Universidade de Brasília
Departamento de Estatística

Um Estudo da Evasão no Curso de Licenciatura em Computação da Universidade de Brasília por meio de Modelos de Análise de Sobrevivência

Richard Wallan Paulino de Sousa

Relatório final apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2021

Richard Wallan Paulino de Sousa

Um Estudo da Evasão no Curso de Licenciatura em Computação da Universidade de Brasília por meio de Modelos de Análise de Sobrevivência

Orientadora: Prof(a). Juliana Betini Fachini Gomes

Relatório final apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2021

Resumo

Este trabalho teve como objetivo utilizar técnicas de análise de sobrevivência para o estudo de fatores que influenciam no tempo, a partir do ingresso do alunos da Universidade de Brasília no curso de Licenciatura em Computação, até a evasão ou não evasão. A metodologia de análise de sobrevivência permite a inclusão da falha (cometer evasão) e também das informações parciais, consideradas como censura (alunos que não evadiram). O conjunto de dados completo é composto por 728 observações e 22 variáveis. O período de estudo considerou alunos que ingressaram no curso de Licenciatura em Computação entre 2012/2 e 2019/2. A análise foi dividida em três perspectivas: a primeira contendo um modelo considerando o banco de dados completo, a segunda com o modelo contendo apenas alunos do sexo masculino e a terceira com o modelo apenas para o sexo feminino. O modelo de regressão Log-normal mostrou-se adequado para analisar os três bancos de dados e observou-se que o conjunto de variáveis explicativas muda para os diferentes bancos de dados. Porém, a variável de indíce de rendimento acadêmico (IRA), variável importante na Universidade para medir o rendimento acadêmico dos estudantes, é significativa em todos os modelos apresentados e apresenta um efeito positivo na curva de sobrevivência dos alunos. Isto é, alunos com maiores valores de IRA possuem maior probabilidade de sobreviver à evasão.

Palavras-chave: Evasão; Ensino Superior; Universidade de Brasília; Licenciatura em Computação; Análise de sobrevivência; Censura; Distribuição Log-Normal; Modelo de regressão; Censo da Educação Superior

Abstract

This work aimed to use survival analysis techniques for the study of factors that influence time, from the entrance of students at the University of Brasilia in the Computer course, until evasion or not evasion. The survival analysis methodology allows the inclusion of failure (committing dropout) and also of partial information, considered as censorship (students who did not drop out). The complete dataset consists of 728 observations and 22 variables. The study period considered students who entered the Computing course between 2012/2 and 2019/2. The analysis was divided into three perspectives: the first containing a model considering the complete database, the second with the model containing only male students and the third with the model only female students. The Log-normal regression model proved to be adequate to analyze the three databases and it was observed that the set of explanatory variables changes for the different databases. However, the academic performance index (ARI) variable, an important variable at the University to measure students' academic performance, is significant in all models presented and has a positive effect on the students' survival curve. That is, students with higher ARI values are more likely to survive dropout.

Keywords: Evasion; University education; University of Brasilia; Degree in Computing; Survival analysis; Censorship; Log-Normal Distribution; Regression model; Higher Education Census

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

A evasão escolar está presente desde os níveis básicos de educação até o ensino superior estando relacionada a fatores de diferentes naturezas tais como a satisfação com a instituição de ensino (RIBEIRO; CORREIA; CAMPOS, 2021) e outras variáveis relacionadas a fatores sociais e individuais do aluno como: desempenho acadêmico, escolha profissional e indicadores econômicos (FRITSCH; ROCHA; VITELLI, 2015).

Sendo assim, "a evasão escolar está relacionada à perda de estudantes que iniciam, mas não concluem seus cursos"(FRITSCH; ROCHA; VITELLI, 2015) ou seja, "é a desistência do aluno do seu curso de origem por qualquer motivo, exceto conclusão ou diplomação"(FRITSCH; ROCHA; VITELLI, 2015). Fritsch, Rocha e Vitelli (2015) também mencionam que o Ministério da Educação (MEC) conceitua evasão como sendo "a saída definitiva do curso de origem sem conclusão, ou a diferença entre ingressantes e concluintes, após uma geração completa".

Contudo, devido a complexidade do entendimento do que se relaciona com a evasão escolar pode-se afirmar que:

> As formas de interpretação não permitem chegar a uma definição precisa de "evasão e abandono escolar", uma vez que esta requer uma compreensão das relações entre os motivos de ingresso e a trajetória dos permanecentes, dos desistentes e egressos desse público (FILHO; ARAÚJO, 2017).

Dado o contexto do que é a evasão escolar e a complexidade de fatores relacionados, no Brasil a lei nº 9.394, de 20 de dezembro de 1996, estabelece as diretrizes e bases da educação nacional (LDB) de tal forma que o artigo 3º descreve que "[...] o ensino será ministrado com base nos princípios de igualdade de condições de acesso e permanência escolar"(FEDERAL, 2005) (FRITSCH; ROCHA; VITELLI, 2015).

Ademais, no Resumo Técnico do Censo da Educação Superior de 2019 (TEIXEIRA; ESTATÍSTICAS, 2019) é demonstrado que o número de alunos ingressantes no ano de 2019 corresponde a um aumento de 5,4% em relação ao ano anterior. Do total de 3.633.320 de ingressantes, 84,6% deles ingressaram em uma instituição privada e 15,4% em uma instituição pública. Ainda, o percentual de vagas novas ocupadas por ingressantes nos cursos de graduação mostra que o aproveitamento de vagas nas instituições públicas é de 80,9%, enquanto que o de vagas privadas é de 23,7%.

Sobre os concluintes entretanto, houve um decréscimo de 1,1% em relação a 2018 e, do total de 1.250.076 concluintes, 20,1% é da categoria pública e 79,9% da categoria privada indicando que apesar do aproveitamento de vagas ser maior na instituição pública, ainda apresenta um percentual de concluintes bem menor do que nas instituições privadas.

O Resumo Técnico também mostra, a partir de uma série de 2010 a 2019, que a taxa de desistência acumulada ao longo dos 10 anos é crescente, variando de 11% no primeiro ano a 59% no último ano sendo constantemente maior do que a taxa de conclusão acumulada que varia de 1% no primeiro ano e 40% no último ano.

Posto isto, segue dos princípios da LDB e dos indicadores do Resumo Técnico do Censo da Educação Superior a importância do estudo dos fatores que auxiliam no entendimento e tomada de atitude a fim de praticar ações que interceptam precocemente um aluno com perfil de possível evadido.

E, ainda, visto que a categoria de ensino - pública ou privada - possui diferença, em números, nos indicadores de ingressantes e concluintes existe a motivação do estudo focado nas instituições públicas mais precisamente na Universidade de Brasília (UnB) pois a UnB possui diferentes formas de ingresso. Como visto em estudos anteriores, a proporção de evadidos difere de acordo com a forma de ingresso do aluno (FRITSCH; ROCHA; VITELLI, 2015) uma vez que "a forma de ingresso pode levar à seleção de alunos com perfis diferentes e, por isso, com desempenhos diferentes"(CABELLO et al., 2021). Também se relacionando ao fato de que é tratada como uma instituição que requer alto desempenho acadêmico e os achados do estudo de Fritsch, Rocha e Vitelli (2015) mostram que a evasão e desempenho acadêmico são inversamente proporcionais.

Como consequência da importância do estudo da evasão escolar, inúmeros estudos investigam fatores que possam estar relacionados com a evasão e o delineamento do perfil desses estudantes com maior risco de evadir. Outrossim, a metodologia aplicada no estudo desses fatores também tem sua importância, assim como Ribeiro, Correia e Campos (2021) indicam em seu estudo: a metodologia adotada no estudo como o modo de coleta, tempo de medição, entre outros fatores, podem influenciar de forma significativa nos resultados.

Sendo assim, a metodologia aplicada no estudo de Ribeiro, Correia e Campos (2021) foi de cunho descritivo utilizando de análise de discurso dos grupos focais e do questionário aplicado. No estudo de (FRITSCH; ROCHA; VITELLI, 2015) a análise foi feita por meio de modelos de regressão logística. Nesse estudo, ainda, é comentado sobre o semestre do aluno indicando que alunos que estão mais próximos do fim do curso tem menor percentual de evasão do que aqueles que estão no início.

Portanto, como já comentado indiretamente pelo último estudo citado, além de estudar se o aluno evadiu ou não evadiu, pode-se também estudar o tempo até que essa evasão ocorra. Então, a aplicação da análise de sobrevivência se torna mais indicada no estudo de evasão já que é possível agregar as informações dos fatores que levam à evasão do aluno com o tempo até que essa evasão ocorra (COLOSIMO; GIOLO, 2006).

Assim sendo, o seguinte estudo tem como objetivo entender como se comporta o tempo até o aluno evadir e implementar um modelo de regressão que, com o auxílio

de variáveis explicativas, possa auxiliar na tomada de decisão da instituição para agir precocemente em alunos com perfis de possível evadido. Para a organização do banco de dados, entendimento das variáveis por meio da análise descritiva e a construção e avaliação do modelo será utilizado o *software* R v4.1.2.

2 Objetivos

Neste capítulo será definido os objetivos que guiarão a realização e conclusão do trabalho.

2.1 Objetivo Geral

Utilizar técnicas de Análise de Sobrevivência para a construção de um modelo de regressão que, com o auxílio de variáveis explicativas, facilite o entendimento do tempo até o aluno evadir e a tomada de decisão da instituição para ajudar os alunos que possam ser futuros evadidos.

2.2 Objetivos Específicos

- Estudar e aperfeiçoar o entendimento sobre o tema de evasão;
- Entender e organizar as informações presentes no banco de dados do estudo;
- Entender a metodologia da Análise de Sobrevivência;
- Estudar a implementação computacional pelo *software* R das técnicas utilizadas;
- Analisar descritivamente as possíveis variáveis do modelo e função de risco e sobrevivência;
- Elaborar um modelo de regressão que melhor se adéqua aos dados, buscando mostrar quais variáveis melhor auxiliam no entendimento da evasão escolar.

3 Revisão de Literatura

3.1 Evasão Escolar

A evasão na educação é compreendida pela saída antecipada do estudante, antes da conclusão do ciclo ou ano, por qualquer motivo. Contudo, o conceito da evasão se torna mais complexo devido aos fatores que podem estar relacionados ao que resulta a saída de um aluno assim como na compreensão entre os tipos de evasão que, se não diferenciadas, podem não deixar claro o objeto de estudo e superestimar as taxas de evasão no ensino superior (CHAGAS, 2019).

Lobo (2012) separa a evasão em três categorias:

- Evasão do curso: a evasão do curso pode ser entendida como a saída definitiva do aluno do seu curso de origem sem tê-lo concluído por qualquer motivo (CHAGAS, 2019): muda de curso, mas permanece na Instituição de Ensino Superior (IES), muda pra outro curso de outra IES ou abandona os estudos universitários (LOBO, 2012). Sendo essa categoria, portanto, a mais geral dentre as outras.
- Evasão da instituição: trata-se da evasão no qual o aluno deixa a instituição de ensino mas permanece no sistema de ensino superior. Nessa categoria, não são considerados os alunos que mudaram de curso mas permanecem na mesma IES (LOBO, 2012).
- Evasão do sistema de ensino superior: essa categoria de evasão consiste no aluno que abandona completamente o ensino superior e não se vincula a outra IES (CHAGAS, 2019). Essa, entre as demais, é a mais difícil de ser rastreada pois as informações disponíveis são reduzidas dado que o estudante abandona sem pedir transferência e não se submete a um novo processo seletivo (LOBO, 2012)(CHAGAS, 2019).

Lobo (2012) diz que em algumas instituições, no caso da evasão do curso, não consideram como evasão aqueles alunos que mudaram de curso mas continuam na mesma IES. Para a realização desse estudo a abordagem utilizada será a de evasão de curso considerando, inclusive, a mudança de curso dentro da Universidade de Brasília.

3.2 Análise de Sobrevivência

A análise de sobrevivência é uma área da estatística que utiliza técnicas para estudar dados relacionados ao tempo decorrido até a ocorrência de um evento de interesse.

Esse tempo é denominado tempo de falha, que corresponde à diferença que entre o tempo em que a falha ocorreu e o tempo em que se iniciou o estudo.

Para definir o tempo de falha, portanto, é preciso:

- fixar o tempo de início do estudo,
- escolher a escala de medida a ser utilizada e
- definir o evento de interesse.

Contudo, dentro de um estudo longitudinal é frequente que existam casos em que não seja possível observar a data de ocorrência do evento de interesse, resultando em dados parciais ou incompletos. Na análise de sobrevivência esses dados são chamados de censura.

Ressalta-se o fato de que, mesmo censurados, essas observações são importantes e devem ser utilizadas na análise estatística já que fornecem informações sobre o tempo de vida de objetos e indivíduos e a omissão das censuras pode acarretar em resultados viesados na análise estatística.

Portanto, a capacidade de incluir os dados de censura é a principal diferença entre a análise de sobrevivência e as técnicas tradicionais de estatística, além de reunir a informação longitudinal no estudo do evento de interesse (COLOSIMO; GIOLO, 2006).

Em vista disso, a censura pode ser classificada em 3 categorias (COLOSIMO; GIOLO, 2006):

- Censura à esquerda: ocorre quando o tempo registrado é maior que o tempo de falha, ou seja, o evento de interesse já ocorreu quando o indivíduo foi observado.
- Censura intervalar: neste caso os elementos tem acompanhamento periódico. O evento de interesse ocorre em um intervalo de tempo, isto é, $T \in (L, U]$.
- Censura à direita: o tempo de ocorrência do evento de interesse está à direita do tempo registrado. Nesse tipo de censura existem, ainda, três mecanismos:
  - Censura do Tipo I (b): o estudo termina após um período previamente estabelecido.
  - Censura do Tipo II (c): o término do estudo acontece após um número preestabelecido de falhas.
  - Censura aleatória (d): são todos os casos em que as observações não experimentam o evento de interesse por motivos não controláveis. Esse é o caso mais geral das censuras à direita.

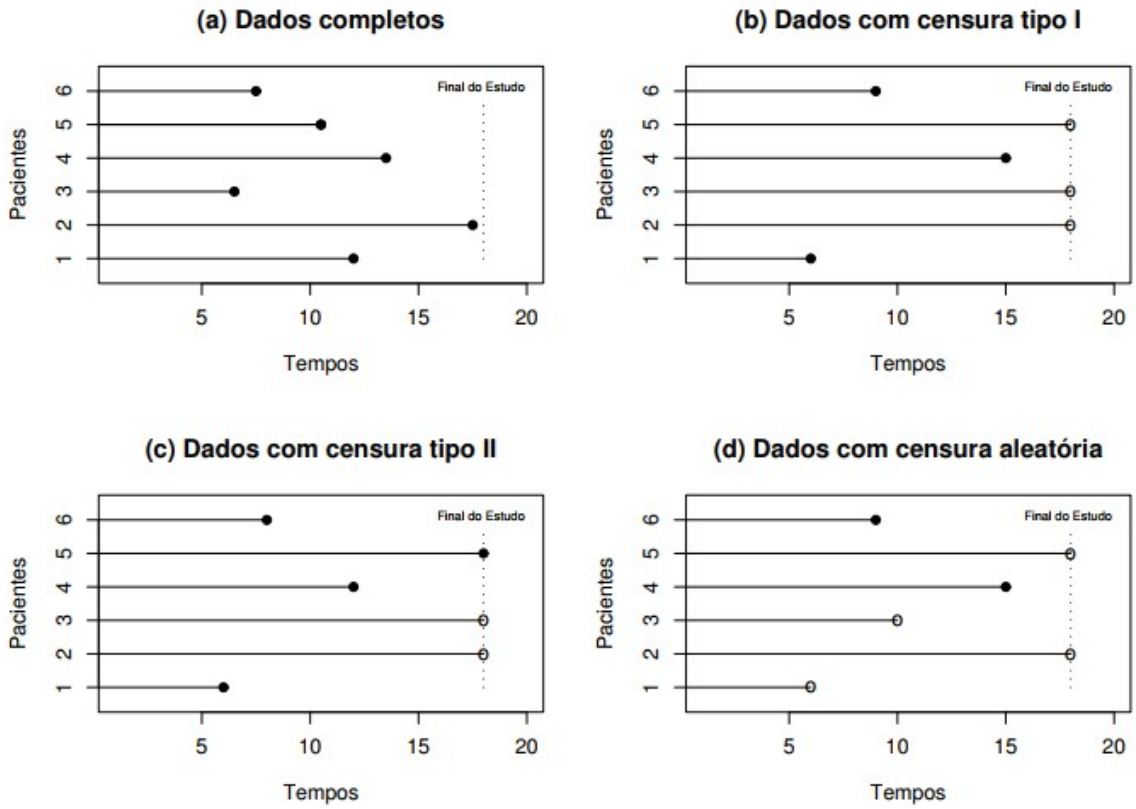

Figura 1: Ilustração de alguns mecanismos de censura em que • é a falha e ∘ é censura.(a) todos os pacientes experimentaram o evento de interesse antes do final do estudo. Fonte: Adaptado de Colosimo e Giolo (2006)

Como nesse trabalho os alunos que não evadiram até o fim do período de estudo serão considerados como uma censura, a categoria a ser utilizada será a de censura à direita utilizando o mecanismo de censura aleatória.

Ainda, outra característica dos estudos de sobrevivência é o truncamento. O truncamento, no entanto, é a exclusão de certos indivíduos ou objetos por alguma condição. Logo, os dados truncados diferem da censura pois não fazem parte da amostra.

Visto o conceito de falha e censura, os dados de sobrevivência para o indivíduo i ($i = 1,...,n$) são representados, em geral, pelo par (t_i,δ_i) sendo t_i o tempo de falha de falha ou de censura e δ_i a variável indicadora de falha ou censura, isto é,

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

Como em outras técnicas tradicionais da estatística tais como, regressão linear e análise de experimentos, a análise de sobrevivência conta também com o uso de outras informações correlacionadas à variável resposta, denominadas covariáveis que influenciam no comportamento do tempo de falha do i-ésimo indivíduo, por exemplo: $\mathbf{x}_i$ = (sexo,

idade, tipo de tratamento). Neste caso os dados ficam representados por $(t_i,\delta_i,\mathbf{x}_i)$ (COLOSIMO; GIOLO, 2006).

Nas próximas seções serão introduzidas as funções utilizadas no estudo de sobrevivência para representar a variável aleatória tempo T.

3.3 Função Densidade de Probabilidade

Seja T uma variável aleatória. Se existe uma função não negativa f definida para todos os reais $t \in (-\infty, \infty)$, então T é uma variável aleatória contínua (ROSS, 1976). Tendo isso em mente, para qualquer intervalo B de números reais,

$$P\{T \in B\} = \int_B f(t)dt. \tag{3.3.1}$$

De tal forma que f satisfaça

$$1 = P\{T \in (-\infty, \infty)\} = \int_{-\infty}^{\infty} f(t)dt. \tag{3.3.2}$$

A função f é chamada como função de densidade de probabilidade (f.d.p.) de uma variável aleatória contínua T. Em outras palavras, a Equação 3.3.1 define a probabilidade de T assumir valores no intervalo B.

3.4 Função de Sobrevivência

Seja $f(t)$ a f.d.p. de T, a função de distribuição acumulada (f.d.c.) que representa a probabilidade de um indivíduo não sobreviver a um tempo t, é

$$F(t) = P(T \leq t) = \int_t^{\infty} f(x)dx. \tag{3.4.1}$$

Sendo assim, a probabilidade de um indivíduo sobreviver a um tempo t é definida pela função de sobrevivência $S(t)$ expressa por:

$$S(t) = P(T \geq t) = \int_t^{\infty} f(x)dx = 1 - F(t) \tag{3.4.2}$$

sendo que $S(t)$ é uma função monótona decrescente e contínua com $S(0) = 1$ e $S(\infty) = lim_{t\to\infty}S(t) = 0$ podendo, em alguns casos, $S(\infty) > 0$ (LAWLESS, 2011).

3.5 Função de Risco ou Taxa de Falha

Dado que um indivíduo i sobreviva até um tempo t_i a função de risco é a probabilidade aproximada de um indivíduo falhar em $[t, t + \Delta t]$ (LAWLESS, 2011). A função de risco $h(t)$ é definida por:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{3.5.1}$$

A $h(t)$ pode, ainda, ser expressa em termos da f.d.p. e da função de sobrevivência,

$$h(t) = \frac{f(t)}{S(t)}. \tag{3.5.2}$$

3.5.1 Função de Taxa de Falha Acumulada

Outra função utilizada nos dados de sobrevivência é a função de risco acumulada que pode ser obtida através de $h(t)$, dada pela expressão

$$H(t) = \int_0^t h(u)du. \tag{3.5.3}$$

Considerando a relação entre $h(t)$, $S(t)$ e $f(t)$ é possível demonstrar que, outra forma de obter $H(t)$ é da seguinte forma:

$$H(t) = -[\log(S(t))]. \tag{3.5.4}$$

Essa função não possui uma interpretação direta como as outras funções $h(t)$ e $S(t)$. Mas pode ser útil na avaliação da função $h(t)$, especialmente na estimação não-paramétrica, já que $H(t)$ apresenta um estimador com boas propriedades e $h(t)$ é difícil de ser estimada (COLOSIMO; GIOLO, 2006).

Dito isso, existem diversas formas que a função de risco $h(t)$ pode assumir para cada modelo probabilístico, são elas: constante, crescente, decrescente, unimodal e forma de U. Logo, é importante utilizar uma metodologia que auxilia na identificação na distribuição de probabilidade que representa a variável aleatória tempo de sobrevivência T.

Com base nas características citadas da função de risco, o gráfico de $H(t)$ é um importante método para identificar qual o modelo probabilístico melhor representa a variável aleatória tempo.

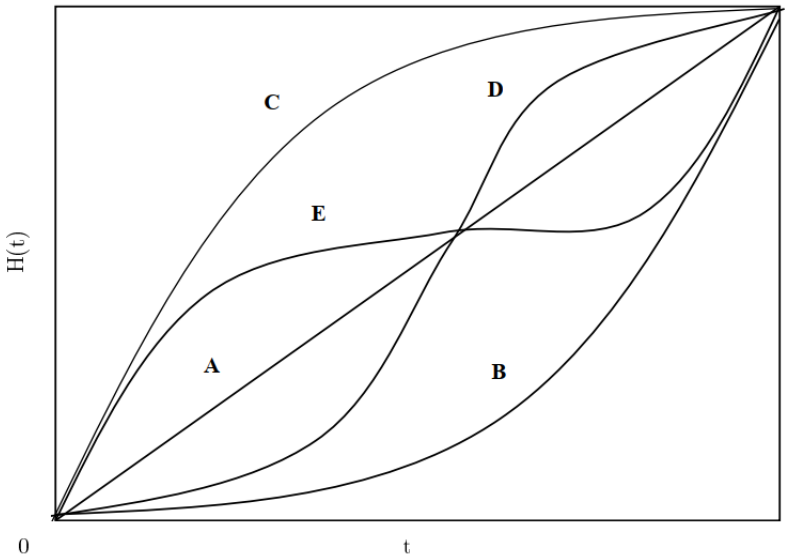

Figura 2: Ilustração de algumas formas da função de risco

Por meio da Figura 2 é possível identificar as formas da função de risco:

- reta diagonal (não necessariamente a reta y=x) (**A**) $\rightarrow \hat{h}(t)$ é constante
- curva convexa (**B**) $\rightarrow \hat{h}(t)$ é monotonicamente crescente
- curva côncava (**C**) $\rightarrow \hat{h}(t)$ é monotonicamente decrescente
- curva convexa e depois côncava (**D**) $\rightarrow \hat{h}(t)$ é unimodal
- curva côncava e depois convexa (**E**) $\rightarrow \hat{h}(t)$ tem forma de **U**

3.6 Gráfico do Tempo Total em Teste

Uma outra forma de investigar o possível comportamento de $\hat{h}(t)$ é usando a curva TTT proposta por Aarset (1987) que é obtida construindo um gráfico de

$$G(r/n) = \frac{[(\sum_{i=1}^{r} T_{i:n}) + (n-r)T_{r:n}]}{(\sum_{i=1}^{n} T_i)}, \tag{3.6.1}$$

por r/n, sendo que $r = 1, ..., n$, e $T_{r:n}, i = 1, ..., n$ são as estatísticas de ordem da amostra (ordenadas de forma crescente).

