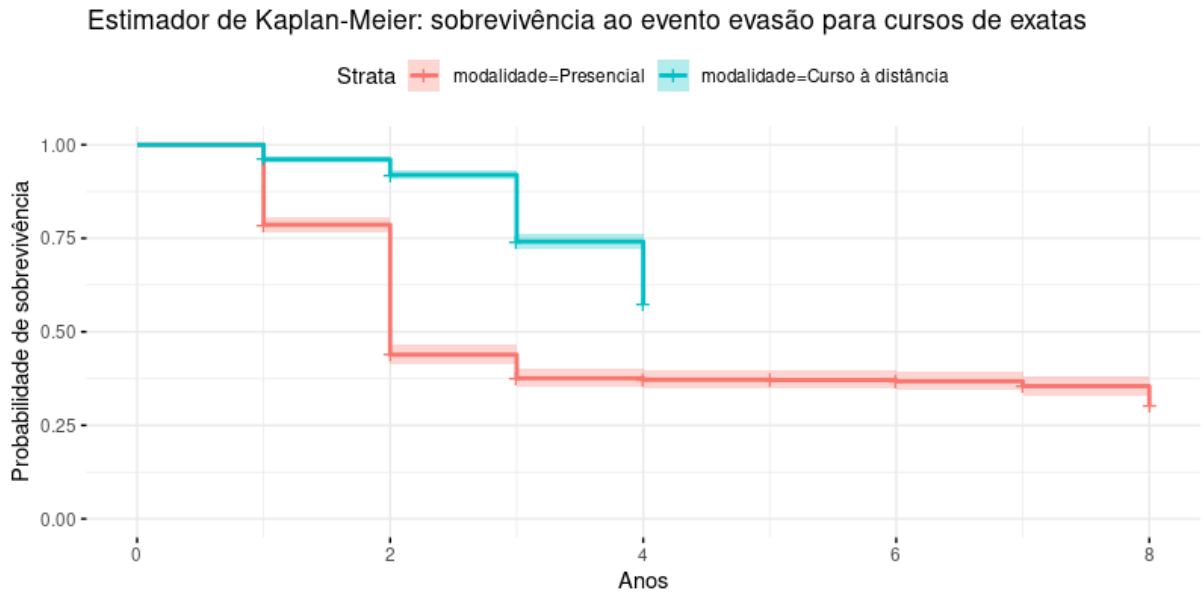
Figura 7 – Estimador de Kaplan-Meier: variável modalidade



Entretanto, devemos considerar que a maior parte dos estudantes inscritos em cursos à distância são de cursos de exatas – mais precisamente, 91,6% dos estudantes. Para que as curvas sejam realmente comparáveis, o Gráfico 8 apresenta a mesma lógica do Gráfico 7, porém considera apenas cursos de exatas. A partir dele, podemos verificar que o nível de evasão dos cursos presenciais de exatas é ainda maior do que o nível de evasão nos cursos presenciais de modo geral, atingindo o cerca de 63% no quarto ano observado. Dentre os cursos à distância de exatas, no mesmo quarto ano, a evasão atinge cerca de 45%. Ainda, é válido mencionar que a curva de EAD é terminada abruptamente no quarto ano – isso porque, de acordo com os registros, nenhum dos 1974 estudantes de cursos de exatas se formou ou evadiu após o ano quatro.

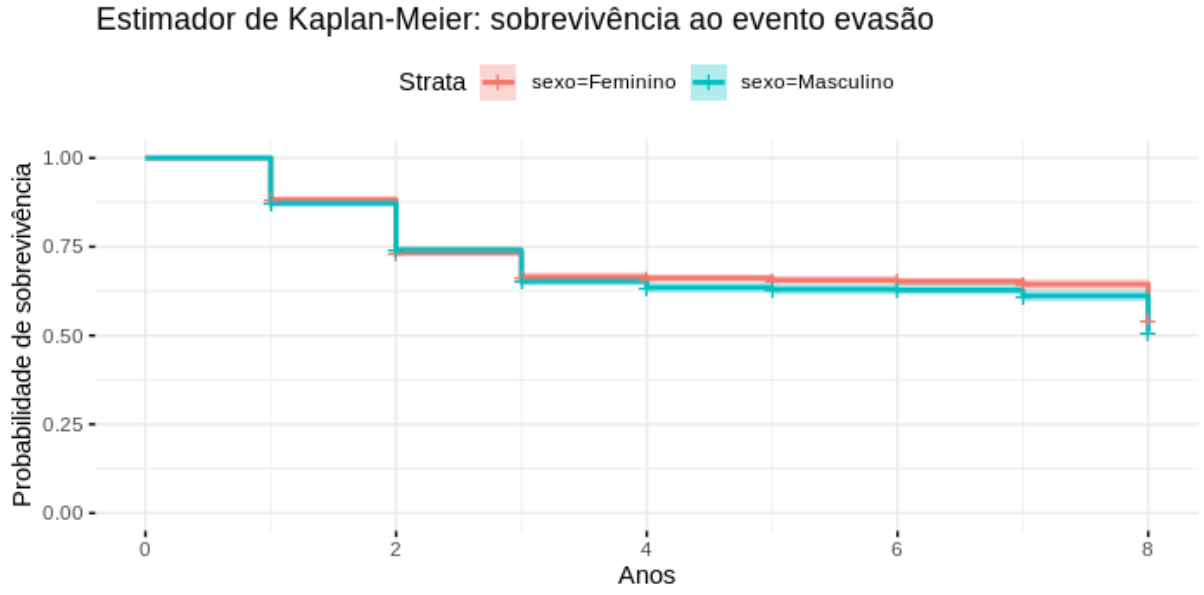
A sobrevivência ao evento evasão agrupada por sexo mostra pouca diferença entre as curvas, ainda que os testes de significância apontem alguma significância (no caso, no teste de Tarone-Ware, que enfatiza os eventos prematuros - ainda que menos que o teste de Gehan -, e no de Logrank, dá igual peso para todas as fases da curva). A Figura 9, apresenta visualmente

Figura 8 – Estimador de Kaplan-Meier para cursos de exatas: variável modalidade



o motivo pelo qual o teste de Gehan, que dá maior peso a eventos prematuros, não apontou diferença significativa.

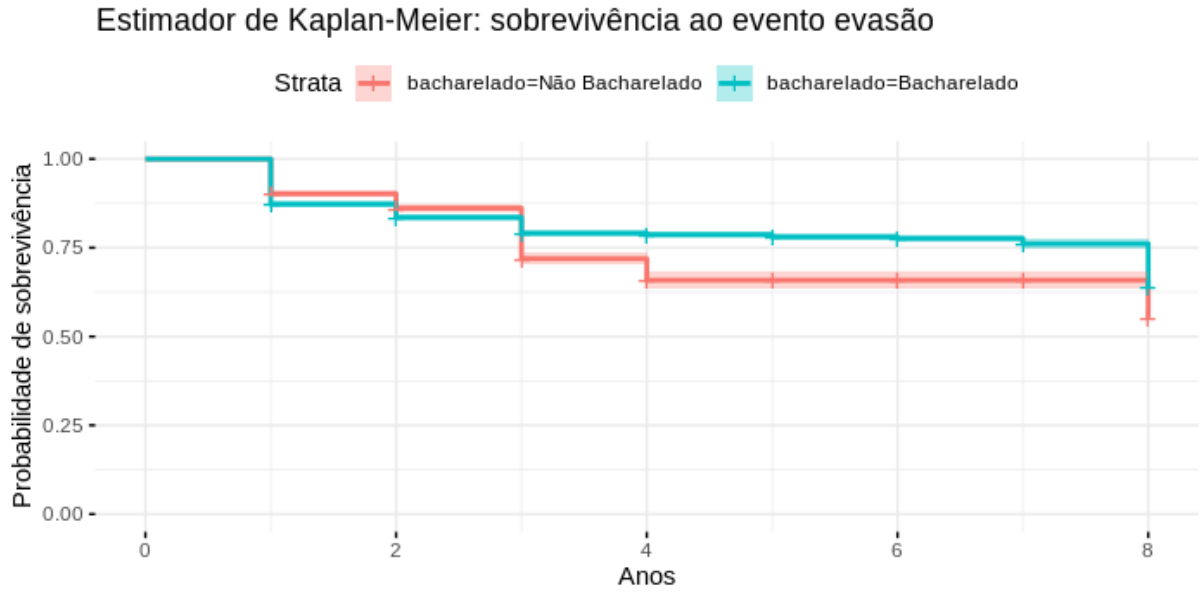
Figura 9 – Estimador de Kaplan-Meier: variável sexo



A variável binária "bacharelado" é verdadeira para alunos inscritos em cursos de bacharelado, e falsa para alunos de cursos de licenciatura e tecnológico, sendo a maior parte de licenciaturas. Conforme a Figura 10, estudantes de bacharelado apresentam evasão ligeiramente maior nos três primeiros anos, quando essa situação se reverte, terminando com quase 10 pontos percentuais de diferença entre a evasão de bacharelado (menor) e não-bacharelado (maior).