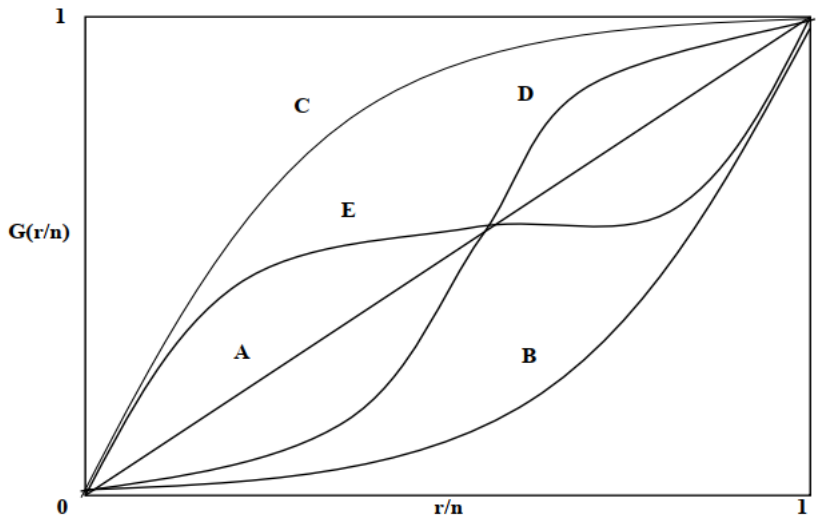

Figura 3: Ilustração de algumas formas da curva TTT

No entanto, a interpretação da curva TTT é o inverso do gráfico de $H(t)$, logo:

- reta diagonal (**A**) $\rightarrow \hat{h}(t)$ é constante
- curva convexa (**B**) $\rightarrow \hat{h}(t)$ é monotonicamente decrescente
- curva côncava (**C**) $\rightarrow \hat{h}(t)$ é monotonicamente crescente
- curva convexa e depois côncava (**D**) $\rightarrow \hat{h}(t)$ tem forma de **U**
- curva côncava e depois convexa (**E**) $\rightarrow \hat{h}(t)$ é unimodal

Também é importante notar que, ao contrário do gráfico da taxa de falha acumulada, a curva TTT não considera a presença de censuras. Sendo assim, na presença de muitas censuras é mais recomendado utilizar o gráfico de $H(t)$.

3.7 Estimação não-paramétrica

Um grande interesse em um estudo estatístico é gerar gráficos e estatísticas descritivas que auxiliem no entendimento inicial dos dados. Nos métodos tradicionais da estatística, a análise descritiva consiste essencialmente em encontrar medidas de tendência central e variabilidade (COLOSIMO; GIOLO, 2006) assim como tabelas de frequência e histogramas, função de distribuição empírica, gráficos de probabilidades e de densidade também são comuns entre diferentes ramos da estatística que, no caso de um estudo de sobrevivência, devido à presença de censura, se torna necessário modificar os métodos tradicionais (LAWLESS, 2011).

Portanto, seguindo da importância da função de sobrevivência e como as outras funções como taxa de falha e a f.d.p se relacionam, será mostrado nas próximas subseções métodos de estimação da função de sobrevivência.

3.7.1 Estimação na Ausência de Censura

No caso em que não há censura, a função de sobrevivência pode ser estimada como a proporção de indivíduos ou objetos que sobreviveram mais que um tempo t. Dada por:

$$\hat{S(t)} = \frac{\#\text{ de observações que não falharam até o tempo } t}{\#\text{ de observações no estudo}}. \tag{3.7.1}$$

De tal forma que $\hat{S}(t)$ é uma função escada com degraus nos tempos observados de falha de tamanho $1/n$, sendo n o tamanho da amostra.

3.7.2 Estimador de Kaplan-Meier

Na prática, são poucos os estudos de sobrevivência em que todas as observações falharam sendo necessária técnicas especializadas para acomodar as informações contidas nas censuras uma vez que a observação censurada informa que o tempo de falha é maior do que aquele que foi registrado (COLOSIMO; GIOLO, 2006).

O estimador mais utilizado em estudos clínicos e que vem ganhando mais espaço em estudos de confiabilidade é conhecido como estimador de Kaplan-Meier e também como estimador produto.

O estimador de Kaplan-Meier é definido por:

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right), \tag{3.7.2}$$

sendo:

- $t_1 < t_2 < ... < t_k$ os k tempos distintos e ordenados de falha que definem os intervalos de tempo,
- cada tempo de falha deve pertencer apenas a um intervalo,
- d_j o número de falhas em t_j, $j = 1, ..., k$, e
- n_j o número de indivíduos sob risco em t_j.

Sua confiabilidade é dada pelas suas propriedades:

- $\hat{S}(t)$ é um estimador não viesado,
- $\hat{S}(t)$ é fracamente consistente e
- Kaplan e Meier justificam a equação (3.7.2) mostrando que ela é o estimador de máxima verossimilhança de $S(t)$.

Alternativas ao estimador de Kaplan-Meier são os de Nelson-Aalen, sendo este mais recente, e a tábua de vida.

3.8 Distribuição de Probabilidade

Como comentado nas Seções 3.5.1 e 3.6, é importante considerar qual distribuição de probabilidade representa a variável aleatória tempo de sobrevivência. Algumas características da variável de tempo são essenciais para considerar um modelo probabilístico, são elas:

- A variável aleatória tempo de sobrevivência é contínua não negativa;
- Distribuição Normal não é adequada para modelar T;
- Frequentemente o tempo T apresenta forte assimetria.

Portanto, serão definidas algumas distribuições que são comumente usadas nos estudos de sobrevivência e que apresentam as características citadas acima.

3.8.1 Log-normal

Conforme é discutido em Lawless (2011), a distribuição log-normal têm sido usada para modelar aplicações nos campos de engenharia, medicina, entre outras áreas. Um tempo T é dito como log-normalmente distribuído se o $\log(T)$ é normalmente distribuído com média μ e variância σ^2.

A partir da f.d.p. da distribuição Normal pode-se facilmente demonstrar que a função de distribuição de probabilidade de uma Log-normal é escrita por:

$$f(t) = \frac{1}{(2\pi)^{1/2}\sigma t} \exp\left[-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^2\right], t > 0, \tag{3.8.1}$$

em que μ é a média do logaritmo do tempo de falha e σ é o desvio padrão.

A sua função de sobrevivência não tem forma explícita, sendo representada por:

$$S(t) = \Phi\left(\frac{-\log(t) + \mu}{\sigma}\right), \tag{3.8.2}$$

em que Φ é a função de distribuição acumulada de uma Normal padrão.

Assim como a S(t), a função de risco não possui forma explícita, sendo representada por:

$$h(t) = \frac{f(t)}{S(t)}. \tag{3.8.3}$$

3.8.2 Log-logística

Uma distribuição comumente utilizada para casos em que a forma da função de risco é unimodal é a log-logística. Se a variável aleatória T possui distribuição log-logística, sua f.d.p. é representada por:

$$f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1}(1 + (t/\alpha)^\gamma)^{-2}, \tag{3.8.4}$$

em que $t > 0$, $\alpha > 0$ é o parâmetro de escala e $\gamma > 0$ o parâmetro de forma.

Sua função de sobrevivência é:

$$S(t) = \frac{1}{1 + (t/\alpha)^\gamma}, \tag{3.8.5}$$

e sua função de risco:

$$h(t) = \frac{\gamma(t/\alpha)^{\gamma-1}}{\alpha[1 + (t/\alpha)^\gamma]}. \tag{3.8.6}$$

3.8.3 Log-logística discreta

Dado a natureza da mensuração dos dados desse trabalho serem em semestres, existe a hipótese de que uma distribuição discreta possa ser mais eficiente ao modelar o tempo.

Utilizando o método de discretização de distribuições contínuas aplicada na f.d.p. 3.8.4, a distribuição de probabilidade da variável aleatória T, representada como p(t), pode ser escrita como:

$$p(t) = \frac{1}{1+(t/\alpha)^{\gamma}} - \frac{1}{1+((t+1)/\alpha)^{\gamma}}, t = 0, 1, 2, ..., \tag{3.8.7}$$

em que t > 0, $\alpha > 0$ é o parâmetro de escala e $\gamma > 0$ é o parâmetro de forma.

Sua função de sobrevivência e função de risco são, respectivamente, escritas como:

$$S(t) = \frac{1}{1+((t+1)/\alpha)^{\gamma}}, t = 0, 1, 2, ..., \tag{3.8.8}$$

$$h(t) = 1 - \frac{1+(t/\alpha)^{\gamma}}{1+((t+1)/\alpha)^{\gamma}}, t = 0, 1, 2, ... \tag{3.8.9}$$

Essa distribuição é comumente utilizada quando a função de risco apresenta formas unimodais e decrescentes.

3.9 Método de Máxima Verossimilhança

Nos modelos probabilísticos apresentados as quantidades desconhecidas são denominadas como parâmetros, que devem ser estimados a partir das observações de uma amostra para que o modelo fique determinado e seja possível responder a perguntas de interesse (COLOSIMO; GIOLO, 2006).

O método de máxima verossimilhança (MLE) é a técnica mais popular na estimação dos parâmetros. O MLE é definido como: para cada observação da amostra, seja $\hat{\theta}$ um valor de parâmetro no qual $L(\theta)$ atinge seu máximo em função de θ, para uma observação fixa da amostra. Um MLE do parâmetro θ com base em uma amostra é $\hat{\theta}$ (CASELLA; BERGER, 2002).

Sendo assim, para uma distribuição com parâmetros α e γ o estimador de máxima verossimilhança escolhe o melhor par (γ, α) que deixe a amostra observada mais provável (COLOSIMO; GIOLO, 2006).

Sendo assim, a função de verossimilhança é dada por:

$$L(\theta) = \prod_{i=1}^{n} f(t_i, \theta)$$

Contudo, no estudo de sobrevivência, além do tempo de falha existe a contribuição das observações censuradas incorporada em S(t) podendo ser esquematizado da seguinte forma:

$$\begin{cases} f(t_i, \theta) & \text{se } t_i \text{ é tempo de falha} \\ S(t_i, \theta) & \text{se } t_i \text{ é tempo de censura.} \end{cases}$$

As observações podem então ser divididas em dois conjuntos, as r primeiras são as não-censuradas, e as $n - r$ seguintes são as censuradas (COLOSIMO; GIOLO, 2006). A função de verossimilhança considerando as censuras é descrita, portanto, por:

$$L(\theta) = \prod_{i=1}^{r} f(t_i, \theta) \prod_{i=r+1}^{n} S(t_i, \theta), \tag{3.9.1}$$

a função de verossimilhança também pode ser escrita em termos da função de taxa de falha devido sua relação com as outras funções. Portanto, segue da equação (3.9):

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} [f(t_i; \theta]^{\delta_i} [S(t_i; \theta]^{i-\delta_i} \\ &= \prod_{i=1}^{n} [h(t_i, \theta)]^{\delta_i} [S(t_i, \theta)], \end{aligned} \tag{3.9.2}$$

sendo δ_i a variável indicadora de falha.

Os estimadores de máxima verossimilhança são os valores de θ que maximizam $L(\theta)$ ou equivalentemente $\log(L(\theta))$. Esses estimadores podem ser encontrados resolvendo o sistema de equações a seguir

$$U(\theta) = \frac{\partial \log L(\theta)}{\partial \theta} = 0.$$

3.10 Critério de Informação de Akaike (AIC)

A seleção de modelos é importante para que seja utilizado o mais adequado. Seja comparando modelos com distribuições diferentes ou com variáveis explicativas diferentes.

O Critério de Informação de Akaike (AIC) é utilizado quando há a comparação de modelos vindo de um mesmo conjunto de dados para escolher um modelo que se ajusta bem aos dados, mas que não seja super parametrizado. Isto é, o modelo mais parcimonioso.

Sendo assim, o AIC pode ser escrito como:

$$AIC = -2 \log(L(\hat{\theta}) + 2p, \tag{3.10.1}$$

em que $\log(L(\hat{\theta})$ é o logaritmo da verossimilhança e p é o número de parâmetros estimados no modelo.

O modelo escolhido deve apresentar o menor valor de AIC dentre todos os modelos considerados.

3.11 Resíduos Cox-Snell

Após a construção do modelo final, é importante avaliar o quão bem esse modelo se ajustou aos dados que estão sendo estudados. Segundo Colosimo e Giolo (2006), a inspeção dos gráficos de resíduos é comumente utilizada para verificar a adequação do modelo.

Uma vez que os dados de sobrevivência possuem censura e estes, por sua vez, não possuem resíduos que seguem uma distribuição Normal e são assimétricos, o resíduo de Cox-Snell é o mais utilizado na verificação de adequação.

Segundo Cox e Snell (1968), esses resíduos auxiliam a examinar o ajuste global do modelo. Eles podem ser definidos como:

$$\hat{e_i} = \hat{H}(t_i|x_i) \text{ ou } \hat{e_i} = \hat{H}(y_i|x_i), \tag{3.11.1}$$

em que $\hat{H}(.)$ é a função de risco acumulada obtida no modelo e $\boldsymbol{x}$ o vetor de covariáveis do indivíduo i.

Os resíduos $\hat{e_i}$ vêm de uma população homogênea e devem seguir uma distribuição exponencial padrão. Sendo assim, quando o gráfico de $\hat{e_i}$ *versus* $\hat{H}(e_i)$ for aproximadamente uma reta para que o modelo exponencial seja adequado e, consequentemente, o modelo (LAWLESS, 2011).

Dado a equação 3.5.4, pode-se escrever $\hat{S}(e_i)$ em função de $\hat{H}(e_i)$. Sendo assim, o gráfico das curvas de sobrevivência desses resíduos, obtidas por Kaplan-Meier e pelo modelo da exponencial padrão, também auxiliam na verificação da qualidade do modelo ajustado.

4 Metodologia

4.1 Banco de dados original

Neste estudo será utilizado um banco de dados disponibilizado pela Secretaria de Tecnologia da Informação (STI) com o auxílio do Instituto de Exatas (IE) da Universidade de Brasília (UnB) com informações dos alunos acerca da sua performance durante o curso e outras intrínsecas a eles como, por exemplo: forma e data de ingresso, data de saída do curso, data de nascimento, CEP, dentre outras.

O banco de dados original para o curso de Licenciatura em Computação tinha 348679 observações e 27 variáveis. É importante ressaltar que os dados foram disponibilizados de forma que a informação do aluno não é única, visto que um mesmo aluno aparece repetidas vezes no banco de dados. Portanto, foi necessário um trabalho de filtro para as informações repetidas do aluno.

As variáveis presentes no banco de dados estão descritas a seguir:

1. Identificação do Aluno;
2. Identificação da Pessoa;
3. Índice de rendimento acadêmico (IRA);
4. Sexo;
5. Data de nascimento;
6. CEP;
7. Estado de nascimento;
8. Sistema de cotas (sim ou não);
9. Cota (qual a cota utilizada);
10. Escola (pública ou privada);
11. Chamada que ingressou na UnB;
12. Curso (Licenciatura em Computação);
13. Período de ingresso na UnB;
14. Período de ingresso no curso;

15. Forma de ingresso na Unb (PAS, Vestibular,...);
16. Período de saída do curso;
17. Forma de saída do curso (informação se o aluno se formou, se está ativo ou se mudou de curso);
18. Período que cursou a disciplina;
19. Média do aluno no semestre;
20. Mínimo de créditos para se formar;
21. Créditos total no período;
22. Total de créditos cursados pelo aluno;
23. Créditos aprovados no período;
24. Código da disciplina;
25. Nome da disciplina;
26. Créditos da disciplina;
27. Menção do aluno na disciplina;

4.2 Limpeza do banco de dados

Para o escopo do estudo foi feito dois filtros iniciais. Um para a variável de curso, visto que o objetivo do estudo é estudar o curso de Licenciatura em Computação. Outro para o período de ingresso no curso, sendo considerado os alunos que entraram no curso entre os períodos de 2012/2 a 2019/2. Foi considerado esse período pois ele conta com dois currículos vigentes no curso. Portanto, habilitaria a possibilidade de estudar o efeito do currículo na evasão do aluno, na qual a hipótese inicial é de que o currículo novo melhora a qualidade de vida dentro do curso e, portanto, aumenta a probabilidade do aluno não evadir.

Do banco de dados original foram descartadas algumas variáveis devido a incoerência ou erros de digitação ou que não foram cogitadas como informativas para o estudo, entre elas se encontram: chamada que ingressou na UnB, média do aluno no semestre, mínimo de créditos para se formar, créditos totais no período, total de créditos cursados pelo aluno, código da disciplina, nome da disciplina, ID do aluno e ID da pessoa.

4.3 Criação de variáveis

Para enriquecer a análise desse estudo, algumas variáveis foram criadas a partir de outras do banco original assim como outras que tiveram alterações em seus fatores.

4.3.1 Tempo

Uma das principais variáveis desse estudo por compor a variável resposta é a de tempo, que foi medida em semestres. Para essa variável foi considerado a diferença entre o ingresso do aluno no curso e o período de saída.

Para a construção do tempo, algumas condições tiveram de ser consideradas:

1. Os alunos que evadiram ou foram censurados durante o semestre de verão (semestre zero), foram realocados para o primeiro semestre do mesmo ano já que a Universidade de Brasília não reconhece o verão como semestre regular;

2. Para os alunos com o *status* de ativo na variável de período de saída do curso considerou-se esse período como 2020/1, pois foi o período máximo observado no banco de dados na variável de período de saída do curso. Esses alunos, portanto, teve sua variável de forma de saída sinalizada com o *status* ativo;

3. No caso dos alunos que evadiram no mesmo semestre que ingressaram não existe uma regra explícita de tratá-los como evasão no tempo 0 ou 1. Então, foi considerado a evasão no tempo 1 já que a evasão não acontece no momento exato 0 e o aluno chega a ter uma experiência dentro da UnB.

Sendo assim, a amplitude observada após a construção da variável de tempo foi de 1 a 16 semestres. A amplitude máxima, portanto, coincide com o tempo máximo de permanência no curso.

4.3.2 Variável de falha e censura

Para a variável indicadora de falha e censura utilizou-se as categorias disponíveis da forma de saída do curso. Foi considerado como censura os alunos que apresentaram estar ativos ou formados, já que estando ativo não foi visto a presença da evasão e, uma vez formado, não há a possibilidade de evadir. Foi considerado como falha os alunos que tiveram a sua saída por qualquer razão que não seja formatura ou esteja ativo.

A tabela abaixo mostra todas as formas de saída e a categoria alocada a elas como falha ou censura:

Tabela 1: Tabela das formas de saída e o *status* considerado pra falha ou censura

Forma de saída	*Status*
Ativo	Censura
Formatura	Censura
Deslig - não cumpriu condição	Falha
Deslig - abandono	Falha
Desligamento voluntário	Falha
Deslig - decisão judicial	Falha
Mudança de curso	Falha
Novo vestibular	Falha
Reprovou três vezes na mesma disciplina	Falha
Transferência	Falha

4.3.3 Distância da residência do aluno até a UnB

A construção dessa variável pode ser dividida em três etapas:

1. Tratamento da variável de CEP
2. Obtenção das informações de latitude e longitude da residência dos alunos
3. Cálculo da distância entre as geolocalizações dos alunos e da UnB

A variável de CEP originalmente estava no formato apenas com números, exemplo: 12345678. Portanto, o primeiro passo foi transformar esse formato para o seguinte: Distrito Federal 12345-678. Entretanto, um cuidado foi tomado ao especificar a unidade da federação (UF) pois alguns alunos tinham CEPs de outros estados, como Goiás e São Paulo.

A etapa intermediária de especificar a UF foi feita pois para utilizar a *Application Programming Interface* (API)[1] e recuperar as informações da geolocalização dos alunos era necessário especificar a UF pertencente ao CEP. Para isso, foi feita uma consulta manual das faixas de CEP que cada UF pertencia acessando o sítio dos Correios (2022).

Após a consulta e alteração do formato dos CEPs, utilizou-se a API do *Google Maps* (INC., 2022) no em um *script* feito em linguagem Python v.3.10 para obter a geolocalização de cada aluno.

[1] "As APIs são um conjunto de padrões que fazem parte de uma interface e que permitem a criação de plataformas de maneira mais simples e prática para desenvolvedores. A partir de APIs é possível criar softwares, aplicativos, programas e plataformas diversas". (TECHTUDO, 2020)

Após os resultados obtidos da API encontrou-se inconsistências para alguns alunos e foi feita uma nova consulta utilizando a biblioteca *RSelenium* (HARRISON; HARRISON, 2020) para acessar o sítio do MapaCEP (2022) e buscar novas informações da geolocalização dos alunos inconsistentes.

Feito isso, a busca da geolocalização foi feita diretamente pelo *Google Maps* e, após, utilizou-se da fórmula de Haversine para calcular a distância, em metros, da residência do aluno até a UnB.

Criada pelo Prof. James Inman, a fórmula de Haversine é uma equação frequentemente utilizada na navegação e criação de *Geographic Information System* (GIS) (UPADHYAY, 2016). A fórmula baseia-se no cálculo da distância entre dois pontos de uma esfera a partir das suas latitudes e longitudes. Portanto, seja r o raio da Terra, d a distância entre os dois pontos, ϕ_1 e ϕ_2 as latitudes dos dois pontos e λ_1 e λ_2 a longitude dos dois pontos. O ponto central de Haversine é dado pela fórmula:

$$haversin\left(\frac{d}{r}\right) = haversin(\phi_2 - \phi_1) + \cos\phi_1 \cos\phi_2 haversin(\lambda_2 - \lambda_1), \tag{4.3.1}$$

a função de Haversine aplicada acima para o ângulo central $\Theta = \frac{d}{r}$ e para as diferenças na latitude e longitude é:

$$hav(\theta) = \sin^2\left(\frac{\theta}{2}\right) = \frac{1 - \cos\theta}{2}. \tag{4.3.2}$$

Finalmente, para resolver a distância d aplica-se o haversine inverso para $h = hav(\Theta)$ ou usa-se a função arcsin:

$$\begin{aligned} d = &\ rarchav(h) = 2r\arcsin(\sqrt{h}) \\ d = &\ 2\arcsin\left(\sqrt{\sin^2\left(\tfrac{\phi_2-\phi_1}{2}\right) + \cos\phi_1\cos\phi_2\sin^2\left(\tfrac{\lambda_2-\lambda_1}{2}\right)}\right) \end{aligned} \tag{4.3.3}$$

(WIKIMEDIA, 2004)

4.3.4 Distância da residência do aluno até a UnB, em metros, categorizada

Após as etapas descritas na subseção 4.3.3, a categorização foi feita utilizando os quartis da variável numérica de distância. Resultando, portanto, em quatro categorias: até 7089,43 metros; de 7089,43 metros até 16552,48 metros; de 14044,27 metros até 23457,23 metros; maior que 23457,23 metros.

4.3.5 Quantidade de reprovações

A quantidade de reprovações é uma variável em que o interesse é contar quantas vezes o aluno i obteve menções: SR, II e MI nas disciplinas.

4.3.6 Soma de créditos reprovados

Na soma de créditos reprovados o interesse não está em quantas vezes o aluno obteve as menções: SR, II ou MI durante o seu curso, mas na soma de créditos que dessas disciplinas.

4.3.7 Quantidade de disciplinas cursadas

Foi considerado como uma disciplina cursada aquelas em que o aluno obteve menções: SR, II, MI, MM, MS ou SS. Após a definição das menções, verificou-se quantas vezes o aluno obteve uma dessas menções.

4.3.8 Soma de créditos cursados

O critério da disciplina cursada é o mesmo feito para construir a quantidade de disciplinas cursadas. Entretanto, o cálculo é feito pela soma da quantidade de créditos de cada disciplina.

4.3.9 Proporção de créditos reprovados

O objetivo dessa variável é levar em consideração o crédito que os alunos cursaram e a proporção desses créditos que eles reprovaram. Sendo assim, utilizando os critérios citados para construir a soma de créditos reprovados e cursados. A fórmula para encontrar a proporção de créditos reprovados para o aluno i é dada por:

$$\text{Proporção de créditos reprovados } = \frac{\text{Soma de créditos reprovados}}{\text{Soma de créditos cursados}}. \tag{4.3.4}$$

Essa variável varia de 0 a 1, visto que os créditos reprovados estão contidos nos créditos cursados.

4.3.10 Média de créditos cursados p/ semestre

Essa variável foi construída para medir quantos créditos, em média, o aluno i cursa por semestre até o fim do seu curso. Também foi considerado o critério em que disciplinas cursadas são aquelas com menções: SR, II, MI, MM, MS ou SS.

O banco de dados foi agrupado por aluno e pelo período que o aluno cursou a disciplina. Após, foi feita a divisão da soma de créditos cursados por semestre pela quantidade de semestres cursados. O cálculo é demonstrado abaixo:

$$\overline{Y} = \frac{\sum_i^n \sum_j^m x_{ij}}{n}, \tag{4.3.5}$$

em que $\overline{Y}$ é a média de créditos cursados por semestre e x é a quantidade de créditos da disciplina j do semestre i.

4.3.11 Quantidade de trancamentos

Foi considerado uma disciplina trancada para aqueles alunos que tiveram menção: TJ ou TR. Logo, verificou quantas vezes o aluno trancou uma disciplina.

4.3.12 Soma de créditos trancados

Considerando o critério para as disciplinas trancadas, foi feita a soma dos créditos de cada disciplina.

4.3.13 Diferença entre o período de entrada na UnB e no curso em semestres

Essa variável foi construída considerando a variável de período de ingresso na UnB e a de período de ingresso no curso. Ambas estavam no formato "anosemestre", portanto, foi feito um trabalho com expressões regulares para extrair o ano e o semestre. Como o formato do ano estava em quatro dígitos, extraiu-se, portanto, os quatro primeiros dígitos e foi considerado como ano. Logo, o quinto dígito é o semestre.

Como o objetivo era manter a unidade de medida igual à variável de tempo, utilizou-se o seguinte cálculo para construir a diferença.

Seja $AnoCurso$ o ano em que o aluno entrou no curso, $AnoUnB$ o ano que o aluno entrou na UnB, $SemestreCurso$ o semestre que o aluno entrou no curso, $SemestreUnB$ o semestre em que o aluno entrou na UnB e $\Delta UnBCurso$ a diferença entre o período de entrada na UnB e no curso, em semestres. Têm-se que:

$$\Delta UnBCurso = (AnoCurso - AnoUnB) \times 2 + (SemestreCurso - SemestreUnB), \quad (4.3.6)$$

sendo que a diferença entre o ano que o aluno entrou no curso e na UnB é multiplicado por 2 para transformar a diferença em anos para a diferença em semestres.

4.3.14 Cursou verão

Cursou verão é uma variável binária que indica se o aluno cursou alguma matéria durante o período de verão ou não. Foi utilizado o período que o aluno cursou a disciplina e utilizado da seguinte forma, sim se o aluno i cursou uma disciplina em que o semestre seja 0 e não caso contrário.

4.3.15 Idade em anos

O banco de dados original não disponibiliza a idade do aluno, mas disponibiliza a data de nascimento do aluno. Foi considerado quantos anos completos o aluno teria no semestre que ele ingressou no curso.

4.3.16 Currículo

Como o estudo considerou o período de ingresso no curso de 2012/2 até 2019/2, o curso de Licenciatura em Computação teve um currículo que iniciou em 2012/2 e durou até 2015/1 e outro que iniciou a partir de 2015/2. Considerando esses intervalos, a variável binária de currículo foi dividida em currículo velho, caso o aluno tenha ingressado em 2012/2 a 2015/1, e currículo novo, caso o aluno tenha ingressado a partir de 2015/2.

4.3.17 Reprovou durante os dois primeiros anos

A motivação para a construção dessa variável foi obtida do estudo do Chagas (2019). No qual destaca que as reprovações obtidas durante os dois primeiros anos exercem uma grande influência sobre a probabilidade de evasão para os recém ingressados.

Portanto, a variável binária foi construída considerando se o aluno obteve menções SR, II ou MI durante os 4 primeiros semestres a partir do período que ingressou no curso. Sendo sim, caso o aluno tenha reprovado e não caso contrário.

4.3.18 Forma de ingresso

A forma de ingresso do aluno na UnB era dividida originalmente em sete categorias, sendo que quatro delas não representavam 10% da amostra de alunos. Como mostra a tabela abaixo:

Tabela 2: Distribuição de frequência da forma de ingresso dos alunos

Forma de ingresso	Qtd de alunos	Percentual
Vestibular	316	43%
Programa de Avaliação Seriada	180	25%
Sisu-Sistema de Seleção Unificada	128	18%
Portador Diplom Curso Superior	61	8%
Enem UnB	27	4%
Transferência Obrigatória	10	1%
Transferência Facultativa	6	1%
Total	728	100%

Portanto, decidiu-se agrupar as categorias: Enem UnB, portador de diploma de curso superior e transferência facultativa e obrigatória em uma categoria chamada "outros". Portanto, as categorias ficaram divididas em: Vestibular, Programa de avaliação seriada (PAS), Sistema de seleção unificada (SISU) e Outros.

4.4 Consolidação do banco de dados

O processo final feito no banco de dados, após a criação de todas as informações necessárias para a análise, foi o de remoção de duplicatas. Inicialmente iria ser utilizada uma das duas variáveis de identificação: ID do aluno ou ID da pessoa. Entretanto, verificou-se que no banco de dados existiam informações idênticas para diferentes IDs de aluno e de pessoa, aproximadamente 513 registros de ID eram duplicados. Portanto, partiu-se da suposição que alunos não teriam todas as outras informações iguais uns aos outros, já que existiam informação de data de nascimento e CEP que são consideravelmente únicas para cada indivíduo.

Algumas variáveis utilizadas para o auxílio da construção das informações finais, como: latitude, longitude e cep, foram retiradas do banco de dados final por serem temporárias. Sendo assim, o banco de dados teve a dimensão de 728 observações e 22 variáveis, sendo elas:

1. IRA;

2. Sexo;
3. Sistema de cota;
4. Escola;
5. Forma de ingresso na UnB;
6. Quantidade de disciplinas reprovadas;
7. Quantidade de disciplinas cursadas;
8. Quantidade de trancamentos;
9. Média de créditos por semestre;
10. Diferença entre o período de entrada na UnB e no curso;
11. Cursou verão;
12. Idade;
13. Currículo;
14. Soma de créditos reprovados;
15. Soma de créditos cursados;
16. Soma de créditos trancados;
17. Proporção de créditos reprovados;
18. Indicador de falha e censura;
19. Tempo;
20. Distância da residência do aluno até a UnB, em metros, categorizada;
21. Distância da residência do aluno até a UnB, em metros;
22. Reprovou durante os dois primeiros anos.

4.5 Análise de dados

A análise de dados primeiramente será realizada por meio das estatísticas descritivas individuais para cada variável como: tabelas, gráficos de colunas, boxplots, curva de sobrevivência, estimativa de Kaplan-Meier.

Depois será feita uma análise de correlações entre as variáveis explicativas. A investigação se dá pelo fato de que variáveis fortemente associadas umas com as outras trazem problemas para a modelagem e não devem ser consideradas juntas.

4.6 Modelagem

Para a modelagem propriamente dita, será considerada as três distribuições apresentadas nas seções 3.8.1 a 3.8.3 a fim de verificar qual a que melhor se ajusta aos dados em estudo.

As etapas a serem consideradas para o processo de modelagem, portanto, serão:

1. Seleção da distribuição de probabilidade que melhor se ajusta aos dados;
2. Análise exploratórias das variáveis explicativas sozinhas no modelo de regressão;
3. Rodadas de seleção de variáveis;
4. Análise do modelo final com a presença das variáveis explicativas no modelo de regressão;

4.6.1 Divisão do banco de dados

A modelagem foi feita considerando três banco de dados: o completo, apenas com alunos do sexo masculino e outro apenas com alunos do sexo feminino. A divisão foi decidida devido a análise descritiva da variável de sexo e ao perceber que a sua mudança de comportamento enquanto está sozinha no modelo e com a presença de outras variáveis. Portanto, a análise consiste em verificar se ao considerar populações diferentes o conjunto de variáveis explicativas mudam.

4.7 Modelo de regressão

De maneira geral, os estudos de sobrevivência têm interesse em estudar quais fatores e como esses fatores influenciam na curva de sobrevivência dos indivíduos que estão em estudo. Portanto, o uso de regressão para a inclusão de covariáveis é necessária.

Portanto, seja $\boldsymbol{x}^T = (x_0, x_1, ..., x_p)$ um vetor formado por observações das $(p+1)$ variáveis regressoras. Ao utilizar uma função de ligação $g(.)$ pode-se conectar a variável resposta às variáveis explicativas. Pode-se definir que o vetor de parâmetros θ será estimado utilizando o vetor $\mathbf{x}$ para um conjunto de $(p+1)$ variáveis regressoras da forma:

$$\theta = g(\boldsymbol{x}^T\boldsymbol{\beta}), \tag{4.7.1}$$

em que $\boldsymbol{\beta}$ é o vetor de coeficientes da regressão.

Seja T uma variável aleatória com distribuição Log-normal, assim como definida na Equação 3.8.1. Ao utilizar o parâmetro μ como $\boldsymbol{x}^T\boldsymbol{\beta}$ e a função de ligação como sendo a função de identidade $I(.)$, têm-se que o modelo de regressão Log-normal pode ser definido por:

$$f(t|x) = \frac{1}{(2\pi)^{1/2}\sigma t}\exp\left[-\frac{1}{2}\left(\frac{log(t) - \boldsymbol{x}^T\boldsymbol{\beta}}{\sigma}\right)^2\right]. \tag{4.7.2}$$

A função de sobrevivência e de risco corresponde, respectivamente:

$$S(t) = \Phi\left(\frac{-\log(t) + \boldsymbol{x}^T\boldsymbol{\beta}}{\sigma}\right) \tag{4.7.3}$$

$$h(t) = \frac{f(t)}{S(t)} \tag{4.7.4}$$

Para estimar os parâmetros do modelo, será utilizado o método de máxima verossimilhança que foi exposto em 3.9.

Para implementar os modelos, análises estatísticas e calcular as estimativas será utilizado o *software* R v.4.1.2.

5 Resultados e discussões

5.1 Análise Descritiva

5.1.1 Variável de falha e censura

Para as observações de falha, foram considerado os alunos que tiveram como forma de saída do curso os tipos: mudança de curso, novo vestibular, transferência, ter reprovado 3 vezes na mesma disciplina obrigatória, desligamento por abandono, decisão judicial, desligamento voluntário ou não cumpriu a condição como evadidos.

Caso o aluno não apresente nenhum desses *status*, seja por estar ativo ou ter concluído a formatura, foram considerados como uma observação censurada.

Tabela 3: Distribuição de frequência dos alunos evadidos (falha) e não evadidos (censura)

Censura	Qtd de alunos	Percentual
Falha	376	52%
Censura	352	48%
Total	728	100%

No banco de dados em estudo, observa-se que 52% de alunos são evadidos contra 48% de censurados. Com o parecer dessa proporção de alunos evadidos somada com a motivação desse estudo nota-se a importância de entender os fatores relacionados a evasão e, para estudos futuros, ampliar a metodologia para outros cursos da UnB.

5.1.2 Variável de tempo

A variável de tempo foi construída para medir, em semestres, o tempo da entrada do curso até a sua falha ou censura.

Tabela 4: Medidas resumo da variável de tempo (em semestres)

Estatística	Valor
Média	5,29
Variância	8,38
Desvio Padrão	2,89
Coeficiente de Variação	0,55
Mínimo	1,00
Máximo	16,00

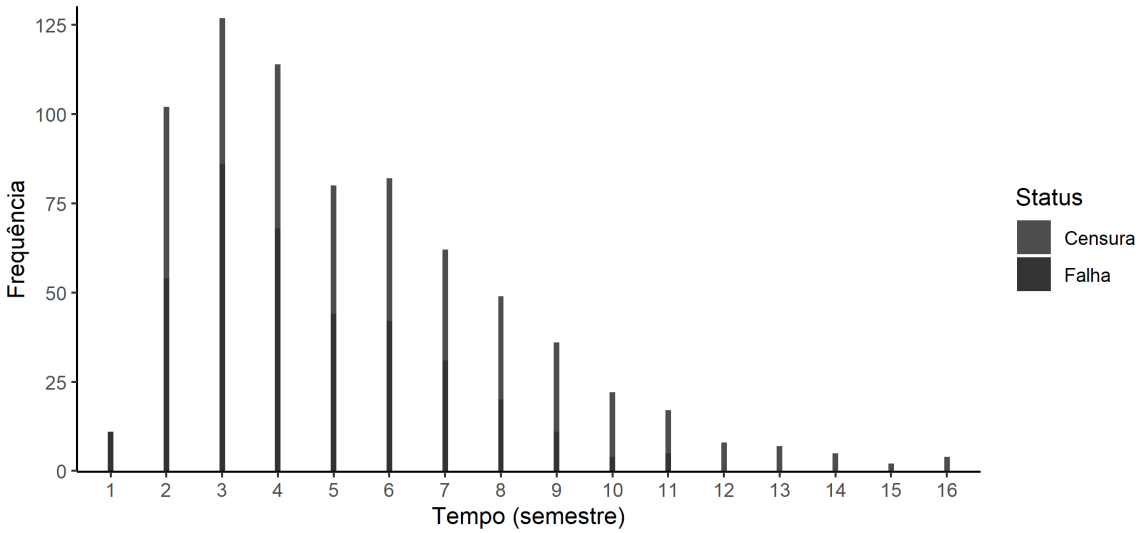

Figura 4: Gráfico de haste da distribuição de frequência da variável de tempo

Analisando a Figura 4, nota-se que há uma maior concentração de alunos entre o 2º e o 4º semestres do curso, tendo poucas observações no 1º pois apesar do estudo estar considerando o período de saída 2012/2 até 2020/1, não há registros de alunos que ingressaram 2020/1. Portanto, a frequência de alunos no 1º semestre são todos evadidos no mesmo semestre, sendo possível notar na Figura 5, já que a função de sobrevivência não é 1 no 1º semestre, pois já há evadidos no mesmo.

Analisando a Figura 5 pode-se analisar que as maiores diferenças na probabilidade de sobrevivência acontece no 3º e no 4º semestre, sendo que quanto mais no fim do curso, a função de sobrevivência tende a cair menos e a quantidade de censuras serem maiores. Também é perceptível que a partir do 11º semestre, não há registros de falhas.

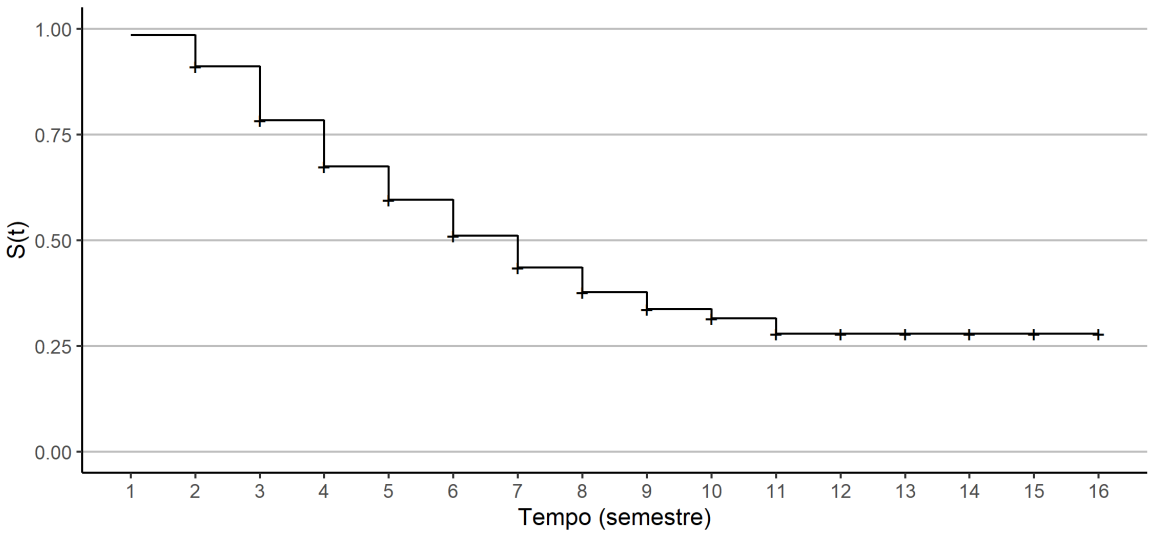

Figura 5: Gráfico de sobrevivência da estimativa de Kaplan-Meier

5.1.3 Sexo dos alunos

O curso de Licenciatura em Computação é majoritariamente composto por homens, sendo 88% dos alunos pesquisados, enquanto que 12% são mulheres.

Tabela 5: Distribuição de frequência do sexo dos alunos

Sexo	Qtd de alunos	Percentual
Masculino	644	88%
Feminino	84	12%
Total	728	100%

Contudo, na Figura 6 nota-se que as curvas de sobrevivência para os sexos não se distanciam bastante uma da outra. Porém, pode-se observar que há uma troca entre as curvas, ou seja, nos 7 primeiros semestres temos que as mulheres possuem uma probabilidade de sobreviver maior que os homens, enquanto que a partir do 8º semestre, os homens tendem a ter maior probabilidade de sobrevivência, o que também segue como motivação a separação dos bancos de dados descrita na subseção 4.6.1.

A Figura 6 também mostra que nos semestres 15 e 16, não há registros de mulheres.

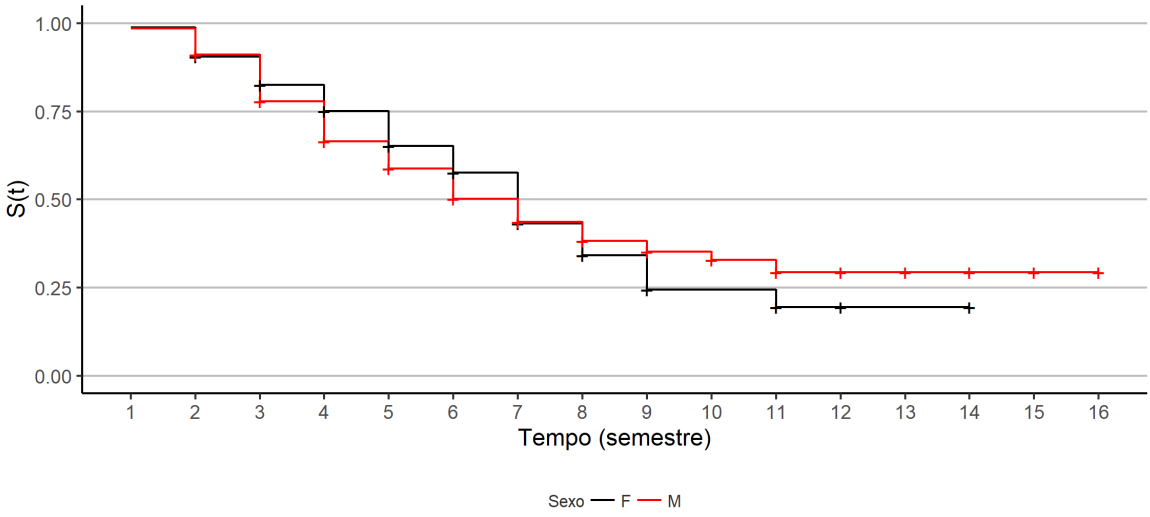

Figura 6: Gráfico de sobrevivência da estimativa de Kaplan-Meier para o sexo dos alunos

Na Figura 7 pode-se observar que a proporção de falha para o sexo masculino e feminino estão próximos dos 50%. Com isso, a análise descritiva não trás evidências de que o sexo esteja influenciando na curva de sobrevivência dos alunos.

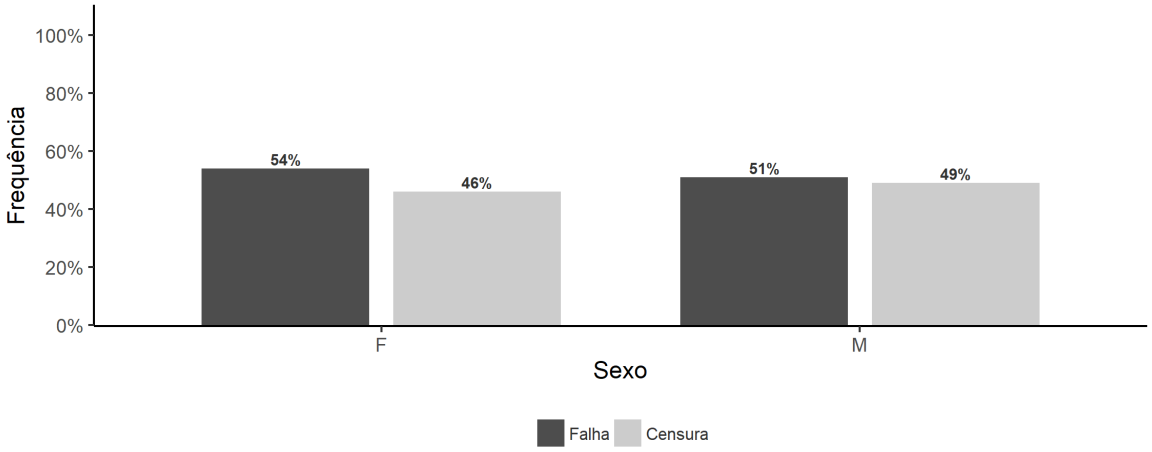

Figura 7: Gráfico de colunas da distribuição de frequência do sexo dos alunos por falha e censura

Entretanto, a variável de sexo está presente indicando que é importante para o estudo do tempo até a falha. Essa mudança pode estar relacionada a pouca informação de mulheres no estudo, tendo então uma motivação para uma nova análise com melhor distribuição dos sexos dos alunos.

A inquietude sobre essa variável também motivou a modelagem utilizando dois bancos de dados: um apenas com homens e outro apenas com mulheres. Trazendo variáveis diferentes para o seus modelos.

5.1.4 Idade

Como descrito anteriormente, essa variável considera a idade que o aluno tinha quando entrou no curso. Sendo assim, o perfil dos alunos têm média de 24 anos, aproximadamente, com um desvio padrão de 8 anos. Pelo coeficiente de variação, nota-se que a idade varia em 34% em torno da média, sendo um perfil levemente homogêneo, como é possível ver pela Tabela 6

Tabela 6: Medidas resumo de idade

Estatística	Valor
Média	24,27
Variância	67,47
Desvio Padrão	8,21
Coeficiente de Variação	0,34
Mínimo	16,00
Máximo	60,00

Pela Figura 8 nota-se que há um cruzamento entre os boxplots, não evidenciando que há uma diferença de idade entre os alunos que evadiram e os alunos que não evadiram. Entretanto, ainda é notável que os alunos que evadiram possuem uma média de idade maior e são mais dispersos.

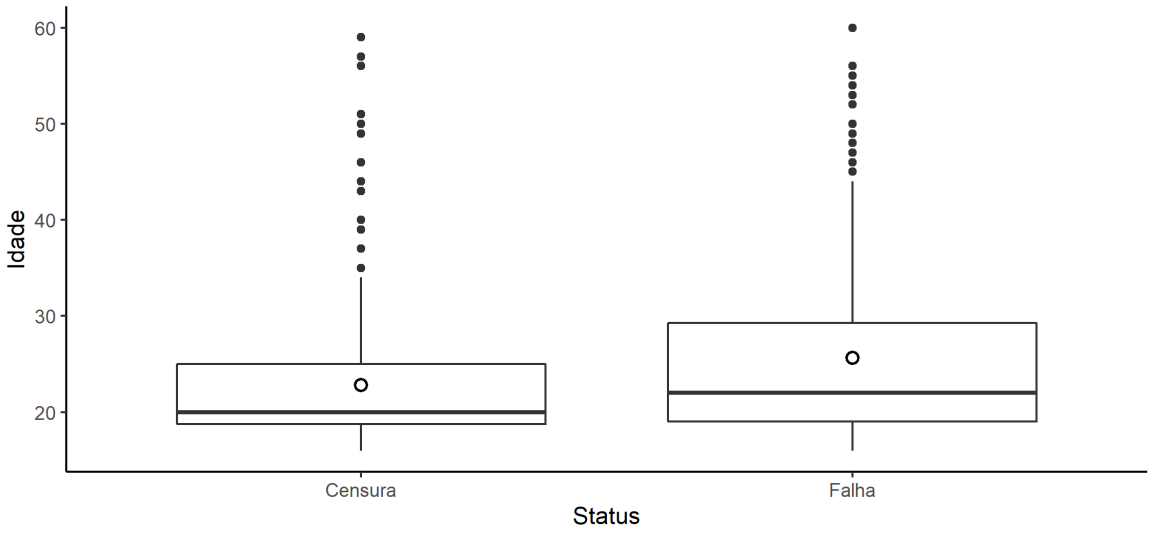

Figura 8: Boxplot da distribuição de idade por falha e censura

5.1.5 Tipo de escola

Os dados acerca do tipo de escola dos alunos se mostrou heterogêneo, sendo composto por 52% de alunos da escola particular e 48% da escola pública.

Tabela 7: Distribuição de frequência do tipo de escola

Escola	Qtd de alunos	Percentual
Particular	377	52%
Publica	351	48%
Total	728	100%

Nota-se também que a curva de sobrevivência para o tipo de escola mostra que há uma probabilidade de sobrevivência ligeiramente menor para os alunos que vieram da escola pública. Contudo, a Figura 5 mostra que as curvas não se distanciam de tal forma para obter evidências de que a diferença entre as funções de sobrevivência seja expressiva. Pode-se notar isso mais claramente na Figura 10 que mostra uma diferença de 4% na proporção de alunos evadidos para aqueles que vieram da escola pública.