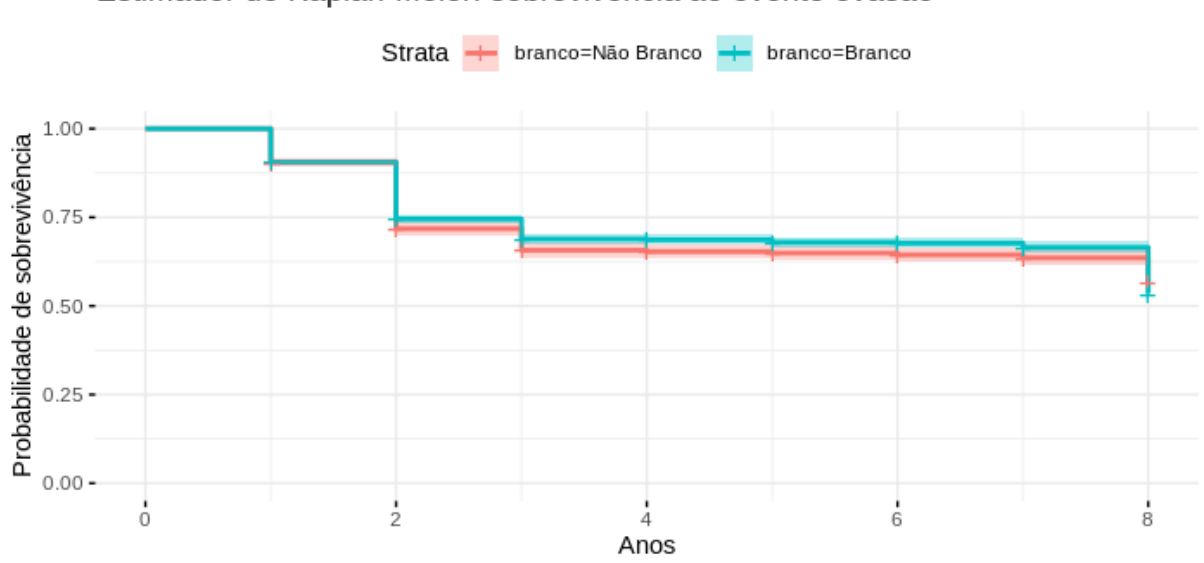
Quanto à cor ou raça auto-declarada dos estudantes, a Figura 11 aponta que estudantes brancos apresentam uma sobrevivência maior do que não brancos (incluindo pretos, pardos, amarelos e indígenas, excluindo os não declarados) para todos os períodos observados, sendo,

Figura 10 – Estimador de Kaplan-Meier: variável bacharelado



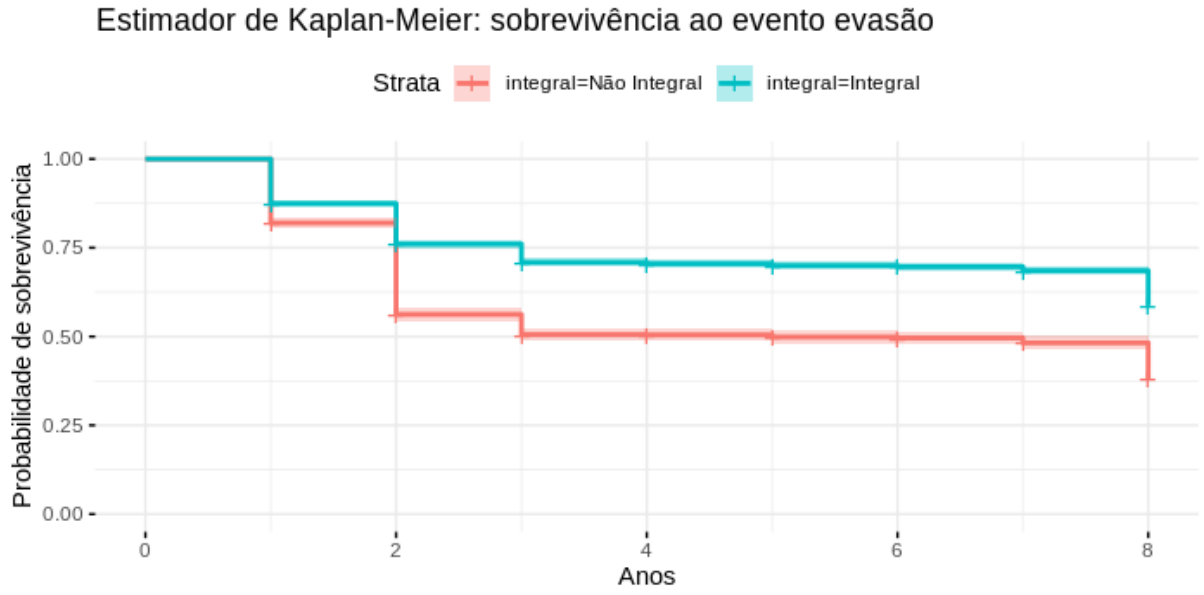
no entanto, nenhuma diferença significativa ao nível de significância de 1% – para Gehan e Tarone-Ware, são significativas a 5%, e para Logrank, a 10%.

Figura 11 – Estimador de Kaplan-Meier: variável branco



Por fim, considerando o turno dos estudantes, a distribuição de probabilidade de sobrevivência ao evento de evasão dentre os alunos matriculados em cursos integrais é consideravelmente maior do que os que cursam graduações em outros turnos (majoritariamente noturno), criando uma diferença estável do segundo ao oitavo ano de mais de 20 pontos percentuais, com entre os matriculados em cursos integrais ao final do período apresentando uma sobrevivência de quase 60%, enquanto alunos de cursos não-integrais, majoritariamente noturnos, finalizam o período com uma sobrevivência de cerca de 38% (Figura 12).

Figura 12 – Estimador de Kaplan-Meier: variável integral



4.3.2 Resultados para o curso de Ciências Econômicas

Avaliando especificamente os resultados para os alunos do curso de Ciências Econômicas da Universidade Federal Fluminense, reproduzimos a técnica de Kaplan-Meier para a as variáveis já abordadas nesta seção, e os respectivos testes. As variáveis binárias "bacharelado", que compara alunos de cursos de bacharelado com seus pares estudantes de licenciaturas e tecnólogos, e "modalidade", que compara cursos à distância e cursos presenciais, apresentam apenas um valor para o curso de ciências econômicas, já que se trata de um curso de bacharelado que não oferece opção de ensino à distância no contexto da Universidade Federal Fluminense.

Os alunos observados, matriculados em 2010 no curso de Ciências Econômicas, são 61,3% homens e 38,7% mulheres; 46,8% são auto-declarados brancos, 15,0% pardos, 3,3% pretos, 0,5% amarelos e 0,7% indígenas, e 33,8% não declararam cor ou raça. Por fim, 28,5% se matricularam no turno noturno, enquanto os 71,5%, no integral.

A Tabela 6 aponta os resultados dos testes de hipótese que verificam a diferença nas distribuições das curvas de sobrevida dos grupos comparados. De acordo com a tabela, para as três variáveis testadas, nenhum dos testes foi capaz de achar uma diferença significativa entre as curvas comparadas.