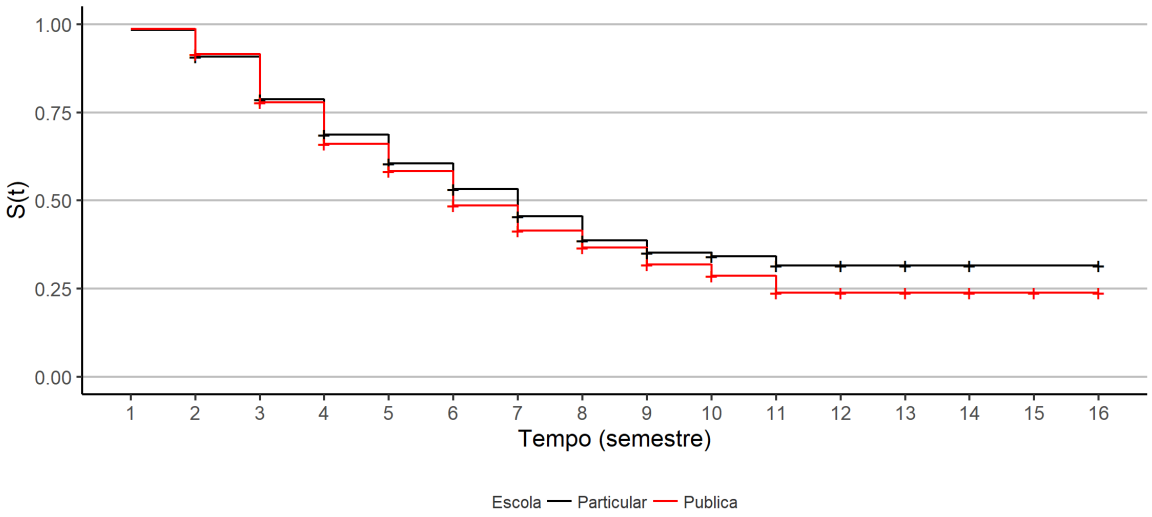

Figura 9: Gráfico de sobrevivência da estimativa de Kaplan-Meier para o tipo de escola

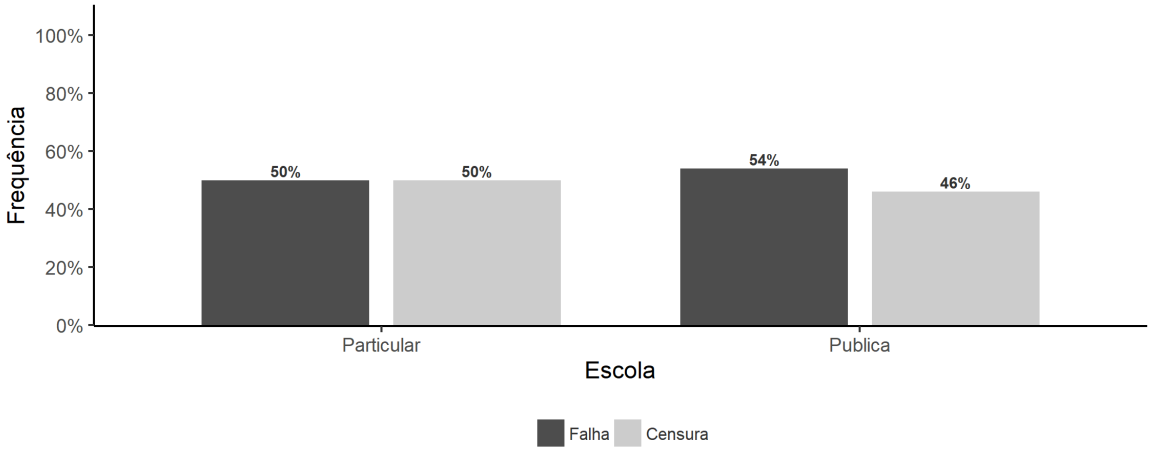

Figura 10: Gráfico de colunas da distribuição de frequência do tipo de escola por falha e censura

5.1.6 Sistema de cota

Analisando os alunos que ingressaram ou não por sistema de cota, 75% dos alunos não utilizaram o sistema de cota. Dos 25%, ainda foi observado durante a construção dessa variável que existem uma minoria de alunos de escola particular que utilizaram sistema de cota. Outro fato é que há alunos da escola pública que não utilizaram do sistema de cota para o ingresso no curso.

Tabela 8: Distribuição de frequência dos alunos que ingressaram utilizando sistema de cota ou não

Sistema de cota	Qtd de alunos	Percentual
Não	545	75%
Sim	183	25%
Total	728	100%

As curvas de sobrevivência, entretanto, não se distanciam, o que indica não ter evidências individualmente que ingressar por sistema de cota ou não influencia na sobrevivência do aluno à evasão, conforme mostrado também na Figura 12. Apesar do sistema de cota não apresentar evidências na influência da função de sobrevivência descritivamente, o modelo final manteve esta variável como significativa.

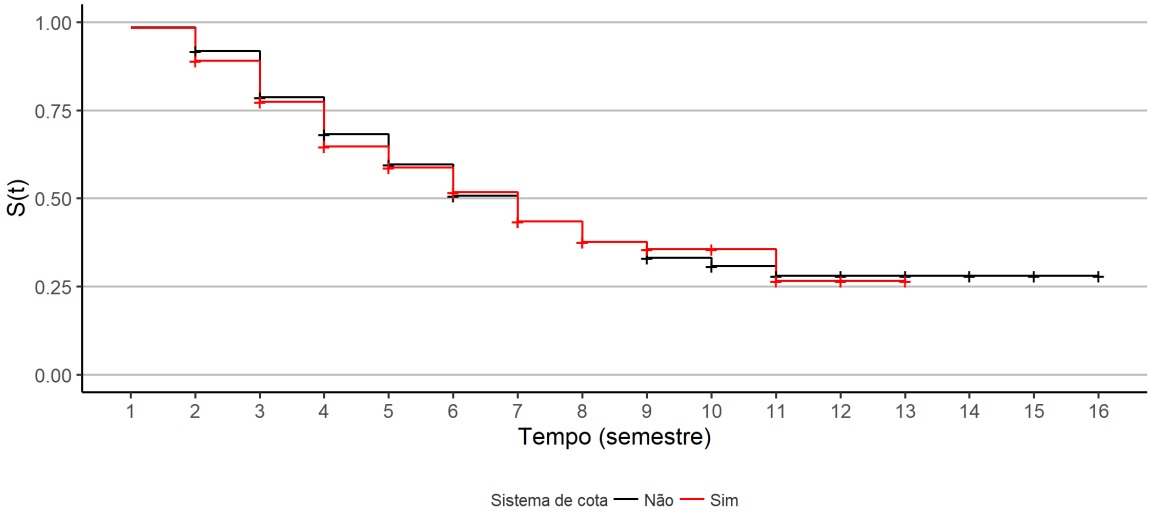

Figura 11: Gráfico de sobrevivência da estimativa de Kaplan-Meier para o sistema de cota

Ainda na Figura 11, nota-se que os alunos que ingressaram com sistema de cota ficam menos tempo no curso, uma vez que não há observações destes a partir do 13º semestre.

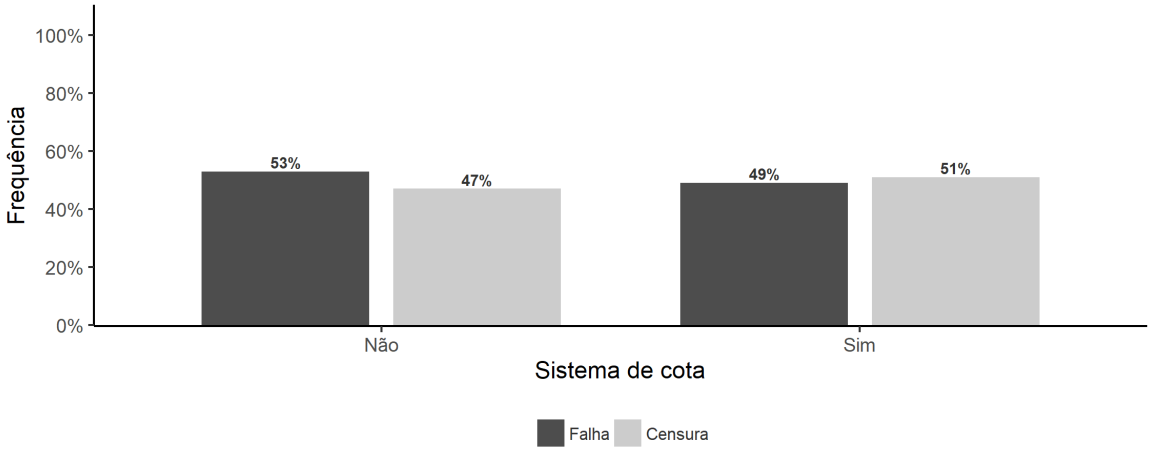

Figura 12: Gráfico de colunas da distribuição de frequência dos alunos que ingressaram com cota ou não por falha e censura

5.1.7 Forma de ingresso

Para a construção da variável de forma de ingresso na UnB, a categoria ”outros”considerou os alunos que ingressaram sendo portadores de diploma de curso superior, por transferência obrigatória ou facultativa e através do Enem UnB.

Tabela 9: Distribuição de frequência da forma de ingresso

Forma de ingresso	Qtd de alunos	Percentual
Vestibular	316	43%
Programa de Avaliação Seriada	180	25%
Sisu-Sistema de Seleção Unificada	128	18%
Outros	104	14%
Total	728	100%

Analisando a Tabela 9 nota-se que a maioria dos alunos ingressaram por meio do vestibular, 43%, seguido por aqueles que entraram pelo PAS (Programa de Avaliação Seriada) com 25% dos alunos, 18% sendo do Sisu e 14% ingressaram de outra forma.

Dessas formas de ingresso, percebe-se pela Figura 13 que os alunos que ingressaram pelo PAS têm maior probabilidade de sobrevivência, seguido por aqueles que ingressaram pelo vestibular que segue como segunda maior probabilidade de sobrevivência em mais da metade do tempo de estudo.

É interessante perceber que nessas duas formas de ingresso a única opção de universidade é a UnB, diferente de Sisu e transferências. Com isso, uma análise a ser

considerada é se os alunos que ingressaram pelo PAS ou vestibular optaram pela UnB antes de optar a forma de ingresso. Logo, uma análise futura é discorrer sobre quando a UnB foi opção para o aluno, ou seja, se foi a primeira opção ou não, ou até mesmo a única.

Outro ponto interessante na Figura 13 é que os alunos pertencentes a categoria de outras formas de ingresso possuem a menor probabilidade de sobreviver à evasão, disso pode-se perceber que parte da composição dessa categoria são alunos transferidos, ou seja, que já evadiram um curso anterior. Portanto, uma análise adicional seria a de verificar se o aluno já evadiu um curso anterior ou não.

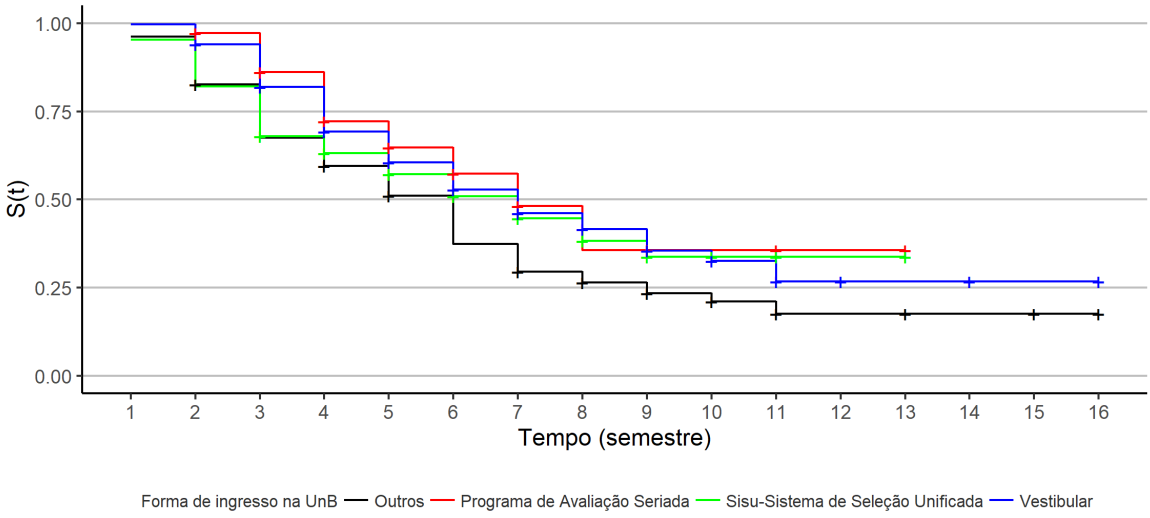

Figura 13: Gráfico de sobrevivência da estimativa de Kaplan-Meier para a forma de ingresso

Na Figura 13 entretanto, verifica-se que a distância entre as curvas só é perceptível entre os ingressos pelo PAS e ”outros”, como também mostra a Figura 14 já que existe uma diferença percentual de 31% de alunos evadidos na categoria de outras formas de ingresso.

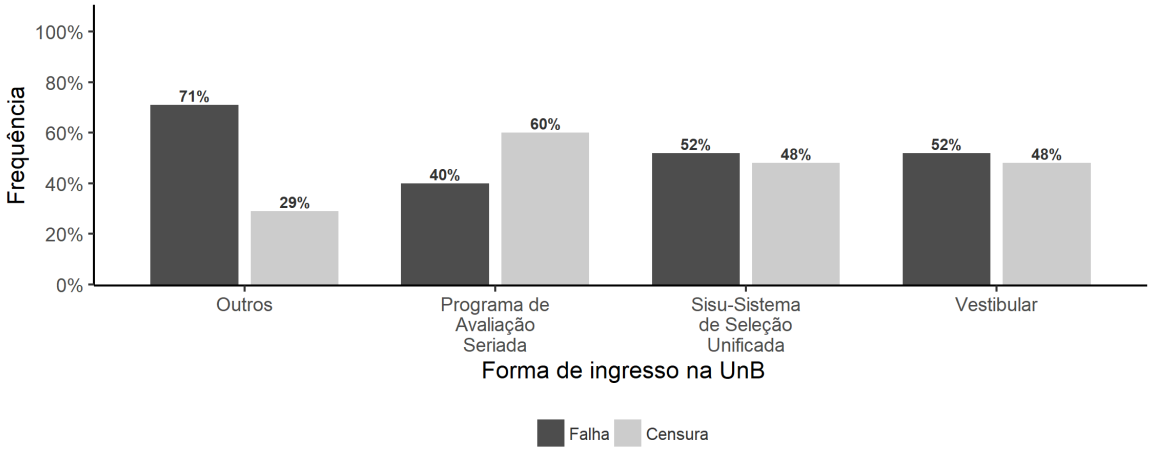

Figura 14: Censura vs forma de ingresso

5.1.8 Diferença entre o período de entrada na UnB e no curso (semestres)

Na Tabela 10 verifica-se que a média da diferença do período de entrada na UnB para o período de entrada no curso é de 0 semestres, ou seja, os alunos que entraram no curso de Licenciatura em Computação não estava em outro curso na UnB em algum período passado.

Entretanto, o coeficiente de variação mostra que os dados são altamente heterogêneos, dado que existe um máximo de 13 semestres de diferença, por exemplo, da entrada da UnB para a entrada do curso.

Na Figura 15 nota-se também que os boxplots apresentam o mesmo comportamento entre os alunos que evadiram e os que não evadiram. Ainda, existem *outliers* presentes nos dois boxplots para valores altos na diferença das entradas.

Tabela 10: Medidas resumo da diferença entre o período de entrada na UnB e no curso (semestres)

Estatística	Valor
Média	0,09
Variância	0,77
Desvio Padrão	0,88
Coeficiente de Variação	9,86
Mínimo	0,00
Máximo	13,00

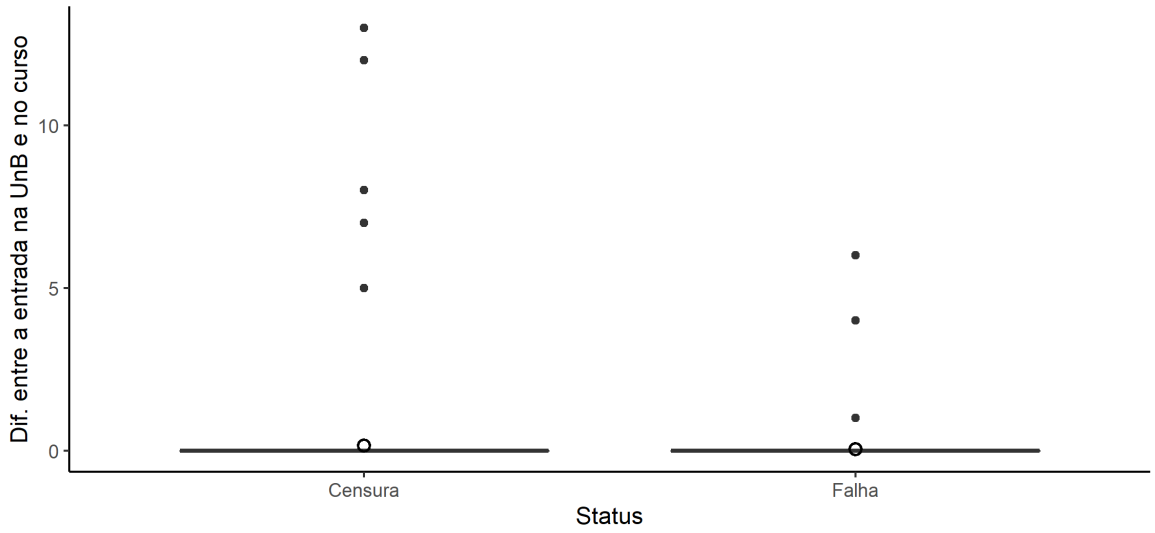

Figura 15: Boxplot da distribuição da diferença entre a entrada na UnB e no curso por falha e censura

5.1.9 Currículo vigente

Devido ao período de estudo considerado, existe a presença de dois currículos vigentes: um para aqueles que ingressaram entre 2012/2 e 2015/1, chamado de currículo velho, e outro para aqueles que ingressaram a partir de 2015/2, chamado de currículo novo. Além disso, parte da motivação para a separação dos currículos é que essa variável é a única do âmbito de infraestrutura acadêmica, ou seja, independe da característica do aluno.

Variáveis como essa podem trazer resultados importantes, como podendo ser visto em um dos estudos de motivação. Ribeiro, Correia e Campos (2021) analisam alguns itens dentro da categoria de infraestrutura acadêmica, por exemplo: grade do curso, conteúdo do curso e qualidade do currículo, que podem estar relacionados a qual currículo vigente do curso.

Tabela 11: Distribuição de frequência dos alunos por currículo vigente

Currículo	Qtd de alunos	Percentual
Novo	391	54%
Velho	337	46%
Total	728	100%

A Tabela 11 mostra uma diferença percentual de 8% entre os alunos que entraram no currículo novo e velho, representando uma diferença de 54 alunos. Ainda que exista essa diferença, em termos do tamanho da amostra pode-se considerar os dois grupos semelhantes.

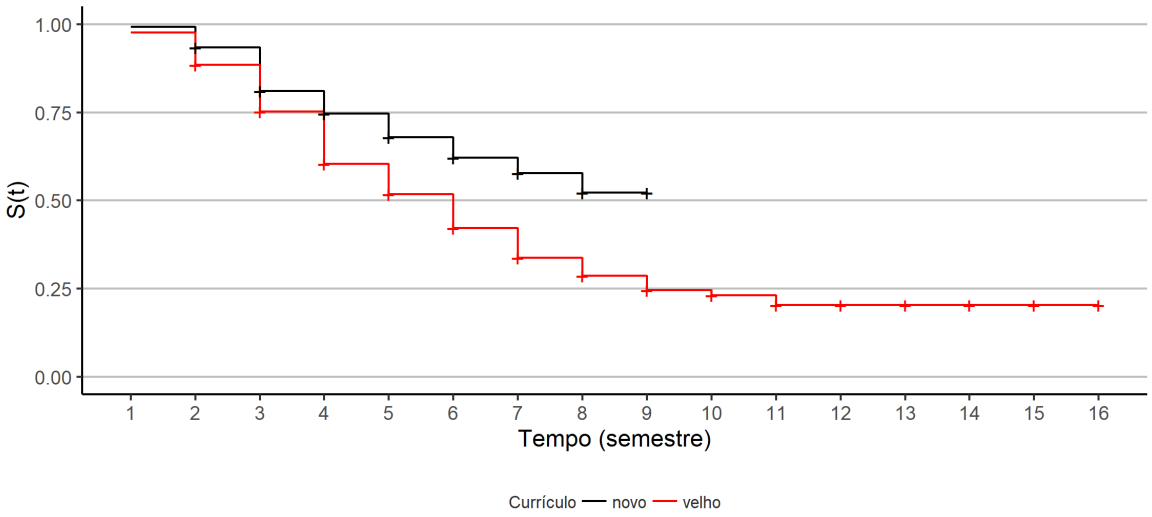

Figura 16: Gráfico de sobrevivência da estimativa de Kaplan-Meier por currículo vigente

Pela Figura 16 pode-se observar que os alunos pertencentes ao currículo novo tem uma probabilidade de sobrevivência maior que os alunos do currículo antigo. Informação que é novamente evidenciada pela Figura 17, em que se pode perceber que menos da metade dos alunos do currículo novo evadiram, enquanto que no currículo antigo 74% evadiram o curso.

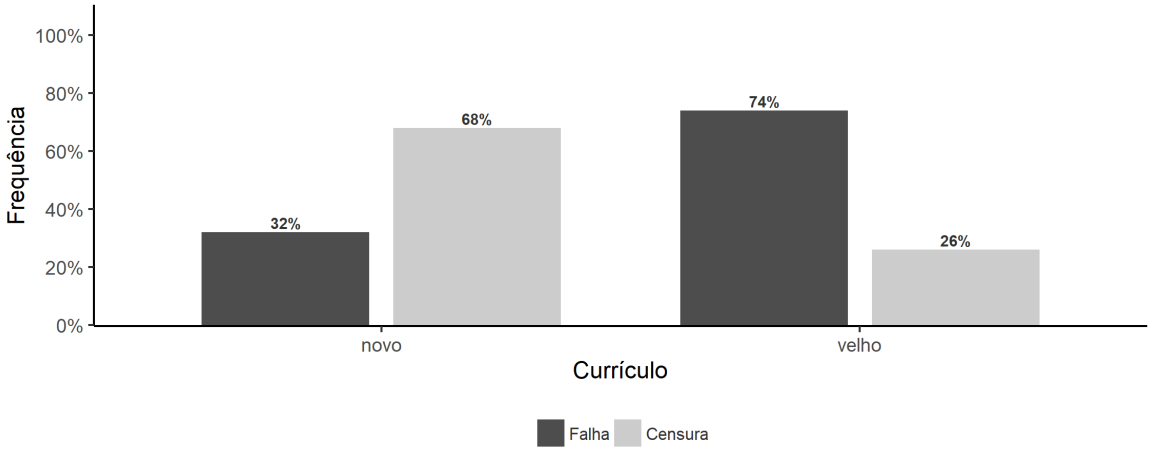

Figura 17: Gráfico de colunas da distribuição de currículo vigente por falha e censura

5.1.10 Cursou verão

Um artifício presente na UnB é a possibilidade de cursar disciplinas com duração menor durante o período de férias, chamadas disciplinas de verão. Na amostra, a minoria dos alunos optaram por cursar pelo menos uma disciplina de verão, enquanto que 76%

dos alunos nunca cursaram no verão.

Tabela 12: Distribuição de frequência dos que cursaram ou não disciplinas de verão

Cursou verão	Qtd de alunos	Percentual
Não	554	76%
Sim	174	24%
Total	728	100%

Assim como a diferença presente na proporção de alunos que optarem por cursar uma disciplina no verão, percebe-se pela Figura 18 que essa minoria apresenta evidências de que a probabilidade de sobreviver, ou seja, de não evadir, é constantemente maior que os alunos que não cursaram nenhuma disciplina durante esse período. Pode-se notar pela Figura 19 que 60% dos alunos que não cursaram alguma disciplina de verão evadiram, enquanto que um quarto dos alunos que cursaram evadiram.

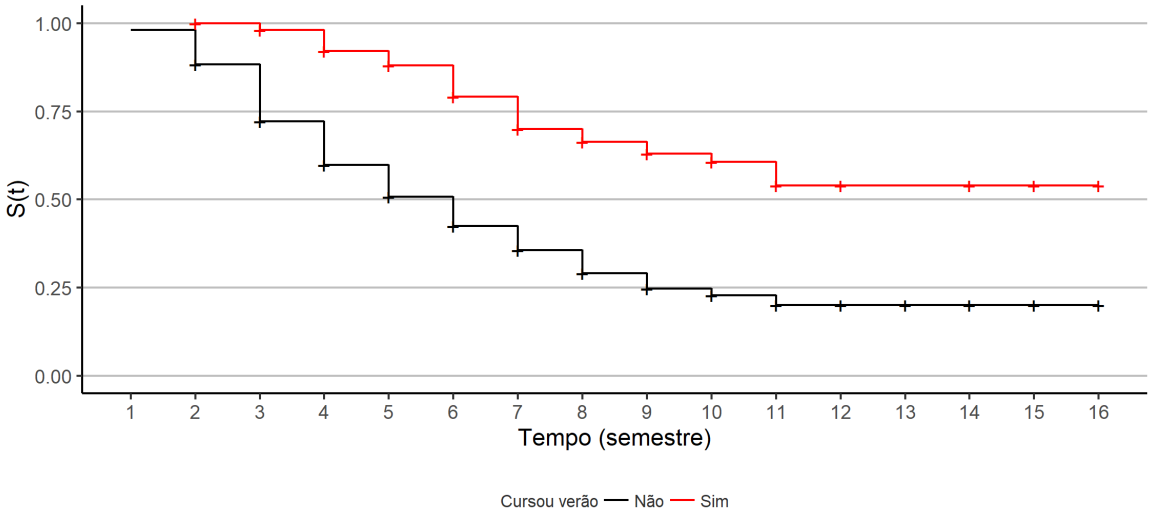

Figura 18: Gráfico de sobrevivência da estimativa de Kaplan-Meier por aluno que cursou ou não no verão

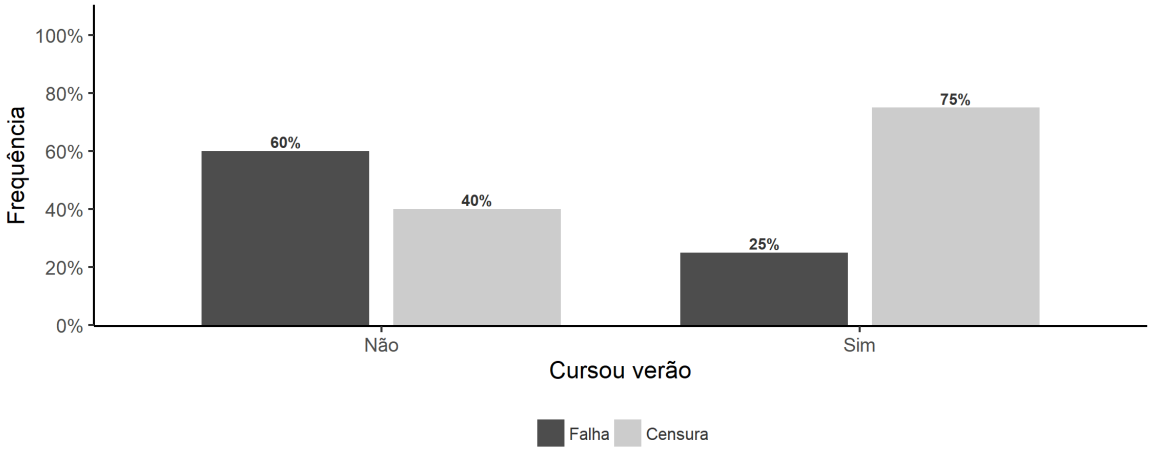

Figura 19: Gráfico de colunas da distribuição de alunos que fizeram ou não uma disciplina de verão por falha e censura

5.1.11 Distância da residência até a UnB (metros)

Supõe-se que a locomoção dos alunos de sua residência até a universidade pode ser um fator que influencie na permanência do aluno na universidade dado que o tempo e a dificuldade na locomoção possa inviabilizar o estudo do aluno.

A Seção 5.1.12 analisará categoricamente os quartis das distâncias, nesta seção a análise será feita na sua natureza quantitativa. Sendo assim, pela Tabela 13 têm-se que, em média, os alunos têm suas residências a 21820 metros de distância da UnB, aproximadamente. Essa variável é bem heterogênea, uma vez que existem alunos que moram no entorno de DF, como planaltina-GO, Valparaíso-GO, etc. Portanto, espera-se uma variância e um desvio padrão grande nos dados.

Outro ponto observado nos dados é que poucos alunos permanecerem com o cep de São Paulo, que possam ter viesado as medidas e os resultados dessa variável no modelo, além de ser razoável admitir erros de aproximação ao obter as latitudes e longitudes e construir o cálculo usando Haversine.

Tabela 13: Medidas resumo da distância da residência até a UnB (metros)

Estatística	Valor
Média	21820,29
Variância	3009069062,83
Desvio Padrão	54854,98
Coeficiente de Variação	2,51
Mínimo	528,44
Máximo	889987,40

Ainda, analisando a Figura 20 pode-se observar que os dois boxplots apresentam os mesmos comportamentos, uma vez que os valores possuem o mesmo nível de grandeza e os *outliers* causados pelos alunos que permanecem com o cep fora do DF causem grande assimetria nos dados.

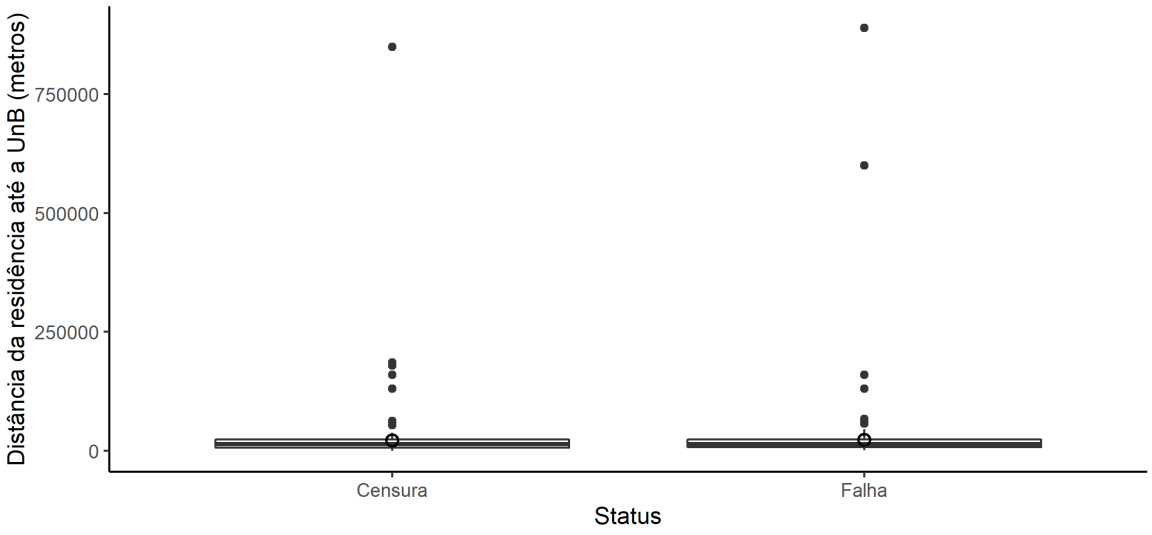

Figura 20: Boxplot da distribuição da distância da residência até a UnB por falha e censura

5.1.12 Distância da residência até a UnB categorizada (metros)

Essa subseção consiste em analisar categoricamente os quartis da variável de distância da residência do aluno até a UnB, minimizando possíveis erros de aproximação da API utilizada. Percebe-se pela Tabela 14 que cada categoria corresponde a 25% dos alunos devido a sua construção.

Tabela 14: Distribuição de frequência dos alunos por distância da residência até a UnB

Distância (metros)	Qtd de alunos	Percentual
Até 7089,43	182	25%
De 7089,43 até 16552,58	182	25%
De 14044,27 até 23457,23	182	25%
Maior que 23457,23	182	25%
Total	728	100%

Analisando as curvas de sobrevivência presentes na Figura 21, observa-se que as curvas estão próximas para todas as distâncias, logo, não há evidências de que há diferença na sobrevivência do aluno para diferentes distâncias da sua residência até a UnB. Ainda, a Figura 22 mostra que para todas as categorias cerca de 50% dos alunos evadiram.

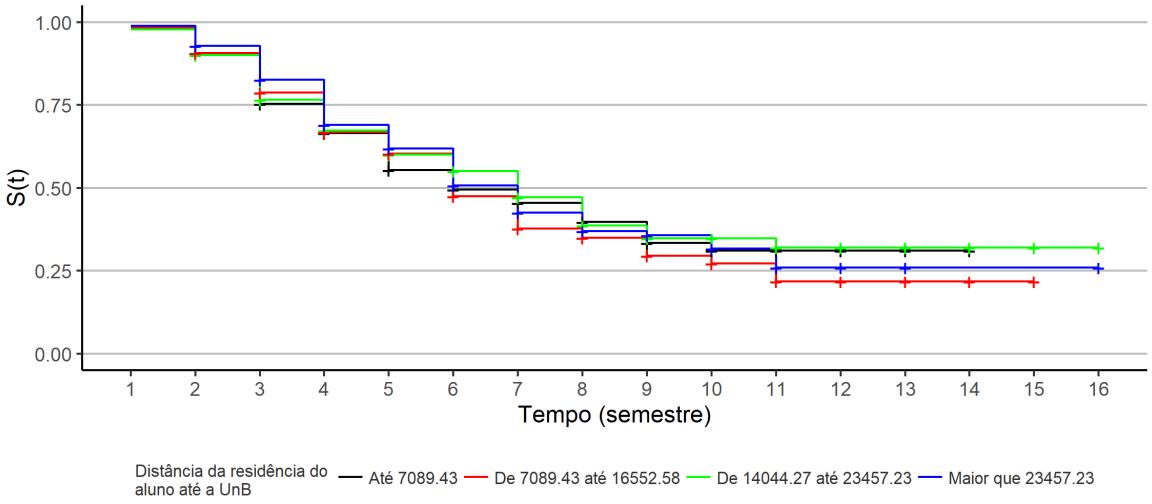

Figura 21: Gráfico de sobrevivência da estimativa de Kaplan-Meier por distância da residência do aluno até a UnB

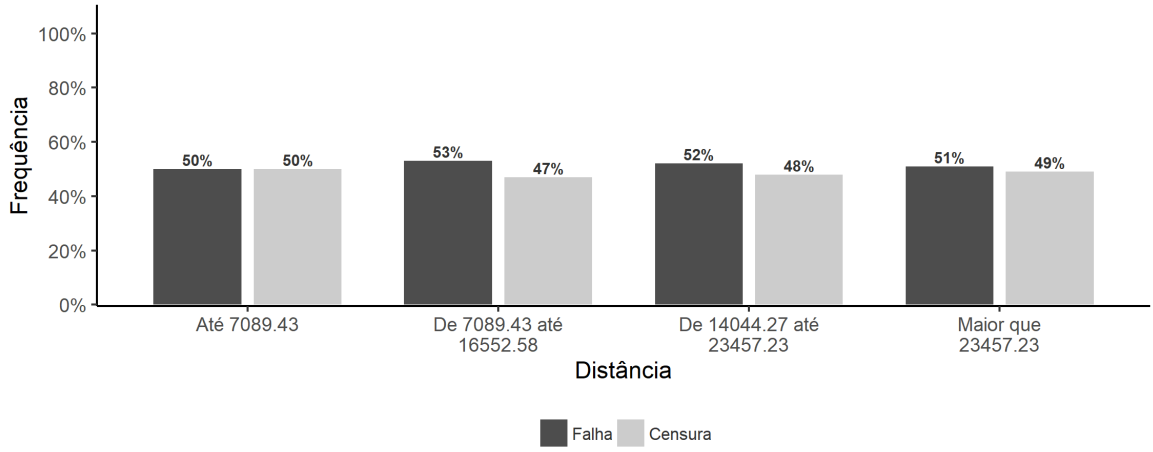

Figura 22: Censura vs distância categorizada

5.1.13 IRA

A variável IRA mede o índice de rendimento acadêmico do aluno, tendo seu cálculo construído pela própria Universidade de Brasília considerando a menção obtida por ele ao cursar uma disciplina, assim como sua carga horária e natureza (disciplina obrigatória ou optativa) podendo variar de 0 a 5.

Tabela 15: Medidas resumo do índice de rendimento acadêmico (IRA)

Estatística	Valor
Média	2,34
Variância	1,54
Desvio Padrão	1,24
Coeficiente de Variação	0,53
Mínimo	0,00
Máximo	5,00

A média de índice de rendimento dos alunos em estudo é de 2,34, ficando na metade da escala possível no IRA. Nota-se que nos dados existe o valor mínimo de 0 e o máximo de 5, podendo explicar a alta dispersão dos dados, mostrado pelo coeficiente de variação de 53%.

O IRA pode ser considerado como uma das variáveis mais importantes para a mensuração do desempenho acadêmico de um aluno na UnB. Portanto, um IRA com valores altos medem um desempenho melhor. Sendo assim, a Figura 23 explicita que os alunos que apresentaram evasão do curso têm seu IRA com valores mais baixos, tendo a

sua média próxima de 2 e uma dispersão maior do índice de rendimento acadêmico.

Os alunos que não evadiram têm seu IRA médio próximo de 3 e apresentam uma assimetria à esquerda, podendo ser notada pelos *outliers* presentes no boxplot.

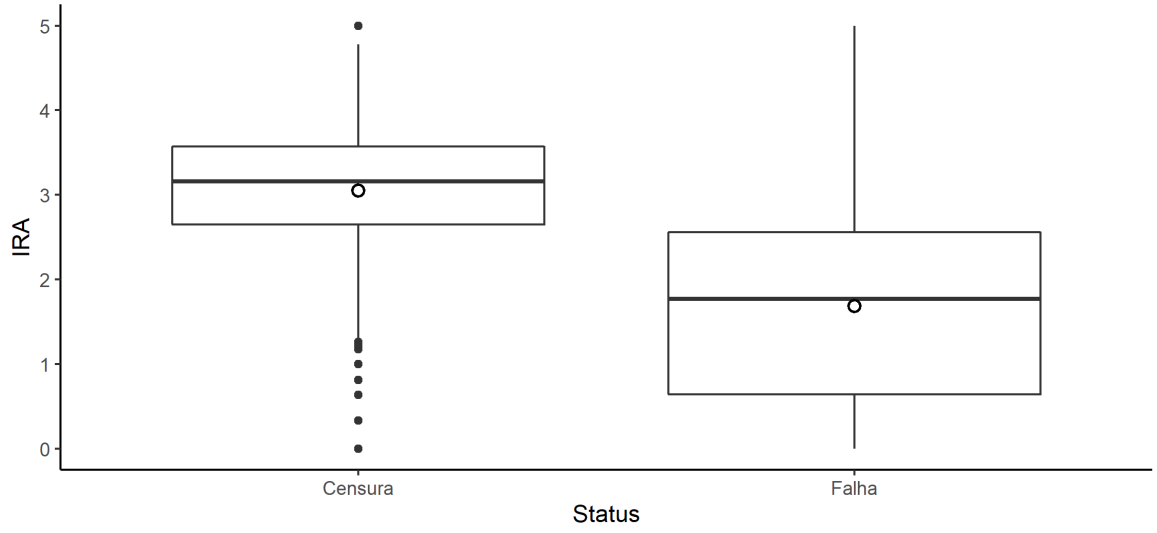

Figura 23: Boxplot da distribuição de IRA por falha e censura

5.1.14 Quantidade de reprovações

Nota-se que a distribuição da quantidade de reprovações dos alunos em estudo é bastante heterogênea, tendo seu coeficiente de variação de 81% tendo alunos que, no máximo, reprovaram 26 disciplinas até o fim do seu período no curso. Ainda, pode-se observar que os alunos em estudo reprovam, em média, aproximadamente 4 disciplinas durante todo o seu período no curso de Licenciatura em Computação.

Tabela 16: Medidas resumo da quantidade de reprovações

Estatística	Valor
Média	4,34
Variância	12,21
Desvio Padrão	3,49
Coeficiente de Variação	0,81
Mínimo	0,00
Máximo	26,00

Apesar que a quantidade de disciplinas reprovadas também está relacionado ao rendimento acadêmico do aluno no seu curso de graduação, a Figura 24 mostra que tanto para os alunos que evadiram quanto para os alunos que não evadiram a distribuição da

quantidade de disciplinas reprovadas se mantêm próximas, tendo em ambas as categorias *outliers* para valores mais altos - mostrando que poucos alunos reprovam mais do que 12 vezes, aproximadamente.

Pode-se notar também que as médias dos dois boxplots estão próximas, apesar dos alunos que não evadiram terem média um pouco menor que os alunos que evadiram. Logo, não há evidências de que a quantidade de reprovações esteja relacionada ao aluno evadir ou não do curso.

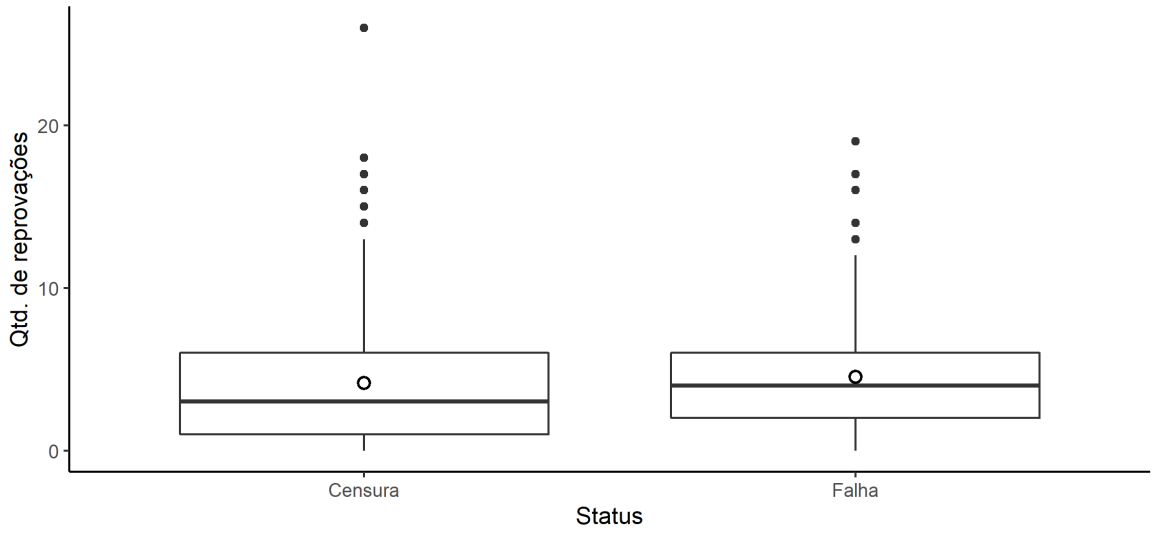

Figura 24: Boxplot da distribuição de quantidade de disciplinas reprovadas por falha e censura

5.1.15 Soma de créditos reprovados

A decisão tomada por se analisar não apenas a quantidade de disciplinas mas sim os créditos deu-se pelo fato de que disciplinas podem ter carga horária diferente uma das outras, resultando em diferentes esforços e impacto na grade horária semestral do aluno. Portanto, para essa seção e para as Seções 5.1.18 e 5.1.22 mudou-se o ponto de vista da análise de quantidade de disciplinas para os créditos dessas disciplinas. Cada disciplina pode variar entre 2, 4 e 6 créditos.

Pela Tabela 17 pode-se analisar que os dados são heterogêneos, variando cerca de 78% em torno da média, que é de aproximadamente 20 créditos reprovados durante todo o período que o aluno esteve no curso. Pela diferença entre a média e o seu máximo de 110 créditos reprovados, pode-se perceber que os dados são assimétricos a direita, assimetria que pode ser observada nos boxplots da Figura 25.

Tabela 17: Medidas resumo da soma de créditos reprovados

Estatística	Valor
Média	19,80
Variância	241,56
Desvio Padrão	15,54
Coeficiente de Variação	0,78
Mínimo	0,00
Máximo	110,00

Na Figura 25 pode-se observar que os alunos que não evadiram o curso possuem uma distribuição mais dispersa e média de créditos reprovados ligeiramente menor que os alunos que evadiram, ainda que os boxplots em suas distribuições estejam próximos um do outro.

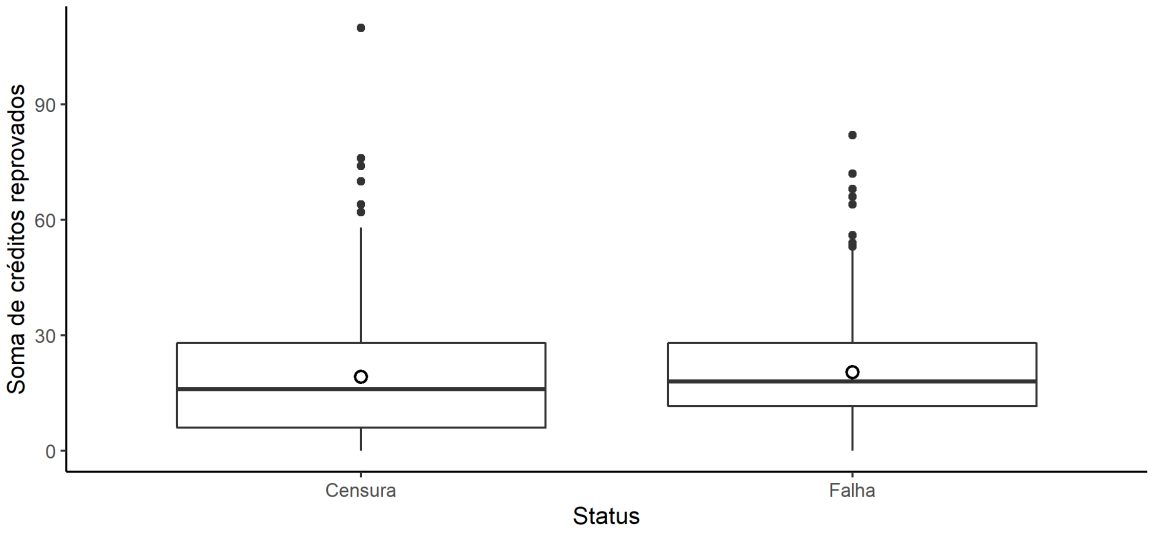

Figura 25: Boxplot da distribuição da soma de créditos reprovados por falha e censura

5.1.16 Reprovou durante os dois primeiros anos

Analisando a Tabela 18 têm-se que 90% dos alunos reprovaram alguma disciplina durante os dois primeiros anos. Os dois primeiros anos é comumente conhecido como sendo os anos com matérias da ”formação básica”no curso de Licenciatura em Computação.

Relembrando a análise da Figura 5, nota-se que as maiores diferenças na curva de sobrevivência acontecem nos quatro primeiros semestres, que correspondem aos dois primeiros anos. A partir disso, cria-se a suposição de que essas matérias sejam determinantes para o aluno continuar sua graduação.

Tabela 18: Distribuição de frequência dos alunos que reprovaram durante os dois primeiros anos ou não

Reprovou durante os dois primeiros anos	Qtd de alunos	Percentual
Sim	653	90%
Não	75	10%
Total	728	100%

Tendo em vista a importância das matérias ditas como "formação básica"do curso, a análise da Figura 26, para os primeiros sete semestres, pode-se analisar que os alunos que reprovaram alguma disciplina durante seus dois primeiros anos possuem uma probabilidade de sobrevivência menor do que os alunos que não reprovaram.

É interessante notar que após o 7º semestre, a curva de sobrevivência dos alunos que não reprovaram se torna mais baixa do que a do outro grupo de alunos. Essa mudança no comportamento pode estar relacionada a alguma correlação com outra variável ou a presença de interação. Resultado também observado para a variável de sexo, analisado na subseção 5.1.3.

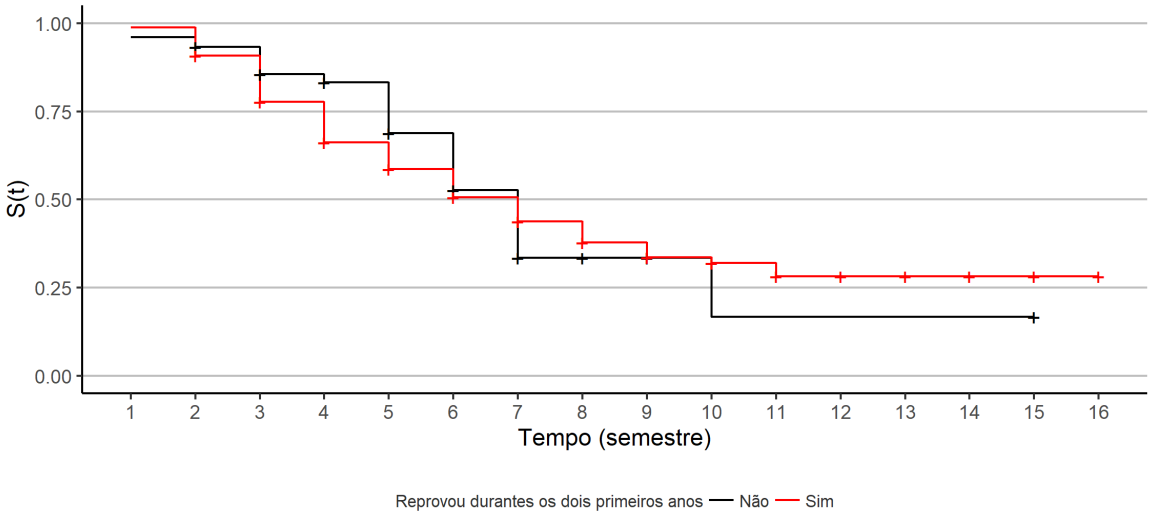

Figura 26: Gráfico de sobrevivência da estimativa de Kaplan-Meier por alunos que reprovaram durante os dois primeiros anos ou não

A análise da Figura 27 trás como análise que a diferença percentual de falha entre os alunos que não reprovaram e os alunos que reprovaram é de 22% para aqueles que reprovaram. Portanto,há evidências que essa variável tenha impacto na evasão do aluno. Em conjunto, têm-se que a análise apontada segundo a Figura 5 para a curva de sobrevivência nos primeiros anos também se torna corroborativa para essa evidência.