Tabela 6 – Testes de Gehan, Tarone-Ware e Logrank para o curso de Economia

Variável	Gehan	Tarone-Ware	Logrank
sexo	0,3758	0,3998	0,4856
branco	0,4000	0,4393	0,6117
integral	0,7434	0,8288	0,9408

Os Gráficos 13, 14 e 15 apresentam justamente o resultado acima, com curvas com

intervalos de confiança se sobrepondo. Ao contrário das curvas apresentadas anteriormente, o intervalo de confiança dos gráficos abaixo são bem maiores dado justamente a quantidade observações, que é consideravelmente menor após o filtro de curso. Por isso, ainda que os gráficos apontem que em média, as mulheres mostrem uma sobrevivência ligeiramente maior do que homens, e que brancos mostrem uma sobrevivência também ligeiramente maior do que não-brancos, não há prova estatística de que há diferença na distribuição dos dois grupos.

Figura 13 – Estimador de Kaplan-Meier para o curso de Economia: variável sexo

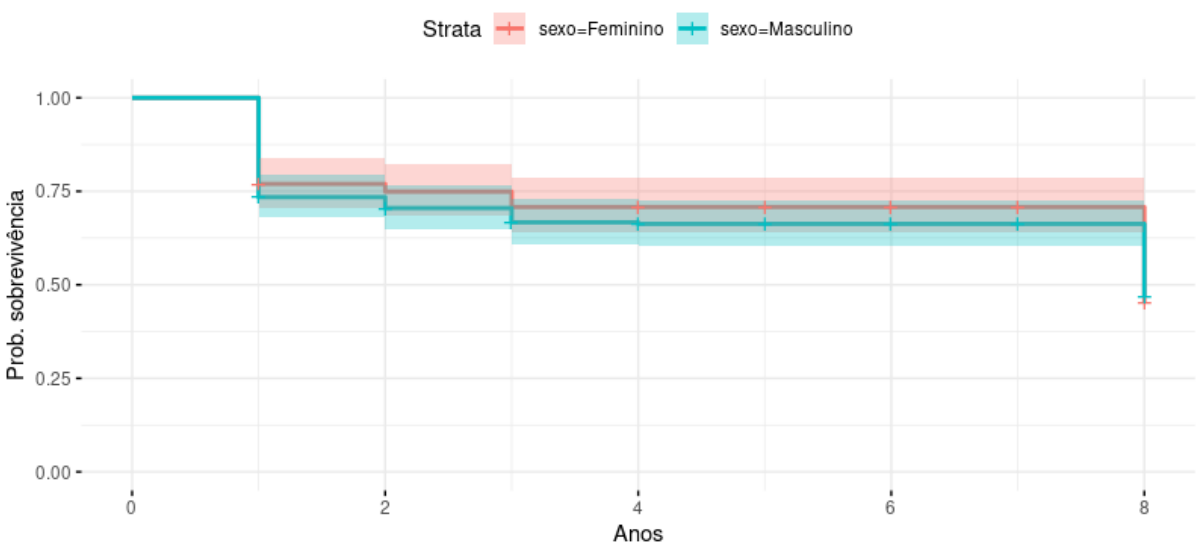


Figura 14 – Estimador de Kaplan-Meier para o curso de Economia: variável branco

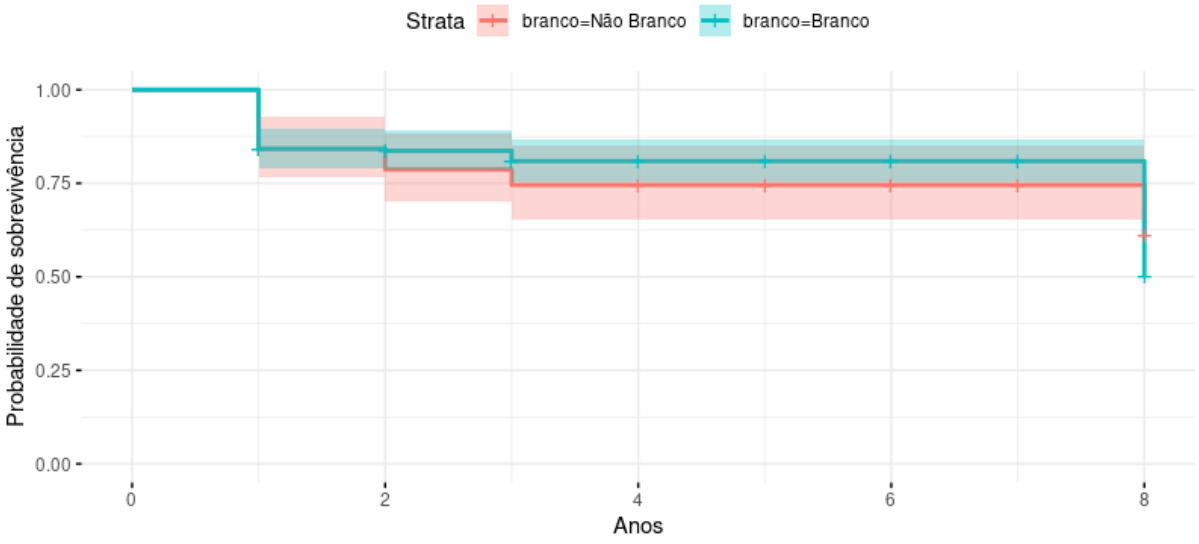
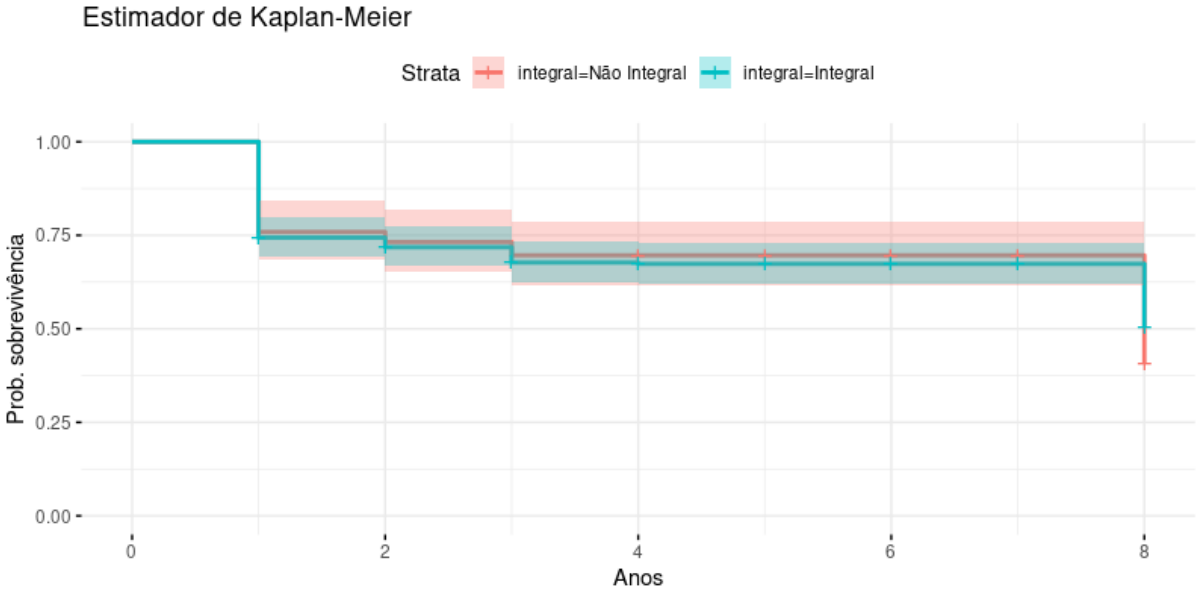


Figura 15 – Estimador de Kaplan-Meier para o curso de Economia: variável integral



4.4 Aplicação da regressão de Cox

Como mostrado no capítulo 3 e visando explorar os determinantes na evasão no ensino superior na UFF, adotamos a técnica da regressão de Cox, método semi-paramétrico. Com base nas estatíticas descritivas escolhemos como explicativas o sexo, a idade de ingresso no curso, a cor, tipo de ingresso, modalidade do curso. Assim, o modelo adquire a forma da equação (4.1). Na Tabela 7 mostramos a respectiva codificação identificado o grupo de controle com o valor 0.