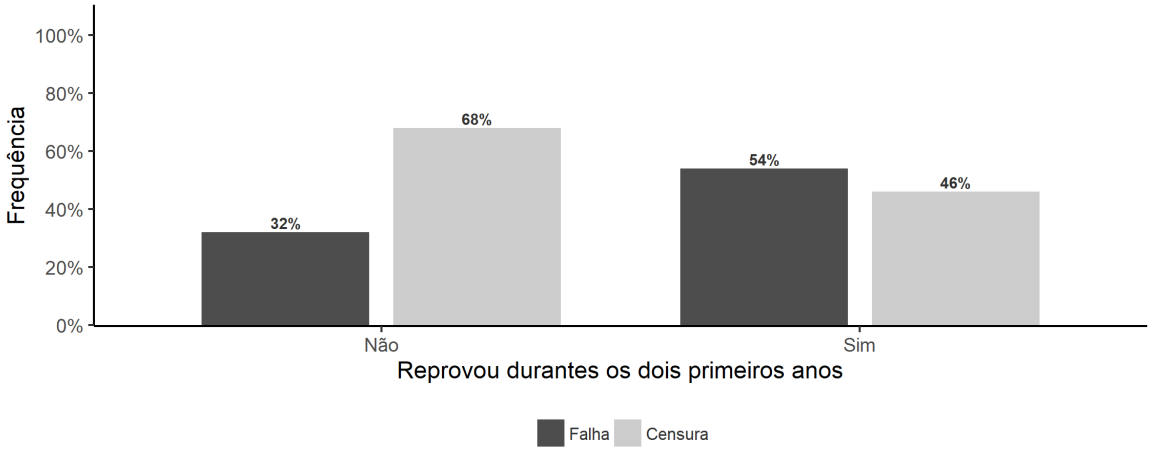

Figura 27: Gráfico de colunas da distribuição de alunos que reprovaram durante os dois primeiros anos ou não por falha e censura

5.1.17 Quantidade de disciplinas cursadas

Para a quantidade de disciplinas cursadas nota-se uma variância grande para na sua distribuição, tendo alta heterogeneidade, com seu coeficiente de variação em 79%. Nos dados, observou-se a presença de 3 alunos alunos que não cursaram nenhuma disciplina e evadiram o curso no 1º semestre, dois deles por não ter cumprido condição, portanto foram desligados, e o outro realizou novo vestibular. Este fato explica o mínimo de 0 disciplinas cursadas, apresentado na Tabela 19.

Tabela 19: Medidas resumo da quantidade de disciplinas cursadas

Estatística	Valor
Média	13,77
Variância	119,37
Desvio Padrão	10,93
Coeficiente de Variação	0,79
Mínimo	0,00
Máximo	62,00

Na Figura 28 observa-se que alunos que não evadiram o curso tem a média de disciplinas cursadas maior que os alunos que evadiram o curso. Entretanto, pode-se notar uma maior dispersão nos dados para aqueles que não evadiram. Ainda, apesar das diferenças apresentadas, os dois boxplots se cruzam, podendo fazer com que as evidências apresentadas não sejam suficientemente fortes para refletir no modelo final.

Ademais, reflete-se se a média maior para os alunos que não evadiram significam uma relação com o engajamento do aluno com o curso. Isto é, alunos mais engajados com o curso cursam mais disciplinas. Outra reflexão se dá quanto à sua correlação com as disciplinas reprovadas, já que uma vez que o aluno reprova, na maioria das vezes ele precisa fazer a mesma disciplina e, no fim, o total é maior mesmo para aqueles alunos que reprovaram e não evadiram.

Outras variáveis relacionadas ao engajamento do aluno com o curso ou universidade tais como: atividades extracurriculares, estágio, programa de iniciação científica (PIBIC) podem ser interessantes para um estudo futuro de evasão escolar.

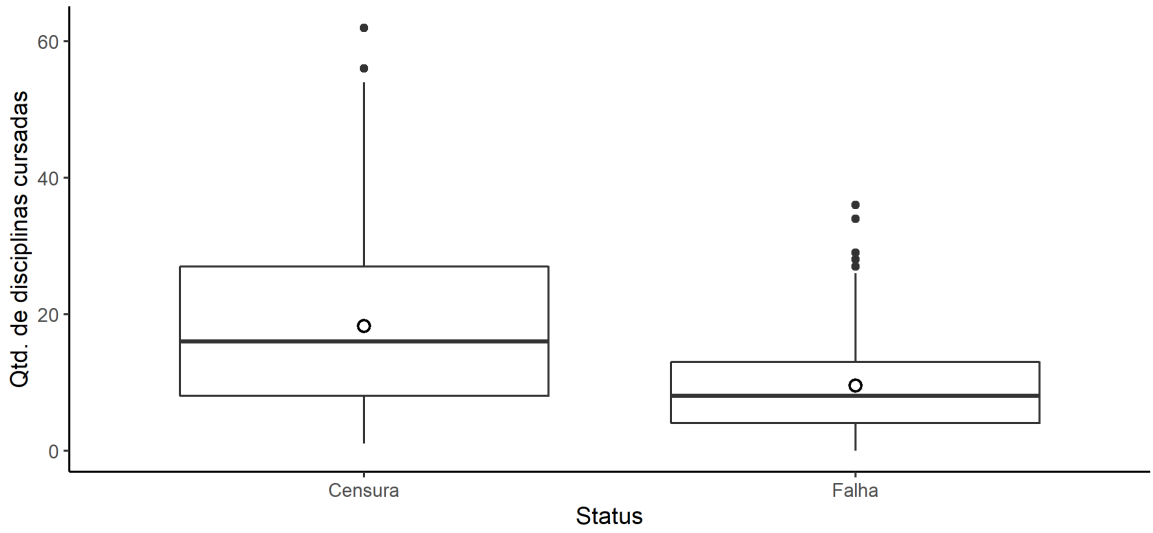

Figura 28: Boxplot da distribuição da quantidade de disciplinas cursadas por falha e censura

5.1.18 Soma de créditos cursados

Quando considerado a soma de créditos cursados ao invés da quantidade de disciplinas cursadas, pode-se observar que os dados se tornam heterogêneos, tendo um coeficiente de variação de 78%. Em média, têm-se que os alunos cursam 58,59 créditos.

Tabela 20: Medidas resumo da soma de créditos cursados

Estatística	Valor
Média	58,59
Variância	2099,69
Desvio Padrão	45,82
Coeficiente de Variação	0,78
Mínimo	0,00
Máximo	256,00

Corroborando com a análise feita na Seção 5.1.17 têm-se que os alunos que não evadiram o curso tem créditos cursados maior que os alunos que evadiram, podendo ser visto não apenas pelas médias mas também pelo 3º quartil dos alunos que evadiram, mostrando que 75% dos alunos que evadiram têm créditos cursados abaixo da mediana dos alunos não evadidos. Como também dito na Seção 5.1.2, para tempos maiores há uma maior presença de censuras, portanto, espera-se que estes alunos tenham uma quantidade de disciplinas cursadas maiores e, consequentemente, a soma de créditos também.

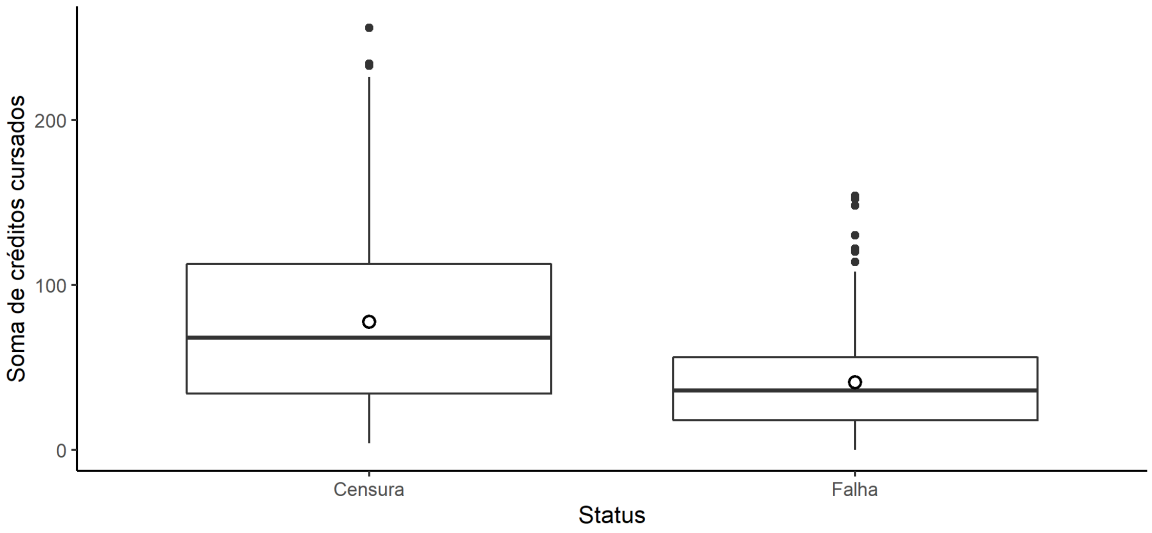

Figura 29: Boxplot da distribuição da some de créditos cursados por falha e censura

5.1.19 Proporção de créditos reprovados

Outro ponto de vista possível para a análise das disciplinas reprovadas pelos alunos é considerar a proporção que esses créditos reprovados representam em todo o seu tempo na academia.

Pode-se observar, portanto, que existe uma média de 43% dos créditos cursados em reprovações e que os dados são heterogêneos, como também pode ser visto na Figura 30 em que os dois boxplots são dispersos.

Ainda na Figura 30, pode-se notar que 50% dos alunos que evadiram possuem uma proporção de créditos reprovados maior que os alunos que não evadiram. Pode-se notar que a proporção média dos alunos evadidos é maior que o 3º quartil dos alunos não evadidos, o que trás uma evidência de que a proporção de créditos reprovados possa influenciar no tempo do aluno até a evasão.

Tabela 21: Medidas resumo da proporção de créditos reprovados

Estatística	Valor
Média	0,43
Variância	0,09
Desvio Padrão	0,30
Coeficiente de Variação	0,71
Mínimo	0,00
Máximo	1,00

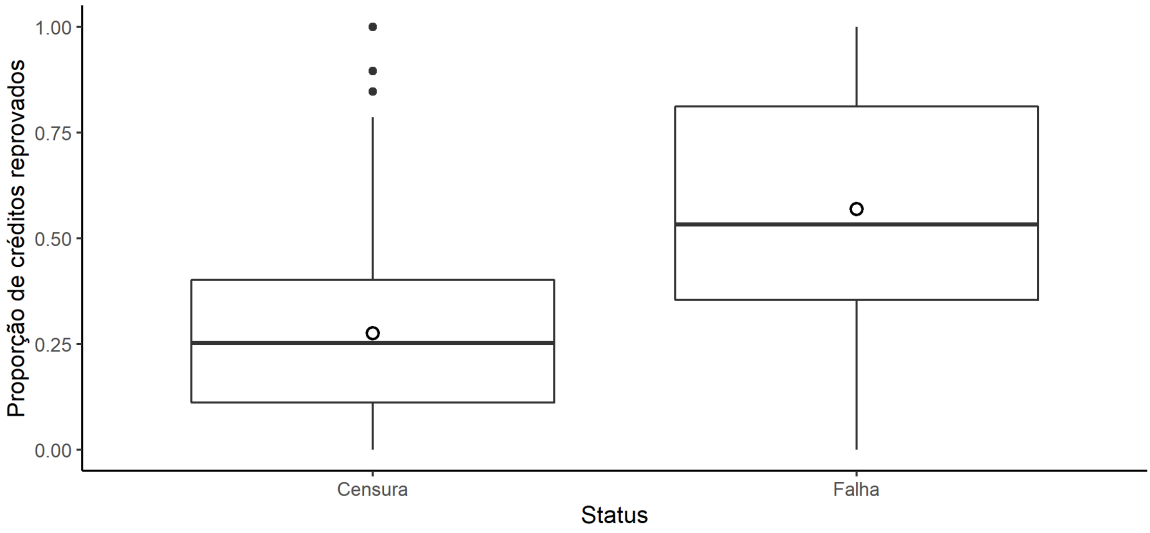

Figura 30: Boxplot da distribuição da proproção de créditos reprovados por falha e censura

5.1.20 Média de créditos cursados p/ semestre

Nessa variável, os 3 alunos que foram citados na Seção 5.1.17 também são responsáveis pelo valor da média mínima de créditos cursados por semestre dos alunos, já que estes não cursaram nenhuma disciplina.

Tabela 22: Medidas resumo da média de créditos cursados por semestre

Estatística	Valor
Média	13,97
Variância	13,26
Desvio Padrão	3,64
Coeficiente de Variação	0,26
Mínimo	0,00
Máximo	30,00

Para a média de créditos cursados por semestre pode-se notar que os dados são homogêneos, tendo o coeficiente de variação de 26%. A média das médias de créditos cursados é de, aproximadamente, 14 créditos por semestre.

Os dados continuam homogêneos quando separamos entre os grupos de alunos que não evadiram e os alunos que evadiram, tendo pouca dispersão nos boxplots apresentados na Figura 31. Também é possível notar que as médias dos dois boxplots estão próximas uma da outra e, além disso, estão aproximadamente em 14 créditos médios cursados por semestre.

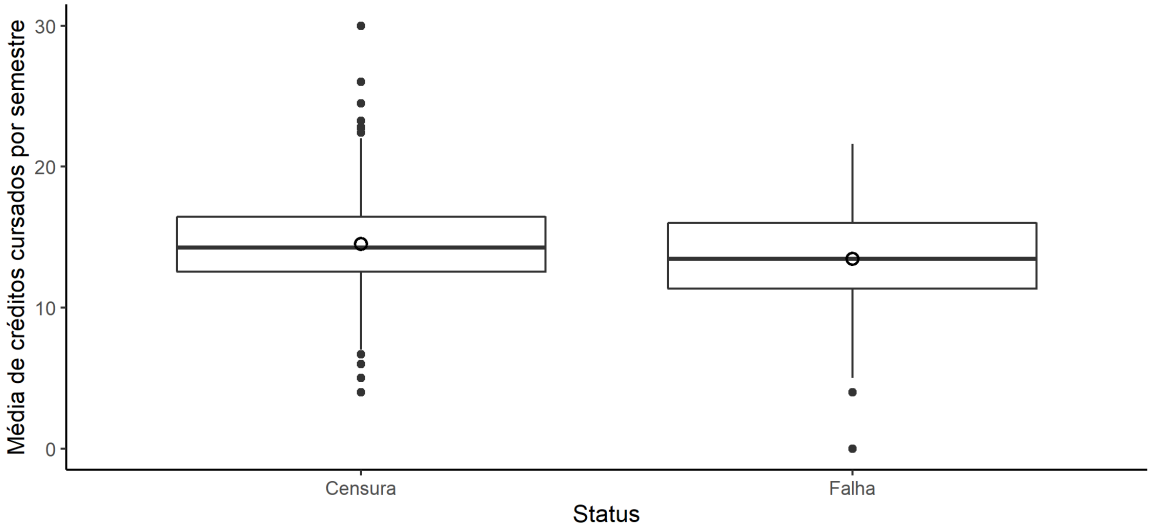

Figura 31: Boxplot da distribuição da média de créditos cursados por semestre por falha e censura

5.1.21 Quantidade de trancamentos

A possibilidade de trancar uma disciplina, seja justificado ou regular, é um recurso importante da carreira acadêmica do aluno, uma vez que este não é considerado como uma reprovação no currículo.

Tabela 23: Medidas resumo da quantidade de trancamentos

Estatística	Valor
Média	1,52
Variância	3,52
Desvio Padrão	1,87
Coeficiente de Variação	1,24
Mínimo	0,00
Máximo	15,00

Nota-se, pela Tabela 23, que, em média, os alunos trancam de 1 a 2 matérias em

todo o seu período de curso. Ainda, é possível observar que existe grande variação nos dados, tornado-os heterogêneos.

Pode-se observar na Figura 32 que os alunos que não evadiram possuem uma média de trancamentos levemente maior que os alunos que evadiram e também são mais dispersos. Entretanto, não aparenta existir evidências de que a quantidade de trancamentos feitos até o fim do período que o aluno esteve no curso está influenciando se o aluno evade ou não o curso.

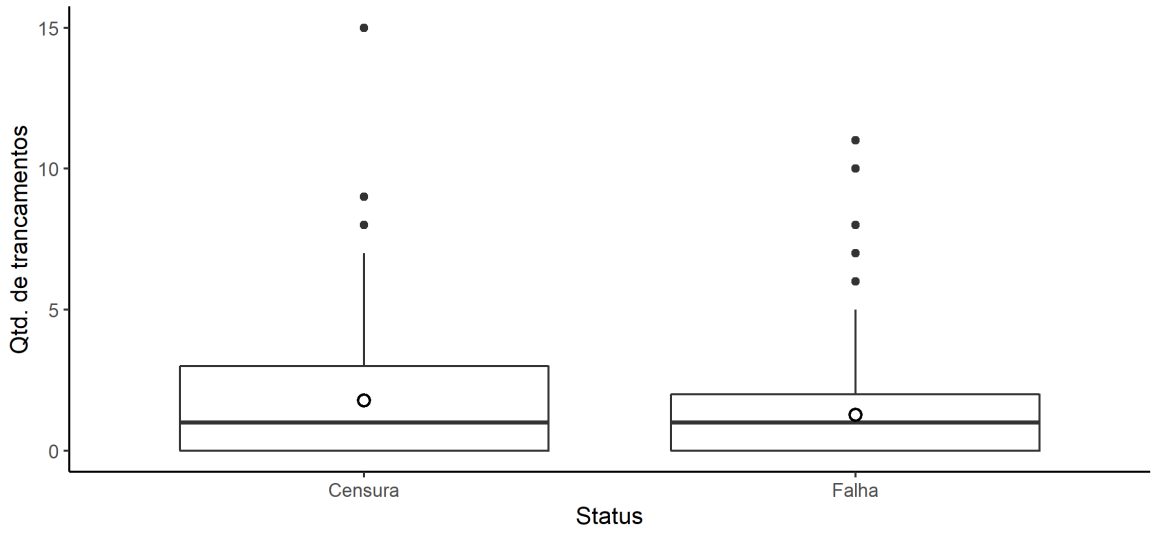

Figura 32: Boxplot da distribuição da quantidade de trancamentos por falha e censura

5.1.22 Soma de créditos trancados

Seguindo com a soma de créditos trancados, nos dados observou-se que aproximadamente 39% dos alunos não trancaram nenhuma matéria durante seu período no curso, o que pode explicar a média de 6,57 créditos trancados e assimetria para os valores mais altos. Além do mais, o coeficiente de variação mostra que os dados são heterogêneos, possivelmente causado pelos *outliers* que também são presentes quando separado entre os alunos que evadiram e que não evadiram, conforme a Figura 33

Tabela 24: Medidas resumo da soma de créditos trancados

Estatística	Valor
Média	6,57
Variância	65,99
Desvio Padrão	8,12
Coeficiente de Variação	1,24
Mínimo	0,00
Máximo	60,00

Pode-se observar na Figura 33 que a distribuição dos dois grupos de alunos são semelhantes, sendo um pouco mais dispersa para os alunos que não evadiram, que também tem uma média de créditos trancados um pouco maior que os alunos que evadiram.

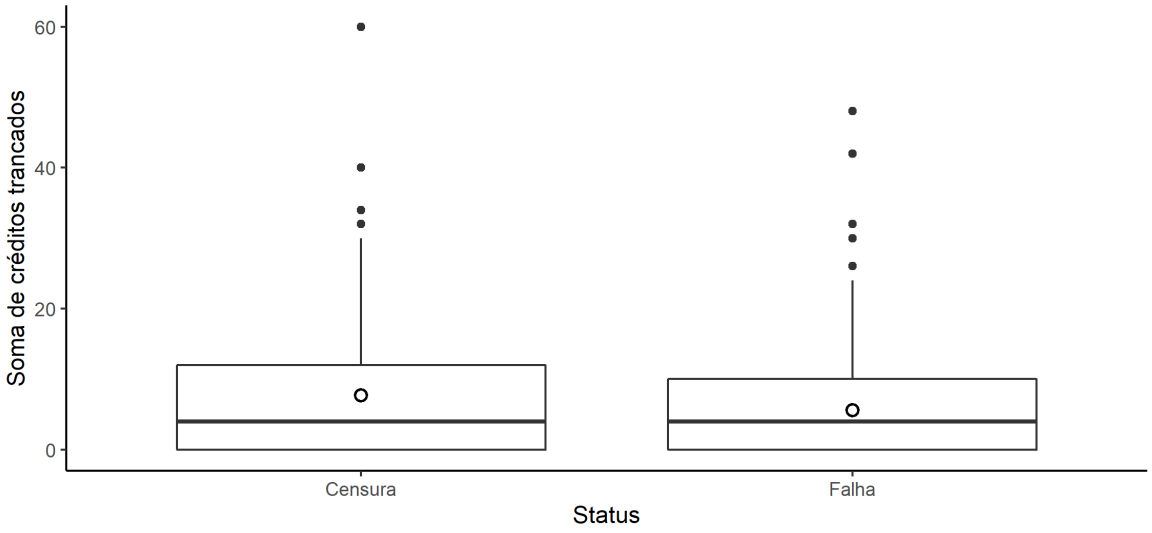

Figura 33: Boxpot da distribuição da soma de créditos trancados por falha e censura

5.1.23 Análise bivariada

Nesta seção será avaliada as associações presentes entre as variáveis qualitativas por meio do coeficiente de contingência modificado, e também entre a correlação entre as variáveis quantitativas utilizando o coeficiente de correlação de Pearson. Ambos os coeficientes podem variar de 0, nenhuma associação, a 1, forte associação.

A análise bivariada têm como objetivo verificar quais são as relações mais fortes entre as variáveis para evitar que, ao colocá-las juntas no modelo, possa ocorrer problemas como o de multicolinearidade, trazendo resultados viesados para a pesquisa.

Nesse estudo considerou relações a partir de 0,5 para considerar evitar colocar as duas variáveis relacionadas ao mesmo tempo no modelo.

Tabela 25: Coeficiente de contingência modificado para medir a associação entre variáveis qualitativas

	Sexo	Sistema de cota	Escola	Forma de ingresso na UnB
Sexo	1,00	0,05	0,04	0,14
Sistema de cota	0,05	1,00	0,61	0,43
Escola	0,04	0,61	1,00	0,13
Forma de ingresso na UnB	0,14	0,43	0,13	1,00
Cursou verão	0,06	0,07	0,01	0,29
Currículo	0,00	0,27	0,02	0,46
Distância da residência até a UnB (metros)	0,05	0,28	0,31	0,24
Reprovou durante os 2 primeiros anos	0,06	0,03	0,02	0,19

Na Tabela 25 nota-se que a associações forte está entre as variáveis de escola do aluno e a de sistema de cota, tendo um associação média forte de 0,61. Esse resultado é esperado, uma vez que os alunos que os alunos que ingressaram utilizando um sistema de cota são de escolas públicas. Entretanto, essa associação não se tornou mais forte pelo fato de que nem todos os alunos de escola pública utilizaram o sistema de cota para ingressar na UnB. Portanto, foi considerado a variável de sistema de cota para ser testada no modelo.

Outras associações presentes que ainda não foram suficientes para retirar uma das duas variáveis associadas estão entre a forma de ingresso na UnB e o sistema de cotas e também entre a primeira com o currículo vigente do aluno, ambas tendo associação média (0,43 e 0,46, respectivamente). É válido investigar com mais atenção essas associações em estudos futuros que sejam mais restritivos para entender se há conhecimento empírico sobre essas associações.

Na Tabela 26 entretanto, não há associações fortes entre as variáveis apresentadas.

Tabela 26: Coeficiente de contingência modificado para medir a associação entre variáveis qualitativas

	Cursou verão	Currículo	Distância da residência até a UnB (metros)	Reprovou durante os 2 primeiros anos
Sexo	0,06	0,00	0,04	0,06
Sistema de cota	0,07	0,27	0,28	0,03
Escola	0,01	0,02	0,34	0,02
Forma de ingresso na UnB	0,29	0,46	0,27	0,19
Cursou verão	1,00	0,15	0,12	0,02
Currículo	0,15	1,00	0,14	0,12
Distância da residência até a UnB (metros)	0,06	0,11	1,00	0,09
Reprovou durante os 2 primeiros anos	0,02	0,12	0,09	1,00

Já entre as variáveis quantitativas apresentadas nas Tabelas 27 a 29, nota-se presença de correlações em mais variáveis estudadas. As correlações de 0,5 ou maiores foram sinalizadas nas tabelas citadas.