$$h_i(y) = h_0(y)e^{\beta_1 sexo + \beta_2 idade\_ingresso + \beta_3 negro + \beta_4 bacharelado + \beta_5 ingresso\_vest + \beta_6 ead} \tag{4.1}$$

Tabela 7 – Variáveis utilizadas para estimação da regressão de Cox

Variável	Descrição
sexo	Os valores são Feminino e Masculino
idade_ingresso	Variável numérica de idade de ingresso no curso
negro	Os valore são Negro (1) e Não-Negro (0)
bacharelado	Os valores são Bacharelado (1) e Não-Bacharelado (0)
ingresso_vest	Os valores são Ingressou via Vestibular Próprio (1) e Outros (0)
ead	Os valores são Curso à Distância (1) e Curso Presencial (0)

Utilizando o pacote `survival` do software R, foi gerado um modelo com a função `coxph`, apropriada para modelos de regressão de riscos proporcionais, tendo como regressores, ou covariantes, as variáveis sexo, idade de ingresso, a dummy negro, a dummy bacharelado, a dummy ingresso_vest e a dummy EAD. Todas as variáveis incluídas são significantes ao nível de 1% para o modelo, conforme a Tabela 8, que é o sumário da regressão, e apresenta também os coeficientes obtidos.

Tabela 8 – Sumário da regressão de Cox

Variável	Coef	exp(Coef)	se(Coef)	z	Pr($>\|z\|$)
sexo_masculino	0,1761	1,1926	0,0443	3,973	0,0000
idade_ingresso	-0,0211	0,9791	0,0035	-6,052	0,0000
negro	-0,2500	0,7788	0,0634	-3,945	0,0000
bacharelado	-0,3817	0,6827	0 0600	-6,361	0,0000
ingresso_vest	-1,1220	0,3256	0,0545	-20,582	0,0000
ead	-0,9910	0,3712	0,0739	-13,406	0,0000

Observando o exp(Coef) do sumário, que é a representação do ϕ, descrito na seção 3.5.1. como a taxa de risco, podemos entender, ponto a ponto, que:

1. A taxa com que as falhas ocorrem para o grupo de homens (sexo_masculino = 1) é quase 20% em relação ao grupo de controle (mulheres, que são sexo_masculino = 0);
2. Para cada ano adicional de idade de ingresso, a taxa de risco diminui em 2,1%;
3. De acordo com o modelo, para estudantes negros (negro = 1), a taxa de risco é menor do que para os demais estudantes (negro = 0), a taxa de risco é 22,1% menor do que a taxa grupo de controle (negro = 0);
4. Estudantes de bacharelado (bacharelado = 1) apresentam uma taxa de risco de evasão 32% menor do que o grupo de controle, representado por estudantes de cursos de licenciatura e tecnólogos (bacharelado = 0);
5. Observando alunos que ingressaram, à época, por vestibular próprio da faculdade, apresentam risco de evasão de 32% em relação ao grupo de controle (estudantes que entraram por outros meios, principalmente ENEM). Como principal forma de ingresso na universidade o ENEM só foi instituído em 2012, sendo, em anos anteriores, muito menos expressivo do que o vestibular próprio da universidade;
6. Por fim, o modelo indica que estudantes de cursos à distância (ead = 1) apresenta uma taxa de risco de evasão muito menor do que estudantes de cursos presenciais, sendo apenas 37% do risco observado no grupo de controle (ead = 0).

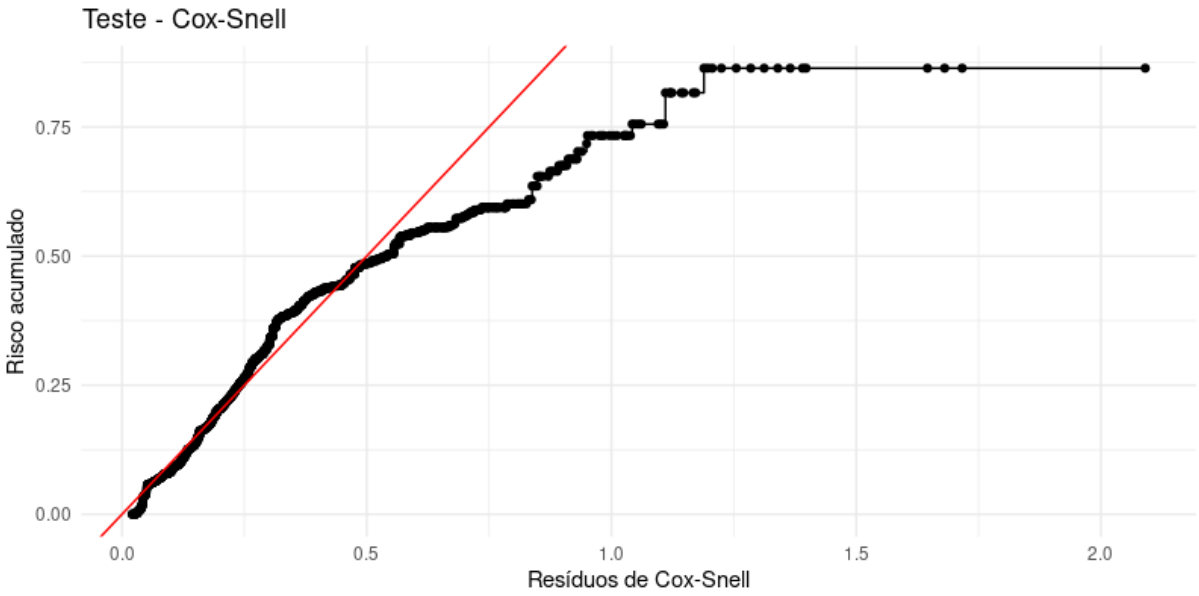
4.4.1 Testes

Para garantir que as hipóteses do modelo estejam sendo cumpridas, uma vez estimado o modelo, foram realizados testes, os já anteriormente mencionados na subseção sobre ajustamento de modelos.

O primeiro teste a ser realizado é o teste de Cox-Snell. De acordo com Colosimo e Giolo (2006, p.97), os resíduos de Cox-Snell são úteis para examinar o ajuste global do modelo final.

A Figura 16 abaixo, plotada a partir da função `gg_coxsnell`, do pacote `ldatools`, revela violação do pressuposto de riscos proporcionais, com os pontos observados sistematicamente se afastando da linha vermelha.

Figura 16 – Resíduos de Cox-Snell



Os resíduos de Martingale, por sua vez, são úteis para a observação das covariáveis numéricas (nesse caso, apenas uma) e suas formas funcionais. A curva LOESS desenhada a partir de seus resíduos deve ser minimamente linear para assumirmos que o pressuposto de linearidade da variável numérica está sendo atendido. Conforme a Figura 17, que foi construída usando a função `ggcoxfunctional` do pacote `survminer`, forma da variável numérica incluída no modelo é, certamente, não linear, nem sua transformação logarítmica.

Por fim, os resíduos deviance, que avaliam a acurácia do modelo para cada sujeito observado, são esperados comportamentos aleatórios distribuídos com alguma simetria ao redor de zero. Na Figura 18 (usando a função `ggcoxdiagnostics`), observamos que eles aparentam se concentrar abaixo de zero, e não estão tão simétricamente distribuídos, com os valores positivos bem distribuídos num intervalo entre 1 e 2, enquanto os valores negativos formam uma nuvem densa em -1.

Além disso, é claro, podemos também observar o não cumprimento do pressuposto de riscos proporcionais nas próprias figuras apresentadas (Figuras 7 a 12) na seção de Estimador de Kaplan-Meier, onde diversas curvas acabam por se cruzar, ou ter proporções razoavalmente diferentes ao longo dos períodos.