As variáveis, portanto, que apresentaram correlações fortes são: IRA, qtd. de reprovações, qtd. de disciplinas cursadas, qtd. de trancamentos, soma de créditos reprovados, soma de créditos cursados, soma de créditos trancados, proporção de créditos reprovados.

Das correlações destacadas, as existentes entre as variáveis de quantidade de disciplinas cursadas, trancadas ou reprovadas com suas respectivas variáveis de soma de créditos, era esperado valores altos, uma vez que a diferença entre elas é a unidade de medida.

Um resultado notável apresentado na Tabela 27 é entre as correlações das variáveis relacionadas às disciplinas cursadas e as variáveis relacionadas às reprovações. Tanto para a quantidade para a soma de créditos, as correlações apresentam valores positivos, ou seja, quanto maior o valor de uma, maior o valor da outra. Esse resultado corrobora com a suspeita levantada pela análise da seção 5.1.17 na qual levanta a hipótese que o aluno que reprovou uma matéria terá que refazê-la e, portanto, terá um somatório de créditos e disciplinas maior mesmo que esse aluno não evada o curso.

Tabela 27: Coeficiente de correlação de Pearson entre variáveis quantitativas

	IRA	Qtd. de reprovações	Qtd. de disciplinas cursadas	Qtd. de trancamentos
IRA	1,00	-0,30	0,37	0,14
Qtd. de reprovações	-0,30	1,00	0,57*	0,38
Qtd. de disciplinas cursadas	0,37	0,57*	1,00	0,51
Qtd. de trancamentos	0,14	0,38	0,51*	1,00
Média de créditos cursados p/ semestre	0,10	0,26	0,37	-0,05
Dif. entre a entrada na UnB e no curso (semestres)	0,08	-0,07	-0,00	0,00
Idade	-0,16	-0,06	-0,10	-0,04
Distância da residência até a UnB (metros)	0,02	0,01	-0,00	-0,03
Soma de créditos reprovados	-0,29	0,99*	0,58*	0,37
Soma de créditos cursados	0,37	0,56*	1,00	0,50
Soma de créditos trancados	0,13	0,37	0,49	0,99*
Proporção de créditos reprovados	-0,94*	0,31	-0,39	-0,19

Tabela 28: Coeficiente de correlação de Pearson entre variáveis quantitativas

	Média de créditos cursados p/ semestre	Dif. entre a entrada na UnB e no curso (semestres)	Idade	Distância da residência até a UnB (metros)
IRA	0,10	0,08	-0,16	0,02
Qtd. de reprovações	0,26	-0,07	-0,06	0,01
Qtd. de disciplinas cursadas	0,37	-0,00	-0,10	-0,01
Qtd. de trancamentos	-0,05	0,00	-0,04	-0,03
Média de créditos cursados p/ semestre	1,00	-0,02	-0,21	0,04
Dif. entre a entrada na UnB e no curso (semestres)	-0,02	1,00	0,01	0,01
Idade	-0,21	0,01	1,00	-0,04
Distância da residência até a UnB (metros)	0,04	0,00	-0,04	1,00
Soma de créditos reprovados	0,27	-0,06	-0,07	0,00
Soma de créditos cursados	0,37	0,00	-0,10	-0,01
Soma de créditos trancados	-0,05	0,01	-0,05	-0,03
Proporção de créditos reprovados	-0,07	-0,07	0,14	-0,01

Outra associação notável presente na Tabela 29 é entre a variável de IRA e a proporção de créditos reprovados, estas duas na sua construção levam em consideração a relação de quais disciplinas foram reprovadas, como carga horária e por sendo obrigatória ou optativa no caso do IRA, por exemplo. Logo, a correlação negativa entre as duas é esperado, já que quanto maior o valor do IRA, espera-se que o aluno tenha menos disciplinas reprovadas, logo, uma proporção de reprovação menor e vice-versa.

Tabela 29: Coeficiente de correlação de Pearson entre variáveis quantitativas

	Soma de créditos reprovados	Soma de créditos cursados	Soma de créditos trancados	Proporção de créditos reprovados
IRA	-0,29	0,37	0,13	-0,94*
Qtd. de reprovações	0,99*	0,56*	0,37	0,31
Qtd. de disciplinas cursadas	0,58*	1,00	0,49	-0,39
Qtd. de trancamentos	0,37	0,50*	0,99*	-0,19
Média de créditos cursados p/ semestre	0,27	0,37	-0,05	-0,07
Dif. entre a entrada na UnB e no curso (semestres)	-0,06	0,00	0,01	-0,07
Idade	-0,07	-0,10	-0,05	0,14
Distância da residência até a UnB (metros)	0,01	-0,01	-0,03	-0,01
Soma de créditos reprovados	1,00	0,58*	0,36	0,31
Soma de créditos cursados	0,58*	1,00	0,48	-0,38
Soma de créditos trancados	0,36	0,48	1,00	-0,18
Proporção de créditos reprovados	0,31	-0,38	-0,18	1,00

5.2 Modelagem para o banco completo

Nesta subseção serão apresentados as etapas realizadas para a construção dos modelos de regressão utilizando o banco de dados completo. Para as análises, utilizou-se o valor do nível de significância de 10%.

5.2.1 Seleção da distribuição

O passo que guia as análise seguintes se dá pela escolha da distribuição que melhor se ajusta aos dados. A princípio, foi considerado que os dados possuíssem tempo discreto então o ponto de partida será a distribuição Log-Logística Discreta e foram consideradas como comparação as distribuições para tempos contínuos Log-Logística e Log-Normal. A análise foi feita utilizando de sobrevivência para ambas as distribuições e a comparação com Kaplan-Meier (KP) e o critério AIC para verificar qual distribuição retorna um modelo mais parcimonioso.

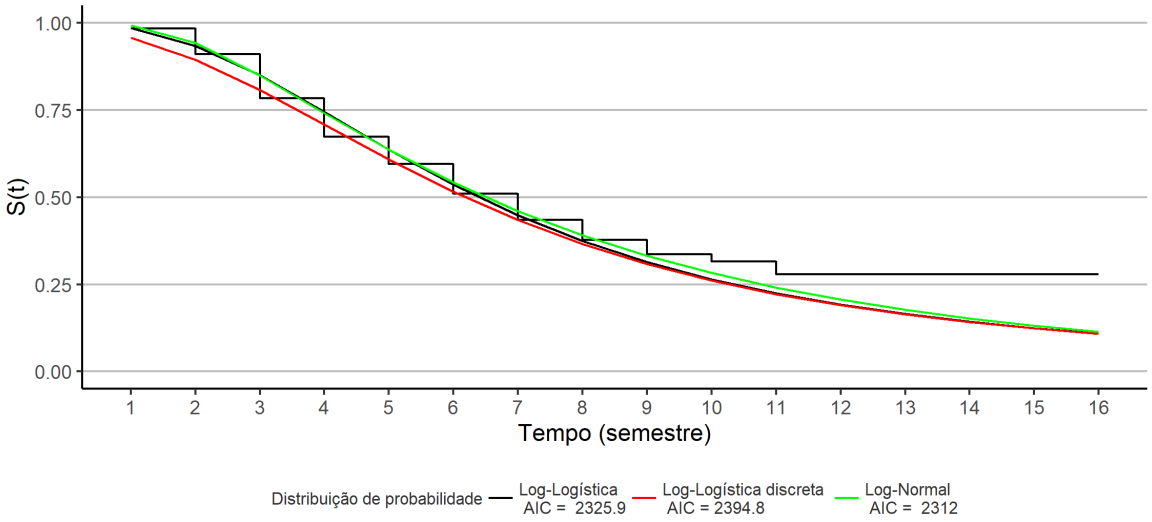

Figura 34: Comparação das curvas de sobrevivência entre as distribuições: Log-Logística, Log-Logística Discreta e Log-Normal

Pela Figura 34, nota-se que as três distribuições possuem comportamento, em geral, semelhante. Contudo, pode-se notar que a Log-Logística Discreta subestima a função de sobrevivência para os tempos menores. Nesse caso, as distribuições contínuas são as melhores candidatas para ajustar os dados.

Entre as distribuições contínuas, nota-se que nos tempos maiores a Log-Normal subestima menos que as demais distribuições. Ela também apresenta o menor AIC, ou seja, apresenta um modelo mais parcimonioso. Sendo assim, a Log-Normal foi selecionada para construir o ajuste dos dados.

5.2.2 Modelos univariados

Uma forma de analisar de forma exploratória se as variáveis explicativas influenciam e como influenciam na variável resposta é analisando-as em um modelo de regressão sem a presença das outras variáveis. Essa análise pode ajudar no diagnóstico do modelo

final já que pode-se estudar possíveis mudanças de interpretação e de significância no modelo.

Essa análise também é crucial para auxiliar no processo de seleção de variáveis a serem testadas no modelo completo já que essas podem definir a precisão e interpretação do modelo.

Para a interpretação do efeito das variáveis na curva de sobrevivência, têm-se que os valores negativos da estimativa contribuem negativamente para a variável resposta, ou seja, com o acréscimo de uma unidade no valor da variável o tempo até a evasão do aluno diminui. E os valores positivos contribuem positivamente para a variável resposta, sendo assim, o aumento de uma unidade no valor da variável aumenta o tempo até a evasão do aluno.

Tabela 30: Coeficientes estimados, erro padrão, estatística do teste e p-valor dos modelos contendo apenas uma variável explicativa quantitativa

Variável	Estimativa	Erro padrão	Estatística do teste	P-valor
IRA	0,3929	0,0207	18,9469	<0,0001
Qtd. de reprovações	0,0515	0,0087	5,9308	<0,0001
Qtd. de disciplinas cursadas	0,0615	0,0025	25,0825	<0,0001
Qtd. de trancamentos	0,1863	0,0159	11,6961	<0,0001
Média de créditos cursados p/ semestre	0,0312	0,0090	3,4776	0,0005
Dif. entre a entrada na UnB e no curso (semestres)	0,1274	0,0730	1,7450	0,081
Idade	-0,0157	0,0036	-4,3077	<0,0001
Distância da residência até a UnB (metros)	-0,0000	0,0000	-0,1315	0,8954
Soma de créditos reprovados	0,0124	0,0019	6,3854	<0,0001
Soma de créditos cursados	0,0146	0,0006	24,9863	<0,0001
Soma de créditos trancados	0,0416	0,0037	11,3439	<0,0001
Proporção de créditos reprovados	-1,5686	0,0821	-19,1026	<0,0001

Pela análise da Tabela 30 nota-se que a única variável que não foi significativa no modelo univariado é a de distância da residência até a UnB, apresentando uma estimativa de 0 e um p-valor de aproximadamente 0,9. Ainda que não significativa sozinha, a variável será testada na rodada de seleção de variáveis

A variável que possui a maior estimativa é a de proporção de créditos reprovados, sendo de -1,5686, com o p-valor <0,0001. A segunda maior estimativa, de 0,3929 com p-valor <0,0001, é dada pela variável de IRA. A interpretação das suas estimativas seguem contrárias uma da outra assim como acontece com sua correlação analisada na Tabela 29. Isto é, a proporção de créditos possui um efeito negativo na probabilidade de sobrevivência do aluno enquanto que a variável de IRA possui um efeito positivo.

Como as variáveis de IRA e proporção de créditos reprovados possuem uma forte correlação, foi tomada a decisão de considerar o modelo utilizando o IRA, uma vez que esta medida já é utilizada no sistema da UnB como mensuração do rendimento acadêmico.

Ainda na Tabela 30, nota-se que a quantidade de reprovações e a soma de créditos reprovados apresentaram efeitos positivos ambos com p-valores <0,0001, o que é contra intuitivo. Pois assim, para o aumento de uma unidade na quantidade de reprovações ou maior a soma de créditos reprovados, maior a probabilidade de sobrevivência. Esse feito pelo fato de que muitos alunos reprovaram pelo menos uma disciplina durante o curso. Logo, existe uma grande proporção de alunos censurados que têm disciplinas reprovadas no seu currículo.

Para as variáveis de quantidade de disciplinas cursadas e créditos cursados, a estimativa apresenta um efeito positivo na curva de sobrevivência, ambas com p-valores <0,0001. Esse efeito retorna o questionamento feito na Seção 5.1.18, na qual alunos mais engajados tendem a cursar mais disciplinas durante seus semestres. Outra explicação pode ser dada pela análise das correlações apresentadas na Tabela 29, na qual as correlações são positivas.

Como as variáveis de quantidade e de soma de créditos têm, em geral, correlações fortes. Decidiu-se por manter as variáveis de soma de créditos nas rodadas de seleção de variáveis.

A decisão foi tomada devido à hipótese de que a soma de créditos é mais informativa para o modelo, uma vez que diferentes disciplinas possuem diferentes cargas horárias. Portanto, os créditos carregam mais dessa informação do que a contagem da disciplina.

Seguindo para análise da Tabela 31, têm-se que duas variáveis são significativas quando sozinha no modelo: é a do aluno ter ou não cursado verão, com um p-valor <0,0001 e o currículo vigente com p-valor de 0,0002.

Tabela 31: Coeficientes estimados, erro padrão, estatística do teste e p-valor dos modelos contendo apenas uma variável explicativa qualitativa

Variável	Estimativa	Erro padrão	Estatística do teste	P valor
Sexo				
M	-0,0291	0,0993	-0,2933	0,7693
Sistema de cota				
Sim	-0,0496	0,0729	-0,6807	0,496
Escola				
Pública	-0,044	0,0628	-0,6996	0,4842
Forma de ingresso				
Programa de Avaliação Seriada	0,3541	0,1021	3,4691	0,0005
Sisu-Sistema de Seleção Unificada	0,0712	0,1051	0,6772	0,4983
Vestibular	0,2859	0,0898	3,1844	0,0015
Cursou verão				
Sim	0,7682	0,0812	9,4556	<0,0001
Currículo				
Velho	-0,2508	0,0665	-3,7712	0,0002
Distância da residência até a UnB (metros)				
De 7089,43 até 14044,27	-0,0095	0,0888	-0,1072	0,9146
De 14044,27 até 23457,23	0,0411	0,0885	0,4643	0,6424
Maior que 23457,23	0,0686	0,0894	0,7673	0,4429
Reprovou durante os 2 primeiros anos				
Sim	-0,0515	0,1195	-0,4309	0,6665

Para a variável de verão, o valor de referência é o aluno que não cursou verão, logo o fator Sim apresenta estimativa positiva de 0,7682. Isto é, se o estudante cursou uma disciplina de verão, ele terá maior probabilidade de sobrevivência.

Já a variável de currículo, o fator de referência é o currículo novo. Portanto, para o fator currículo velho, têm-se que a estimativa é de -0,2508. Isto é, se o estudante ingressou em um período que o currículo velho estava em vigência, a probabilidade desse alunos sobreviver é menor do que a probabilidade dos alunos que ingressaram com o currículo novo.

Como a variável de distância da residência do aluno até a UnB também foi construída com sua natureza quantitativa, decidiu-se por utilizar a variável quantitativa ao invés da qualitativa. O critério utilizado foi da hipótese de que a variável quantitativa pudesse trazer resultados mais fieis à realidade dos alunos, já que a diferença em metros não captadas pelo agrupamento feito na variável qualitativa podem apresentar dificuldades diferentes para os alunos na sua locomoção até a UnB.

Outra decisão que precisou ser tomado foi a de testar os modelos contendo escola e outro contendo o sistema de cota. Entretanto, na seleção de variáveis apenas o sistema de cota se mostrou significativo com a presença de outras variáveis, portanto, não foi considerado a escola dos alunos para a análise final.

5.2.3 Modelo final

Para a rodada de seleção de variáveis para a construção do modelo final, foi considerado a lista de variáveis abaixo:

Tabela 32: Variáveis consideradas utilizadas do modelo completo para as rodadas de seleção de variáveis

Variáveis
IRA
Sexo
Sistema de cota
Forma de ingresso
Cursou verão
Currículo
Média de créditos cursados por semestre
Dif. entre a entrada na UnB e no curso
Distância da residência até a UnB
Soma de créditos reprovados
Soma de créditos trancados
Idade
Reprovou nos 2 primeiros anos

A análise apontada na Figura 26 trouxe evidências de uma possível interação da variável que indica se o aluno reprovou durante os dois primeiros anos. Dado isso, foi feita uma análise dos boxplots das variáveis de média de crédito por semestre e soma de créditos reprovados pela variável que indica se o aluno reprovou durante os dois primeiros anos (Apêndices B e C).

A análise dos boxplots mostrou evidências de que o aluno ter reprovado ou não durante os dois primeiros anos influenciam nas duas variáveis citadas apesar de ter o R^2 baixo. Outra evidência de uma possível interação se dá pela análise do modelo sem interação (Apêndice A) mostrou que o coeficiente da média de créditos cursados por semestre há mudança de interpretação do modelo univariado quando colocada em conjunto da variável de reprovados durante os 2 primeiros anos.

Sendo assim, decidiu-se construir o modelo completo como referência para as rodadas de seleção de variáveis considerando as seguintes interações:

- Reprovou nos 2 primeiros anos e média de créditos por semestre;
- Reprovou nos 2 primeiros anos e soma de créditos reprovados

O algoritmo utilizado para a seleção de variáveis foi o *backwise*, que consiste em

retirar a variável com maior p-valor e maior que o nível de significância até que nenhuma variável precise ser retirada do modelo.

Após a rodada de seleção de variáveis, o modelo selecionado é apresentado na Tabela 33:

Tabela 33: Coeficientes estimados, erro padrão, estatística do teste e p-valor para o modelo final do banco de dados completo

Variável	Estimativa	Erro padrão	Estatística do teste	P-valor
β_0	-0,0675	0,1695	-0,3982	0,6905
IRA	0,3784	0,0180	21,0266	<0,0001
Sexo - M	0,1141	0,0586	1,9472	0,0515
Sistema de cotas - Sim	-0,0827	0,0443	-1,8684	0,0617
Verão - Sim	0,1327	0,0562	2,3624	0,0182
Currículo - Velho	-0,1436	0,0400	-3,5864	0,0003
Média de créditos p/ semestre	0,0269	0,0126	2,1350	0,0328
Soma de créditos reprovados	0,1193	0,0345	3,4604	0,0005
Soma de créditos trancados	0,0167	0,0029	5,8222	<0,0001
Reprovou nos 2 primeiros anos - Sim	0,8106	0,1769	4,5824	<0,0001
Média de créditos p/ semestre: Reprovou nos 2 primeiros anos - Sim	-0,0467	0,0137	-3,4180	0,0006
Soma de créditos reprovados: Reprovou nos 2 primeiros anos - Sim	-0,1025	0,0344	-2,9764	0,0029
log(scale)	-0,8720	0,0371	-23,5285	<0,0001

Partindo da análise da Tabela 33, verifica-se que o modelo final conta com a presença das duas interações analisadas nos boxplots presentes nos apêndices B e C. Sendo assim, a interação entre a média de créditos por semestre com a variável de reprovação durante os dois primeiros anos tem efeito negativo na probabilidade de sobrevivência de -0,0467 com p-valor de 0,0006. Ou seja, a medida que a média de créditos cursados por semestre dos alunos que reprovaram durante os dois primeiros anos cresce, a probabilidade de sobrevivência desses é menor do que aqueles que não reprovaram.

Ademais, a interação entre a soma de créditos reprovados e se o aluno reprovou nos dois primeiros anos também tem efeito negativo na curva de sobrevivência. A estimativa é de -0,1025 com p-valor de 0,0029, sendo o efeito maior que a interação anterior. Sendo assim, quanto maior a soma de créditos reprovados dentre os alunos que reprovaram nos dois primeiros anos a probabilidade de sobrevivência diminui em relação aos alunos que não reprovaram.

Os resultados dessas interações são importantes no modelo, já que sua interpretação é significativa. Essa importância também se dá pelo conhecimento obtido do estudo do Chagas (2019), que o resultado do seu modelo indica que quanto maior o número de reprovações nos dois primeiros anos, maior a probabilidade de evasão do aluno.

Ainda que o efeito individual da variável que indica se o aluno reprovou durante os dois primeiros anos seja positivo, a significância da sua interpretação é menor para a variável resposta, já que existe o efeito da interação, como citado no parágrafo anterior. Contudo, para estudos mais aprofundados, vale a pena verificar se ao considerar a variável de forma quantitativa, o resultado ainda difere ao analisar essa variável individualmente.

A segunda variável com maior efeito na variável resposta é o IRA, tendo um coeficiente de 0,3784 com p-valor<0,0001. Sendo assim, maiores valores de IRA aumentam a probabilidade de sobrevivência do aluno.

É visto também que os alunos do sexo masculino tem probabilidade de sobrevivência maior que os alunos do sexo feminino. A estimativa do efeito é de 0,1141 com p-valor de 0,0515. Vale a pena mencionar que se o estudo fosse mais restritivo, utilizando nível de significância de 5%, por exemplo, essa variável poderia não entrar no modelo. Sendo assim, é válido investigar mais afundo em estudos futuros considerando ter uma amostra maior de mulheres, já que são minoria nos dados (análise da Tabela 5).

O sistema de cotas apresentou ter um efeito negativo na curva de sobrevivência dos alunos, ou seja, alunos que ingressaram utilizando sistema de cotas tem probabilidade de sobreviver menor que os alunos que não ingressaram com sistema de cotas. A Tabela 33 mostra que a estimativa é de -0,0827 com p-valor de 0,0617. Assim como a variável de sexo, a interpretação dessa variável poderia não ser válido para um modelo mais restritivo.

Como descrito na Seção 5.1.10, os alunos que cursaram alguma disciplina de verão tem maior probabilidade de sobrevivência do que alunos que não cursaram. Esse resultado pode ser verificado na Tabela 33, observando seu coeficiente positivo de 0,1327 com p-valor de 0,0182.

Partindo para a variável do currículo vigente, que foi citada na Seção 5.1.9 como sendo relacionada à infraestrutura acadêmica. Têm-se que os alunos que ingressaram no currículo velho possui probabilidade de sobreviver menor que os alunos que ingressaram no currículo novo. A estimativa do seu efeito na curva de sobrevivência é de -0,1436, com p-valor de 0,0003.

A média de créditos cursados por semestre possui efeito positivo de 0,0269, com p-valor de 0,328. Sendo assim, à medida que a média de créditos cursados por semestre cresce, a probabilidade de sobrevivência aumenta. Entretanto, é preciso tomar cuidado ao analisar essa variável já que a sua interação possui efeito contrário.

Assim como a soma de créditos reprovados, que também possui efeito positivo na curva de sobrevivência. Ou seja, sua estimativa de 0,1193 (p-valor de 0,0005) indica que quanto maior o total de créditos reprovados, maior a probabilidade de sobrevivência do aluno. Esse resultado é contra intuitivo, uma vez que existe a possibilidade do aluno evadir por ter reprovado três vezes na mesma disciplina obrigatória. Apesar disso, a

frequência que alunos evadiram por esta forma é pequena.

A soma de créditos trancados apresentou um efeito positivo para a probabilidade de sobrevivência do aluno. Isto é, alunos com maiores quantidades de créditos trancados possuem maior probabilidade de sobrevivência. Nos dados, é pouca a presença de alunos que trancaram suas disciplinas e, ainda, o trancamento possui a dinâmica de que o aluno, na maioria das vezes, escolhe o trancamento, ao contrário da reprovação, que o aluno, geralmente, não escolhe reprovar uma matéria.

Ainda que existam coeficientes contra intuitivos que despertem a curiosidade de estudos futuros, a análise da adequação global do modelo feita pela Figura 35 mostra que o modelo se ajustou bem aos dados. O seu AIC também foi o menor dentre outros modelos testados, como o sem interação e os modelos com apenas com uma das interações. Sendo assim, há evidências de que os resultados obtidos no modelo descrito na Tabela 33 são válidos.