Figura 17 – Resíduos de Martingale - curva LOESS de idade_ingresso e log(idada_ingresso)

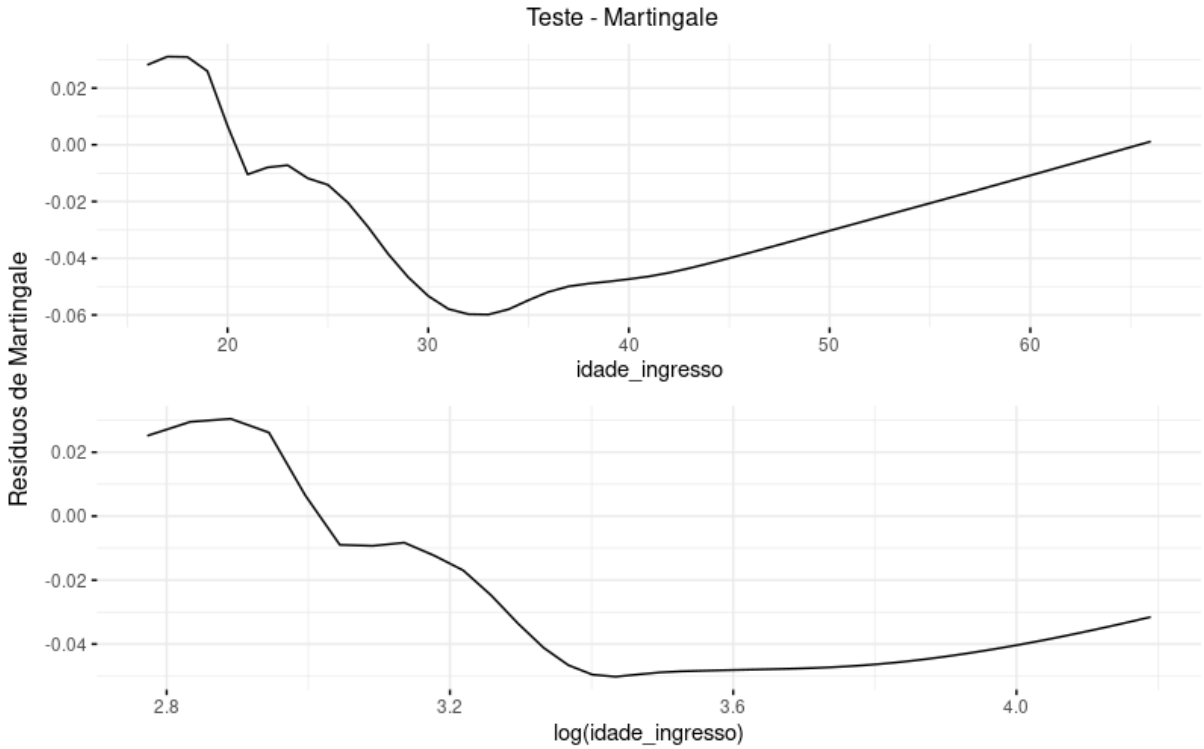
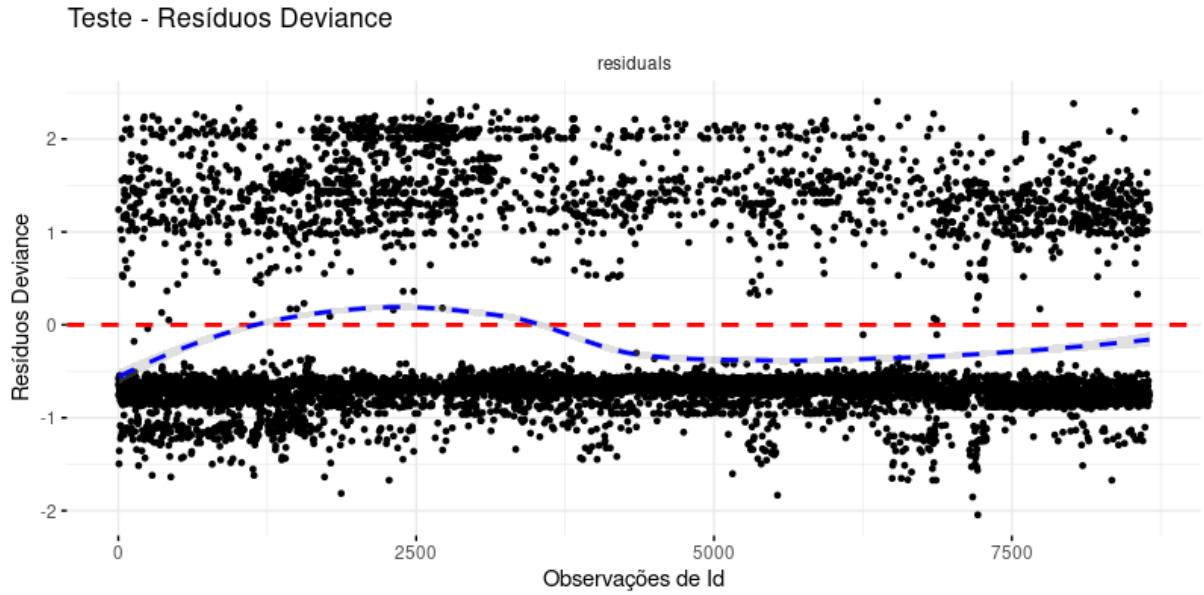


Figura 18 – Resíduos deviance



4.5 Considerações parciais

Variáveis que seriam de muita riqueza para tal regressão, como renda familiar, coeficiente de rendimento, dentre outros, não figuram entre as disponíveis no Censo, e ainda variáveis que teoricamente disponíveis, como escola de origem (pública ou privada), e o registro de atividades complementares (como estágio, iniciação científica, etc), não eram devidamente preenchidas à época, com uma quantidade enorme de observações vazias, ao menos para a instituição observada.

No que se refere aos resultados da aplicação da análise de sobrevida para os cursos da UFF, podemos observar que no que trata da modalidade do curso, foi estimada uma evasão bem maior para os cursos presenciais nos três primeiros anos, se comparada aos cursos à distância; no quarto ano, entretando, a evasão para cursos presenciais se estabiliza, enquanto nos cursos à distância, ainda há uma queda; a diferença entre as duas curvas chega a ser de mais de 25 pontos percentuais (no segundo ano), mas se estabiliza, ao final dos 8 períodos observados, com uma diferença bem pequena.

Quanto à avaliação das curvas de sobrevivência de homens e mulheres, há pouca diferença em termos de sobrevivência, sendo significativa em apenas ao nível significância de 5% no teste de Logrank e de 10% no teste de Tarone-Ware.

Cursos de bacharelado, ao final do tempo observado, tendem ter uma evasão menor do que outros cursos (majoritamente licenciaturas); até o terceiro ano, o comportamento é muito similar, com a sobrevivência em cursos de bacharelado sendo ligeiramente pior, mas a partir do terceiro ano, a evasão do outro grupo supera a curva de bacharelado, e a diferença de 12,5 pontos percentuais se mantém relativamente constante até o final do período observado.

A variável binária *branco* é tem valor 1 para pessoas auto-declaradas brancas, e valor 0 para pessoas auto-declaradas de outra etnia ou raça; observando a curva de sobrevivência desses dois grupos até o evento de evasão, é possível observar que as formas de ambas são muito similares, com a curva de sobrevivência do grupo não-branco sempre ligeiramente abaixo do grupo branco, com uma diferença consistente de cerca de 3 pontos percentuais, mas com o intervalo de confiança se tocando. Ao fim do período, o teste de Logrank conseguiu identificar significativa entre as curvas a um nível de significância de 10%, e os testes de Gehan e Tarone-Ware, a 5%.

Por fim, as curvas que mais se destacam, em termos de diferença, são as dos grupos integral e não-integral – as duas categorias da variável binária que classifica se o sujeito observado estava ou não matriculado num curso integral. A maioria dos usuários classificados no grupo não-integral são do turno noturno. As curvas empíricas apontam que, já no primeiro ano, o grupo não-integral tem uma sobrevivência um pouco menor do que o grupo integral, de menos de 5 pontos percentuais, mas a partir do segundo ano, essa diferença salta para mais de 20 pontos percentuais, e se mantém até o fim do período observado, sendo significativa a nível de 1% para

todos os testes rodados.

A partir de tais achados, podemos entender melhor como a evasão escolar afeta os estudantes de distintos contextos e características, e, a partir desse conhecimento, pensar em políticas de permanência que visem tais públicos, de modo a promover efetivamente a inclusão e garantir o acesso ao ensino superior.

5 CONSIDERAÇÕES FINAIS

Nesta monografia, abordamos a análise de sobrevida para avaliar um grave problema enfrentado pelo sistema de educação superior, a evasão; como visto no capítulo 2, tal metodologia é amplamente utilizada na literatura nacional para a compreensão do fluxo escolar, através de estudos como Lima Junior, Silveira e Ostermann (2012), Mello et al (2015) e Franca e Saccaro (2018).

A respeito do arcabouço teórico empregado, discorremos sobre os conceitos básicos, passando por diferentes métodos, distribuições, estimadores e testes, conforme observado no Capítulo 3. Neste capítulo foram apresentadas, além de outros técnicas, o estimador de Kaplan-Meier e a regressão de Cox, aplicadas nos dados utilizados para a construção do trabalho.

No Capítulo 4, foi apresentada a base de dados utilizada, o Censo da Educação Superior, e as estatísticas descritivas – o perfil – dos estudantes observados ao longo de 8 anos: aqueles matriculados na Universidade Federal Fluminense no ano de 2010. A partir dos dados, pudemos aplicar o estimador de Kaplan-Meier em distintos grupos para compreender melhor a curva de evasão dos estudantes, de onde pudemos concluir que grupos como estudantes de cursos noturnos, ou cursos de licenciatura ou tecnólogos, tendem a ser mais vulneráveis – isto é, apresentando uma evasão maior no período observado – do que seus opostos, os estudantes de cursos integrais, ou cursos de bacharelado.

Por fim, ainda no Capítulo 4, desenvolvemos a respeito da regressão de Cox estimada para os dados disponíveis e os testes realizados para a verificação do cumprimento dos pressupostos do modelo. Conforme apontado, os principais pressupostos foram violados, levando a conclusão de o modelo não é apropriado. Uma consideração interessante é que, ainda que o modelo não tenha sido bem-sucedido, por falta de mais variáveis adequadas, a exposição do método e da interpretação trazem ganho de conteúdo para o trabalho e contribuem para o propósito do mesmo.

6 REFERÊNCIAS

BORGES, A. *Análise de Sobrevivência com o R. Dissertação*. Dissertação (Mestrado) — Programa de Pós-Graduação em Matemática/Universidade da Madeira, 2014. Citado na página 30.

CÂNDIDO, L. *Evasão universitária na graduação presencial brasileira: um panorama apartir do Censo da Educação Superior (2009-2018)*. Monografia (Bacharelado) — Faculdade de Economia/UFF, 2019. Citado 3 vezes nas páginas 14, 15 e 17.

COLOSIMO, E.; GIOLO, S. *Análise de sobrevivência aplicada*. São Paulo: Edgard Blucher, 2006. Citado 5 vezes nas páginas 21, 24, 25, 30 e 46.

DESAFIOS DA EDUCACÃO. *Evasão diminui no ensino superior, aponta levantamento*. [S.l.], 2019. Disponível em: <https://desafiosdaeducacao.grupoa.com.br/evasao-dos-estudantes-tem-queda/>. Acesso em: 2020-10-12. Citado na página 11.

DOBSON, A.; BARNETT, A. *An Introduction to Generalized Linear Models*. Flórida, Estados Unidos: CRC Press, 2018. Citado 5 vezes nas páginas 22, 24, 27, 28 e 31.

FORPLAD. *Indicadores*. Ouro Preto, 2015. Disponível em: <http://www.uff.br/sites/default/files/indicadores_do_forplad.pdf>. Acesso em: 2020-10-16. Citado na página 14.

FORPLAD. *Taxa de Sucesso, Evasão e Retenção nas IFES*. Vitoria, 2016. Citado na página 11.

FORPLAD. *Nota Técnica MEC/SE Nº 4/2018*. [S.l.], 2018. Citado na página 13.

FRANCA, M.; SACCARO, A. *Gastos governamentais no ensino superior e evasão: um estudo de análise de sobrevivência para os estudantes dos cursos de ciências naturais e engenharias em instituições públicas e privadas*. [S.l.], 2018. Disponível em: <https://ideas.repec.org/p/anp/en2016/67.html>. Citado 5 vezes nas páginas 12, 18, 19, 22 e 51.

G1. *País perde R$9 bilhões com evasão no ensino superior, diz pesquisador*. [S.l.], 2011. Disponível em: <http://g1.globo.com/educacao/noticia/2011/02/pais-perde-r-9-bilhoes-com-evasao-no-ensino-superior-diz-pesquisador.html>. Acesso em: 2020-10-19. Citado 2 vezes nas páginas 12 e 15.

HOED, R. *Análise da evasão em cursos superiores: o caso da evasão em cursos superiores da área de Computação*. Dissertação (Mestrado) — Departamento de Ciência da Computação/UnB, 2016. Citado na página 12.

INEP. *Sinopses Estatísticas da Educação Superior – Graduação*. Brasília, 2018. Disponível em: <http://inep.gov.br/web/guest/sinopses-estatisticas-da-educacao-superior>. Acesso em: mar. 2019. Citado na página 32.

KLEIN, J.; MOESCHBERGER, M. *Survival Analysis: techniques for censored and truncated data*. Nova Iorque, Estados Unidos: Springer, 2003. Citado na página 24.

LIMA JUNIOR, P.; SILVEIRA, F.; OSTERMANN, F. Análise de sobrevivência aplicada ao estudo do fluxo escolar nos cursos de graduação em física: um exemplo de uma universidade brasileira. *Revista Brasileira de Ensino de Física*, scielo, v. 34, p. 1 – 10, 03 2012. ISSN

1806-1117. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1806-11172012000100014&nrm=iso>. Citado 7 vezes nas páginas 12, 17, 18, 19, 22, 23 e 51.

MEC. *REUNI - Diretrizes Gerais*. Brasília, 2007. Disponível em: <http://portal.mec.gov.br/sesu/arquivos/pdf/diretrizesreuni.pdf>. Acesso em: 2019-11-07. Citado na página 11.

MELLO, L. et al. Estudo do tempo depermanência de cursistas em um curso realizado na modalidade à distância. 2015. Disponível em: <https://cancri.ead.unesp.br/sigeve/paginas/baixar_trabalho_aprovado.php?id=545>. Acesso em: 2020-10-12. Citado 3 vezes nas páginas 18, 19 e 51.

PONTES, D. *Retenção: perfil do aluno retido e suas percepções sobre as políticas existentes na UFF*. Niteroi, 2015. Disponível em: <http://www.uff.br/?q=noticias/29-06-2015/pesquisa-inedita-analisa-causas-da-retencao-de-alunos-da-uff>. Acesso em: mar. 2019. Citado na página 12.

R CORE TEAM. *R: A Language and Environment for Statistical Computing*. Viena, Áustria: R Foundation for Statistical Computing, 2020. Disponível em: <https://www.R-project.org/>. Acesso em: 2020-10-15. Citado na página 39.

SILVA, A. et al. Modelos de sobrevivência aplicados à evasão dos alunos de estatística da ufpb. *InterScientia*, v. 6, p. 135 – 145, 2018. Disponível em: <https://periodicos.unipe.br/index.php/interscientia/article/view/860>. Citado 2 vezes nas páginas 18 e 19.

VITELLI, R.; FRITSCH, R. Evasão escolar na educação superior: de que indicador estamos falando? *Estudos em Avaliação Educacional*, São Paulo, v. 27, p. 908–937, 2016. Citado na página 14.

A SCRIPT EM R COM COMENTÁRIOS

Carregando os pacotes necessários:

```
library(tidyverse)
library(devtools)
library(survival)
library(survminer)
library(coin)
library(openxlsx)
library(lubridate)
install_github("adibender/ldatools", force = TRUE)
library(ldatools)
```

Carregando as bases de 2010 a 2017 já tratadas. É vaĺido comentar que um problema frequentemente lidado foi a mudança de nome das variáveis ao longo do tempo.

```
load("./Bases/df10.RData")
load("./Bases/df11.RData")
load("./Bases/df12.RData")
load("./Bases/df13.RData")
load("./Bases/df14.RData")
load("./Bases/df15.RData")
load("./Bases/df15.RData")
load("./Bases/df16.RData")
load("./Bases/df17.RData")
```

Filtrando os alunos que ingressaram em 2010:

```
df10 <- df10 %>%
  filter(NU_ANO_INGRESSO == 2010)

idsUnicos <- df10 %>%
  select(CO_ALUNO,
         CO_CURSO) %>%
  distinct()
```

Filtrando os IDs dos alunos que ingressaram em 2010 nos anos posteriores:

```
df11 <- inner_join(df11,
                   idsUnicos,
                   by = c("CO_ALUNO" = "CO_ALUNO",
                   "CO_CURSO" = "CO_CURSO"))
```

```
df12 <- inner_join(df12,
                   idsUnicos,
                   by = c("CO_ALUNO" = "CO_ALUNO",
                   "CO_CURSO" = "CO_CURSO"))

df13 <- inner_join(df13,
                   idsUnicos,
                   by = c("CO_ALUNO" = "CO_ALUNO",
                   "CO_CURSO" = "CO_CURSO"))

df14 <- inner_join(df14,
                   idsUnicos,
                   by = c("CO_ALUNO" = "CO_ALUNO",
                   "CO_CURSO" = "CO_CURSO"))

df15 <- inner_join(df15,
                   idsUnicos,
                   by = c("CO_ALUNO" = "CO_ALUNO",
                   "CO_CURSO" = "CO_CURSO"))

df16 <- inner_join(df16,
                   idsUnicos,
                   by = c("CO_ALUNO" = "CO_ALUNO",
                   "CO_CURSO" = "CO_CURSO"))

df17 <- inner_join(df17,
                   idsUnicos,
                   by = c("CO_ALUNO" = "CO_ALUNO",
                   "CO_CURSO" = "CO_CURSO"))
```

Juntando todos os anos em um único dataset:

```
all <- rbind(df10,
             df11,
             df12,
             df13,
             df14,
             df15,
             df16,
             df17) %>%
  arrange(CO_ALUNO)
```

Corrigindo um problema da base: para algumas pessoas, o ano de ingresso em databases

posteriores não está preenchido.

```
all$NU_ANO_INGRESSO[is.na(all$NU_ANO_INGRESSO)] <- 2010

all <- all %>% filter(NU_ANO_INGRESSO == 2010)
```

Ao longo do tempo, as variáveis foram mudando, e algumas até saiam do book de variáveis disponíveis no Censo. Aqui, pego algumas variáveis da base de 2010, ano de matrícula dos alunos a serem acompanhados:

```
csv10 <- read.csv("./Bases/dm_aluno_uff10.csv", sep = "|") %>%
  select(CO_ALUNO,
         CO_CURSO,
         NO_CURSO,
         CO_PAIS_ORIGEM_ALUNO,
         CO_UF_NASCIMENTO,
         CO_MUNICIPIO_NASCIMENTO,
         NU_ANO_INGRESSO = ANO_INGRESSO) %>%
  filter(NU_ANO_INGRESSO == 2010) %>%
  select(-NU_ANO_INGRESSO)

all <- left_join(all,
                 csv10,
                 by = c("CO_ALUNO" = "CO_ALUNO",
                 "CO_CURSO" = "CO_CURSO"))

rm(csv10, df10, df11, df12, df13, df14, df15, df16, df17, idsUnicos)

ano2010 <- all %>%
  filter(NU_ANO_CENSO == 2010)
```

Dados invariantes dos declarados no ano de matricula. Infelizmente, variáveis ricas como se o estudante participou de monitorias, iniciações científicas, teve auxílio permanência, realizou estágio profissional, dentre outros, não eram adequadamente preenchidas na época. Além disso, dado que a aprovação da Lei das Cotas data de 2012, apenas para os alunos ingressantes de 2013 há o preenchimento das variáveis de vagas reservadas (cotas); em 2010 ainda não há esse mecanismo.

```
invariant <- ano2010 %>% select(CO_ALUNO,
                                TP_SEXO,
                                CO_CURSO,
                                CO_CURSO_POLO,
                                NU_IDADE,
                                TP_GRAU_ACADEMICO,
```

```
                                TP_COR_RACA,
                                TP_MODALIDADE_ENSINO,
                                TP_TURNO,
                                IN_INGRESSO_VESTIBULAR,
                                IN_INGRESSO_ENEM,
                                IN_INGRESSO_TOTAL) %>%
  mutate(TP_SEXO = factor(TP_SEXO,
                          levels = c(1, 0),
                          labels = c("Feminino", "Masculino")),
         TP_GRAU_ACADEMICO = factor(TP_GRAU_ACADEMICO,
                                    levels = c(1, 2, 3),
                                    labels = c("Bacharelado",
                                               "Licenciatura",
                                               "Tecnologico")),
         TP_COR_RACA = factor(TP_COR_RACA,
                              levels = c(1, 2, 3, 4, 5, 6, 0),
                              labels = c("Branca",
                                         "Preta",
                                         "Parda",
                                         "Amarela",
                                         "Indigena",
                                         "Nao␣dispoe␣da␣informacao",
                                         "Nao␣declarado")),
         TP_MODALIDADE_ENSINO = factor(TP_MODALIDADE_ENSINO,
                                       levels = c(1, 2),
                                       labels = c("Presencial",
                                                  "Curso␣a␣distancia")),
         TP_TURNO = factor(TP_TURNO,
                           levels = c(1, 2, 3, 4),
                           labels = c("Matutino",
                                      "Vespertinho",
                                      "Noturno",
                                      "Integral")),
         IN_INGRESSO_VESTIBULAR = factor(IN_INGRESSO_VESTIBULAR,
                                         levels = c(0, 1),
                                         labels = c("Nao",
                                                    "Sim")),
         IN_INGRESSO_ENEM = factor(IN_INGRESSO_ENEM,
                                   levels = c(0, 1),
                                   labels = c("Nao",
                                              "Sim")),
         IN_INGRESSO_TOTAL = factor(IN_INGRESSO_TOTAL,
```

```
                                    levels = c(0, 1),
                                    labels = c("Nao",
                                               "Sim")))
```

Dados variantes:

```
variants <- all %>% select(CO_ALUNO,
                           CO_CURSO,
                           TP_SITUACAO,
                           DT_INGRESSO_CURSO,
                           IN_MATRICULA,
                           IN_CONCLUINTE,
                           NU_ANO_INGRESSO,
                           NU_ANO_CENSO)

variants <- left_join(invariant,
                      variants,
                      by = c("CO_ALUNO" = "CO_ALUNO",
                             "CO_CURSO" = "CO_CURSO"))
```

Removendo os alunos falecidos:

```
falecidos <- data.frame(table(variants$CO_ALUNO[variants$TP_SITUACAO == 7]),
                        stringsAsFactors = F) %>%
  mutate(CO_ALUNO = as.numeric(as.character(Var1))) %>%
  select(CO_ALUNO)

variants <- anti_join(variants,
                      falecidos,
                      by = "CO_ALUNO")

data <- variants %>%
  select(CO_ALUNO,
  CO_CURSO,
  CO_CURSO_POLO,
  TP_SEXO,
  NU_IDADE,
  TP_GRAU_ACADEMICO,
  TP_COR_RACA,
  TP_MODALIDADE_ENSINO,
  TP_TURNO,
  IN_INGRESSO_VESTIBULAR,
  IN_INGRESSO_ENEM,
  TP_SITUACAO,
```

```
  IN_CONCLUINTE,
  NU_ANO_INGRESSO,
  NU_ANO_CENSO)

names(data) <- c("id",
                 "curso",
                 "polo",
                 "sexo",
                 "idade_ingresso",
                 "grau",
                 "cor",
                 "modalidade",
                 "turno",
                 "ingresso_vest",
                 "ingresso_enem",
                 "situacao",
                 "concluinte",
                 "ano_ingresso",
                 "ano_censo")

grau <- data %>%
  select(curso, grau) %>%
  distinct()

data <- data %>%
  select(-grau)

data <- left_join(data,
                  grau,
                  by = "curso")


data$id <- as.character(data$id)

ids <- as.character(unique(data$id))

classif <- data.frame(id = NA,
                      censurado = NA,
                      formado = NA,
                      evadido = NA,
                      tempo_observado = NA)
```

Classificação dos alunos:

```
for(i in 1:length(ids)){
  aluno <- data %>%
    filter(id == ids[i])

  if(any(c(last(aluno$situacao) == 2,
           last(aluno$situacao) == 3,
           is.na(last(aluno$situacao))))){

    censurado = 1} else {censurado = 0}

  if(!is.na(last(aluno$situacao))){
    if(last(aluno$situacao) == 6){
      formado = 1} else {formado = 0}
  } else {formado = 0}


  if(!is.na(last(aluno$situacao))){
    if(any(c(last(aluno$situacao) == 4,
             last(aluno$situacao) == 5))){
      evadido = 1} else {evadido = 0}
  } else {evadido = 0}

  tempo_observado = nrow(aluno)

  classif <- rbind(classif,
                   c(ids[i], censurado,
                     formado, evadido,
                     tempo_observado))
}



final <- data %>%
  select(id,
         curso,
         sexo,
         idade_ingresso,
         cor,
         modalidade,
         turno,
         ingresso_vest,
```

```
        ingresso_enem,
        grau) %>%
  distinct() %>%
  left_join(classif, by = "id")
```

Dicotomizando variáveis ainda não binárias:

```
final$branco <- ifelse(final$cor == "Branca", 1, 0)
final$branco <- factor(final$branco,
                       levels = c(0, 1),
                       labels = c("Nao␣Branco",
                                  "Branco"))
final$branco[final$cor == "Nao␣dispoe␣da␣informacao" |
               final$cor == "Nao␣declarado"] <- NA

final$negro <- ifelse((final$cor %in% c("Preta", "Parda")), 1, 0)
final$negro <- factor(final$negro,
                      levels = c(0, 1),
                      labels = c("Nao␣Negro",
                                 "Negro"))

final$preto <- ifelse((final$cor %in% c("Preta")), 1, 0)
final$preto <- factor(final$preto,
                      levels = c(0, 1),
                      labels = c("Nao␣Preto",
                                 "Preto"))

final$noturno <- ifelse(final$turno == "Noturno", 1, 0)
final$noturno <- factor(final$noturno,
                        levels = c(0, 1),
                        labels = c("Nao␣Noturno",
                                   "Noturno"))

final$integral <- ifelse(final$turno == "Integral", 1, 0)
final$integral <- factor(final$integral,
                         levels = c(0, 1),
                         labels = c("Nao␣Integral",
                                    "Integral"))

final$ingresso_vest <- factor(as.numeric(final$ingresso_vest),
                              levels = c(1, 2),
                              labels = c("Outro",
                                         "Ingresso␣via␣Vestibular␣Proprio"))
```

```
final$bacharelado <- ifelse(final$grau == "Bacharelado", 1, 0)
final$bacharelado <- factor(final$bacharelado,
                            levels = c(0, 1),
                            labels = c("Nao␣Bacharelado",
                                       "Bacharelado"))

final$licenciatura <- ifelse(final$grau == "Licenciatura", 1, 0)
final$licenciatura <- factor(final$licenciatura,
                             levels = c(0, 1),
                             labels = c("Nao␣Licenciatura",
                                        "Licenciatura"))

final$tempo_observado <- as.numeric(final$tempo_observado)
final$censurado <- as.numeric(final$censurado)
final$formado <- as.numeric(final$formado)
final$evadido <- as.numeric(final$evadido)
```

Incluindo os nomes dos cursos já corrigidos (vários erros de digitação ou formatação foram observados):

```
nomes <- read.csv("./nomes_cursos.csv")

final <- left_join(final, nomes, by = "curso")

write.csv(final, paste0("./data_final_", today(), ".csv"))

evadidos <- final %>%
  filter(evadido == 1)
```

Estatísticas:

```
table(final$sexo)
cor <- data.frame(table(final$cor))
cor$prop <- cor$Freq*100/sum(cor$Freq)
mean(final$idade_ingresso)
quantile(final$idade_ingresso)
table(final$modalidade)
turno <- data.frame(table(final$turno))
turno$prop <- turno$Freq*100/sum(turno$Freq)
grau <- data.frame(table(final$grau))
grau$prop <- grau$Freq*100/sum(grau$Freq)
table(final$ingresso_vest)
table(final$ingresso_enem)
```

Estimador de Kaplan-Meier e testes:

```
surv_object <- Surv(time = final$tempo_observado, event = final$evadido)

ggsurvplot(survfit(surv_object ~ sexo, data = final),
           data = final,
           pval = FALSE,
           conf.int = TRUE,
           ggtheme = theme_minimal())

ggsurvplot(survfit(surv_object ~ modalidade, data = final),
           data = final,
           pval = FALSE,
           conf.int = TRUE,
           ggtheme = theme_minimal())

ggsurvplot(survfit(surv_object ~ branco, data = final),
           data = final,
           pval = FALSE,
           conf.int = TRUE,
           ggtheme = theme_minimal())

ggsurvplot(survfit(surv_object ~ integral, data = final),
           data = final,
           pval = FALSE,
           conf.int = TRUE,
           ggtheme = theme_minimal())

ggsurvplot(survfit(surv_object ~ bacharelado, data = final),
           data = final,
           pval = FALSE,
           conf.int = TRUE,
           ggtheme = theme_minimal())

logrank_test(surv_object ~ sexo,
             final,
             type = c("logrank"))
logrank_test(surv_object ~ branco,
             final,
             type = c("logrank"))
logrank_test(surv_object ~ modalidade,
             final,
             type = c("logrank"))
```

```
logrank_test(surv_object ~ integral,
             final,
             type = c("logrank"))
logrank_test(surv_object ~ bacharelado,
             final,
             type = c("logrank"))

logrank_test(surv_object ~ sexo,
             final,
             type = c("Gehan-Breslow"))
logrank_test(surv_object ~ branco,
             final,
             type = c("Gehan-Breslow"))
logrank_test(surv_object ~ modalidade,
             final,
             type = c("Gehan-Breslow"))
logrank_test(surv_object ~ integral,
             final,
             type = c("Gehan-Breslow"))
logrank_test(surv_object ~ bacharelado,
             final,
             type = c("Gehan-Breslow"))

logrank_test(surv_object ~ sexo,
             final,
             type = c("Tarone-Ware"))
logrank_test(surv_object ~ branco,
             final,
             type = c("Tarone-Ware"))
logrank_test(surv_object ~ modalidade,
             final,
             type = c("Tarone-Ware"))
logrank_test(surv_object ~ integral,
             final,
             type = c("Tarone-Ware"))
logrank_test(surv_object ~ bacharelado,
             final,
             type = c("Tarone-Ware"))
```

Regressão de Cox e testes:

```
final$ead <- ifelse(final$modalidade == "Curso␣a␣distancia", 1, 0)
final$ingresso_vest <- ifelse(final$ingresso_vest == "Outro", 0, 1)
```

```
final$negro <- ifelse(final$negro == "Negro", 1, 0)
final$bacharelado <- ifelse(final$bacharelado == "Bacharelado", 1, 0)

cox <- survival::coxph(surv_object ~ sexo +
                         idade_ingresso +
                         negro +
                         bacharelado +
                         ingresso_vest +
                         ead,
                       data = final)

summary(cox)

gg_coxsnell(cox) +
  theme_minimal() +
  geom_abline(intercept=0, slope=1, col=2)

survminer::ggcoxfunctional(formula = surv_object ~ idade_ingresso,
                           data = final,
                           fit = coxph(surv_object ~ idade_ingresso,
                                       data = final),
                           ggtheme = theme_minimal())

ggcoxdiagnostics(cox, type = "deviance",
                 linear.predictions = FALSE,
                 ggtheme = theme_minimal())
```