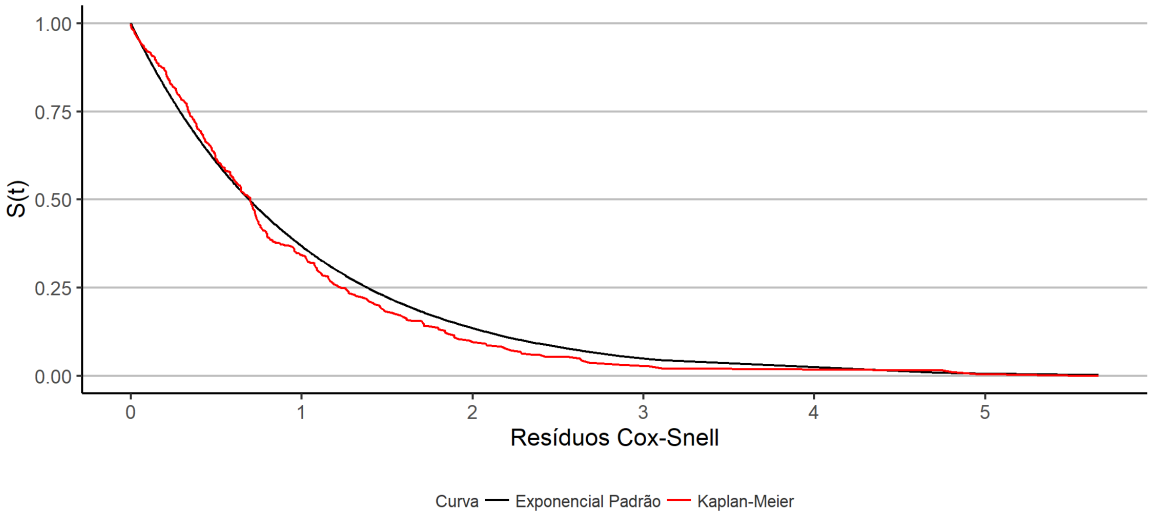

Figura 35: Resíduos de cox-snell do modelo final para o banco de dados completo

5.3 Modelagem para o banco separado por sexo

Nessa etapa, o banco de dados foi separado com o objetivo de investigar se as variáveis explicativas se mantém as mesmas do modelo para o banco de dados completo quando dividimos a população de origem do modelo. As etapas apresentadas na Seção 32 foram repetidas para cada uma dos bancos de dados.

5.3.1 Modelo para o sexo masculino

Pela Figura 36 nota-se que a distribuição Log-normal se mostrou igualmente melhor do que as distribuições Log-logística e Log-logística discreta.

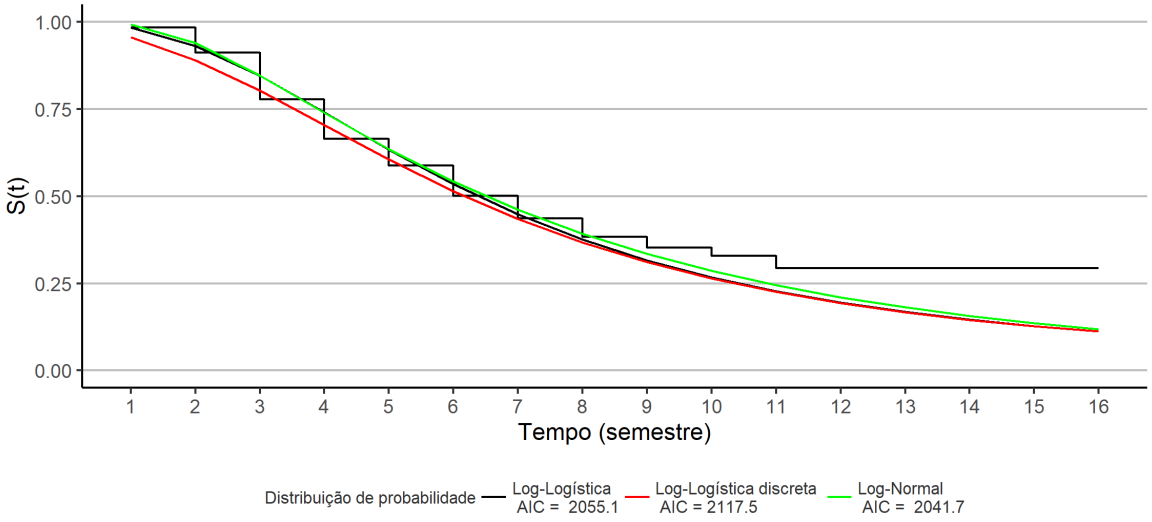

Figura 36: Comparação das curvas de sobrevivência entre as distribuições: Log-logística, Log-logística discreta e Log-normal para a população de alunos do sexo masculino

A análise exploratória do modelo univariado para das covariáveis considerando a população de alunos do sexo masculino (Apêndice D) não trouxe interpretações diferentes no efeito individual das variáveis.

Utilizando os mesmos critérios para a modelagem feita para o banco de dados completos do modelo completo de referência para a rodada de seleção de variáveis. Têm-se que o modelo final é descrito na tabela abaixo:

Tabela 34: Coeficientes estimados, erro padrão, estatística do teste e p-valor para o modelo final do banco de dados de alunos do sexo masculino

Variável	Estimativa	Erro padrão	Estatística do teste	P-valor
β_0	0,3779	0,1031	3,6652	0,0002
IRA	0,3911	0,0182	21,4483	<0,0001
Forma de ingresso na UnB - PAS	-0,2000	0,0717	-2,7881	0,0053
Forma de ingresso na UnB - SISU	-0,1712	0,0681	-2,5130	0,0120
Forma de ingresso na UnB - Vestibular	-0,1189	0,0599	-1,9838	0,0473
Cursou verão - Sim	0,1493	0,0616	2,4253	0,0153
Currículo - Velho	-0,1384	0,0452	-3,0657	0,0022
Soma de créditos reprovados	0,1208	0,0347	3,4772	0,0005
Soma de créditos trancados	0,0160	0,0030	5,4035	<0,0001
Reprovou nos 2 primeiros anos	0,2867	0,0909	3,1526	0,0016
Soma de créditos reprovados: Reprovou nos 2 primeiros anos - Sim	-0,1042	0,0347	-3,0043	0,0027
log(scale)	-0,8714	0,0395	-22,0796	<0,0001

Analisando a Tabela 34 e comparando com a Tabela 33, nota-se que as variáveis de sistema de cotas, média de créditos cursados por semestre e a interação entre a média de créditos por semestre e se o aluno reprovou nos dois primeiros anos não estão presentes no modelo final para a população de alunos do sexo masculino.

Ainda, diferentemente do modelo apresentado para o banco de dados completo, na Tabela 34 nota-se a presença da variável de forma de ingresso na UnB. Considerando que o fator de comparação é a categoria de outras formas de ingresso, nota-se que os alunos que ingressaram pelo PAS, SISU e vestibular têm probabilidade de sobrevivência menor que os alunos ingressantes por outras formas.

É interessante notar que a construção da variável de forma de ingresso (detalhado na Seção 4.3.18) também considerou os alunos que já tinham um diploma ou foram transferidos de um curso anterior. Portanto, há a suposição de que esses alunos sejam mais assertivos com o curso de Licenciatura em Computação, uma vez que esses alunos possuem experiência do ambiente universitário nos seus cursos anteriores. Outro detalhe é que esse efeito é contrário do que foi evidenciado na Seção 5.1.7.

Analisando a interação presente na tabela 34, nota-se que os alunos para valores maiores da soma de créditos reprovados e se o aluno reprovou durante os 2 primeiros anos a probabilidade de sobrevivência é menor do que quando o aluno não reprovou durante os dois primeiros anos com uma estimativa de -0,1042 e p-valor de 0,0027.

De acordo com a Tabela 34, as variáveis de IRA, soma de créditos reprovados, soma de créditos trancados, os alunos que cursaram uma disciplina no verão e a variável que indica se o aluno reprovou durante os dois primeiros anos - quando analisada individualmente - possuem efeitos positivos na curva de sobrevivência, assim como encontrados

na análise da Tabela 33.

Ainda com a análise da Tabela 34, os alunos que entraram com o currículo velho em vigência possuem menor probabilidade de sobrevivência, assim como apresentado no modelo considerando o banco de dados completo.

Para validar os resultados obtidos na Tabela 34, a análise dos resíduos de Cox-Snell será utilizada.

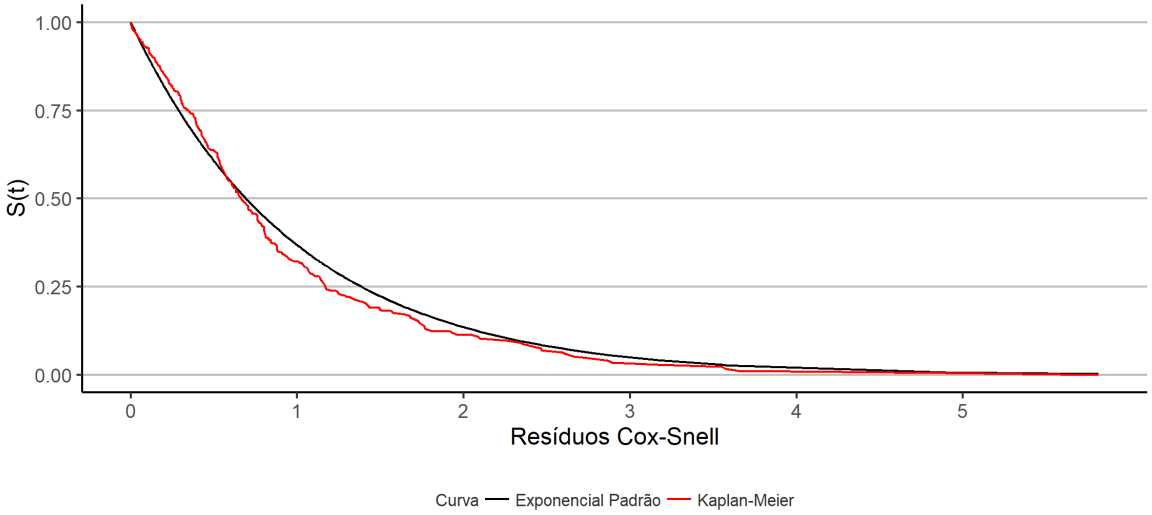

Figura 37: Resíduos de cox-snell do modelo final para o banco de dados de alunos do sexo masculino

Estudando a Figura 37, nota-se que há evidências para considerar o modelo adequado à população de alunos do sexo masculino. Isto é, os resultados obtidos possuem um bom ajuste global.

5.3.2 Modelo para o sexo feminino

Analisando a Figura 38 nota-se que a distribuição Log-logística e a Log-normal possuem comportamento e AIC semelhantes, ambas performam melhor que a Log-logística discreta. Apesar da semelhança, decidiu-se manter a distribuição Log-normal para manter a mesma distribuição das análises anteriores.

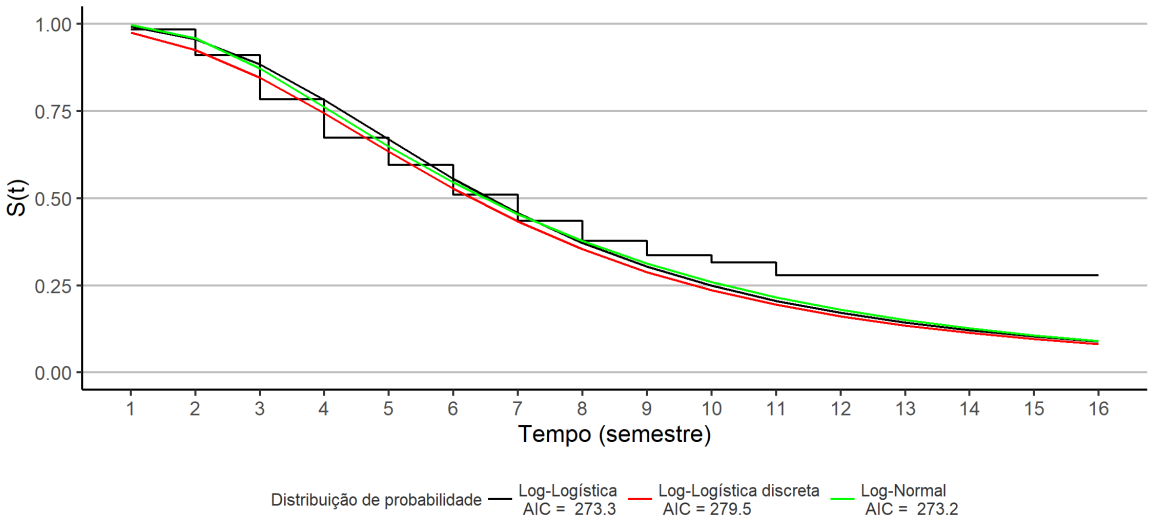

Figura 38: Comparação das curvas de sobrevivência entre as distribuições: Log-logística, Log-logística discreta e Log-normal para a população de alunos do sexo feminino

A análise feita da tabela do modelo individual para cada variável explicativa (Apêndice E) não apresentou, em geral, interpretações diferentes nos efeitos individuais em relação aos modelos utilizando o banco de dados completo.

Utilizando os critérios das análises anteriores para o modelo completo de referência na rodada de seleção de variáveis. A tabela 35 descreve o modelo final para os alunos do sexo feminino:

Tabela 35: Coeficientes estimados, erro padrão, estatística do teste e p-valor para o modelo final do banco de dados de alunos do sexo feminino

Variáveis	Estimativa	Erro padrão	Estatística do teste	P-valor
β_0	-2,3642	0,6337	-3,7310	0,0002
IRA	0,2703	0,0465	5,8119	<0,0001
Currículo - Velho	-0,3190	0,1026	-3,1081	0,0019
Média de créditos p/ semestre	0,3802	0,0813	4,6791	<0,0001
Soma de créditos reprovados	0,0118	0,0036	3,2567	0,0011
Soma de créditos trancados	0,0313	0,0068	4,6151	<0,0001
Reprovou nos 2 primeiros anos - Sim	3,6259	0,6518	5,5629	<0,0001
Média de créditos p/ semestre: Reprovou nos 2 primeiros anos - Sim	-0,4116	0,0825	-4,9864	<0,0001
Log(scale)	-1,0918	0,1065	-10,2538	<0,0001

Comparando com os resultados obtidos da Tabela 33, nota-se que o modelo apresentado na Tabela 35 não conta com a presença das variáveis de sistema de cotas, cursou verão e a interação entre a soma de créditos reprovados e se o aluno reprovou alguma disciplina nos 2 primeiros anos. Comparando também com a Tabela 34, a forma de ingresso na UnB não se tornou presente no modelo para os alunos do sexo feminino.

Entre os resultados observados, os alunos que reprovaram nos dois primeiros anos e possuem valores altos de média de créditos cursados por semestre têm probabilidade menor de sobreviver do que os alunos que possuem valores altos de média de créditos por semestre mas que não reprovaram. Sua estimativa é de -0,4116 com p-valor <0,001.

Ainda que a variável que indica se o aluno reprovou durante os dois primeiros anos, quando analisada mantendo as outras variáveis constantes, esteja com sua estimativa positiva. Segue ainda que a interpretação da sua interação é mais significativa no modelo. Pois como citado anteriormente, uma vez verificado a presença de interação, a análise individual se torna menos importante.

Das outras variáveis comuns aos outros modelos, nota-se que IRA, média de créditos por semestre, soma de créditos reprovados e soma de créditos trancados continuam com efeito positivo na probabilidade de sobrevivência dos alunos. Já os alunos do sexo feminino que ingressaram com o currículo velho também possuem probabilidade menor de sobrevivência.

Para esse modelo, nota-se que não houve o acréscimo de variáveis diferentes do modelo para o banco de dados completo. Um estudo futuro com um número maior da amostra de alunos do sexo feminino se torna importante pois os alunos do sexo feminino representam 12% do total de alunos, como analisado na tabela 5.

Ainda que a proporção seja menor, analisando a Figura 39 para validar o resultado e adequabilidade do modelo. Têm-se que os resíduos de Cox-Snell indicam que o modelo segue bem ajustado. É possível ver algumas fugas na sobrevivência estimada que podem estar relacionadas ao numero de alunos do sexo feminino.

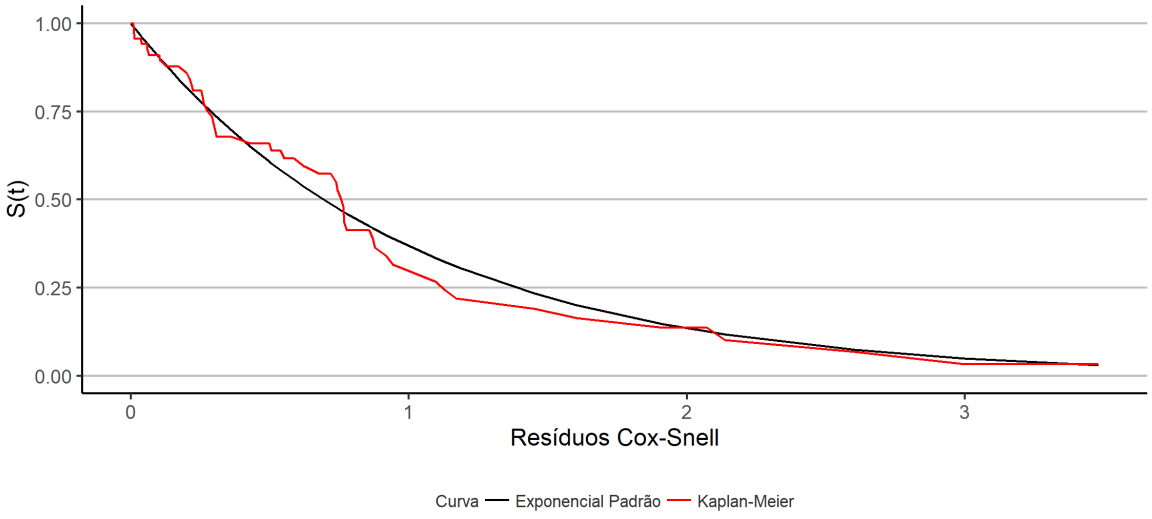

Figura 39: Resíduos de cox-snell do modelo final para o banco de dados de alunos do sexo feminino

6 Conclusões e considerações finais

Sabe-se que a evasão escolar está presente em todos os níveis de educação. No Brasil, a igualdade do acesso e permanência do ensino está descrito nas diretrizes da educação nacional. Portanto, a motivação desse trabalho surgiu de um problema real, a evasão escolar no ensino superior, tendo o Resumo Técnico do Censo da Educação Superior de 2019 como indicativo que a problemática continua na realidade brasileira.

O objetivo desse trabalho, portanto, é contribuir no entendimento dos fatores relacionados à evasão escolar formulando um modelo de regressão com técnicas de Análise de Sobrevivência, mais especificamente dos alunos do curso de Computação.

Os dados trabalhados foram medidos em semestres e os alunos podem ser divididos em três grupos: 1. Alunos que evadiram, 2. Alunos que estão cursando e 3. Alunos que se formaram. A partir desses dados, a variável resposta foi definida como o tempo, desde o ingresso do aluno no curso de Computação, até a evasão do referido curso.

Ainda no escopo do estudo, considerou-se os alunos que ingressaram entre os períodos de 2012/2 e 2019/2. Ademais, foi definido como falha os alunos que evadiram o curso, pertencentes ao grupo 1. E censura os alunos do grupo 2 e 3.

Para a análise dos dados foi utilizada técnicas descritivas, de sobrevivência e o modelo de regressão Log-normal considerando uma aproximação a tempos contínuos. Como apresentado na subseção 4.6.1, a modelagem foi construída para três bancos de dados, um contendo os dados completos - tanto para homens quanto para mulheres, um contendo apenas alunos do sexo masculino e outro contendo apenas alunos do sexo feminino.

Analisando os parâmetros obtidos nos três modelos, pode-se observar que a variável de IRA possui grande efeito na curva de sobrevivência dos alunos. Isto é, para valores maiores de IRA, os alunos têm maior probabilidade de sobreviver à evasão. Além disso, também foi visto para os três bancos que o aluno reprovar durante os dois primeiros anos possui grande importância na probabilidade de sobrevivência dos alunos, tanto suas interações quanto seu resultado *ceteris paribus*, ou seja, mantendo todo o resto constante.

O resultado da variável que indica se o aluno reprovou durante os dois primeiros anos possui um efeito *ceteris paribus* positivo. Isto é, alunos que reprovaram durante os dois primeiros anos possui maior probabilidade de sobreviver a evasão do que os alunos que não reprovaram. Esse resultado, entretanto, é contra intuitivo e também vai contrário ao resultado obtido no estudo de Chagas (2019). Apesar disso, as interações entre as variáveis de média de créditos cursados por semestre e soma de créditos reprovados com o aluno ter reprovado durante os dois primeiros anos mostram um efeito negativo estando mais

de acordo com estudos anteriores. Sendo assim, valores maiores da média de créditos por semestre ou da soma de créditos reprovados sendo que o aluno reprovou durante os dois primeiros anos representa uma probabilidade menor de evasão do que esses valores para alunos que não reprovaram. Os efeitos citados se repetem para os três bancos de dados.

Outro resultado observado é que os alunos que cursaram uma disciplina de verão tem, de modo geral, maior probabilidade de sobrevivência à evasão. Somado a isso, observou-se que os alunos que ingressaram com o currículo novo vigente, ou seja, a partir de 2015/2, tem maior probabilidade de sobrevivência.

Comparando os três modelos, nota-se que o conjunto de variáveis explicativas muda para a população de alunos do sexo masculino e também para o sexo feminino. No modelo com o sexo masculino, nota-se a presença da variável de ingresso na UnB, seja por portador de diploma, por transferência obrigatória ou facultativa e pelo Enem UnB têm maior probabilidade de sobrevivência do que aqueles que ingressaram pelo PAS, Sisu ou vestibular. Já o modelo do sexo feminino, percebe-se que não há presença de novas variáveis.

De maneira geral, ambos os modelos de regressão Log-normal tiveram bons ajustes aos dados e, ainda que alguns resultados sugerem estudos aprofundados, os resultados foram coerentes. Como proposta para trabalhos futuros, sugere-se:

- Ampliar a coleta de dados para pesquisar a possibilidade de outras covariáveis significativas para o tempo de sobrevivência dos alunos do curso de Computação;
- Incorporar informações que estejam relacionadas à infraestrutura acadêmica, como avaliação dos alunos sobre sua experiência e infraestrutura do curso.
- Incorporar informações sobre atividades extracurriculares, como estágio, PIBIC, etc.;
- Estudar a variável que indica se o aluno reprovou durante os dois primeiros anos com sua natureza quantitativa;
- Melhorar a precisão do cálculo de distância da residência do aluno até a UnB;
- Realizar estudos semelhantes para outros cursos de graduação a fim de comparação;
- Propor outros modelos de sobrevivência para os dados.

Referências

AARSET, M. V. How to identify a bathtub hazard rate. *IEEE Transactions on Reliability*, IEEE, v. 36, n. 1, p. 106–108, 1987.

CABELLO, A. et al. Formas de ingresso em perspectiva comparada: por que o sisu aumenta a evasão? o caso da unb. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, SciELO Brasil, v. 26, p. 446–460, 2021.

CASELLA, G.; BERGER, R. L. *Statistical inference*. [S.l.]: Duxbury, 2002.

CHAGAS, T. M. Análise da evasão dos alunos dos cursos da unb: um estudo no âmbito da graduação. 2019.

COLOSIMO, E. A.; GIOLO, S. R. *Análise de sobrevivência aplicada*. [S.l.]: Editora Blucher, 2006.

CORREIOS. *Correios - Busca CEP*. 2022. <https://buscacepinter.correios.com.br/app/faixa_cep_uf_localidade/index.php>. Acesso em 04 abr. 2022.

COX, D. R.; SNELL, E. J. A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 30, n. 2, p. 248–265, 1968.

FEDERAL, S. Lei de diretrizes e bases da educação nacional. *Diário Oficial [da] República Federativa do Brasil, Poder Legislativo, Brasília, DF*, v. 19, p. 26, 2005.

FILHO, R. B. S.; ARAÚJO, R. M. de L. Evasão e abandono escolar na educação básica no brasil: fatores, causas e possíveis consequências. *Educação por escrito*, v. 8, n. 1, p. 35–48, 2017.

FRITSCH, R.; ROCHA, C. S. da; VITELLI, R. F. A evasão nos cursos de graduação em uma instituição de ensino superior privada. *Revista Educação em Questão*, v. 52, n. 38, p. 81–108, 2015.

HARRISON, J.; HARRISON, M. J. *Package 'RSelenium'*. 2020.

INC., G. *Python Client for Google Maps Services*. 2022. <https://googlemaps.github.io/google-maps-services-python/docs/index.html>. Acesso em 04 abr. 2022.

LAWLESS, J. F. *Statistical models and methods for lifetime data*. [S.l.]: John Wiley & Sons, 2011. v. 362.

LOBO, M. Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. *Abmes Cadernos*, ABMES Brasília, v. 25, p. 9–58, 2012.

MAPACEP. *MapaCEP - Sistema de Busca CEP*. 2022. <https://www.mapacep.com.br>. Acesso em 04 abr. 2022.

RIBEIRO, I. M.; CORREIA, W. F. M.; CAMPOS, F. Setores acadêmicos que interferem na satisfação do aluno no ensino superior. *Acta Scientiarum. Education*, v. 43, p. e50121–e50121, 2021.

ROSS, S. M. *A first course in probability*. [S.l.], 1976.

TECHTUDO. *O que é API e para que serve? Cinco perguntas e respostas*. 2020. <https://www.techtudo.com.br/listas/2020/06/o-que-e-api-e-para-que-serve-cinco-perguntas-e-respostas.ghtml>. Acesso em 04 mai. 2022.

TEIXEIRA, A.; ESTATÍSTICAS, D. de. Resumo técnico do censo da educação superior 2019. *Brasília: Instituto*, 2019.

UPADHYAY, A. Haversine formula-calculate geographic distance on earth. *URL https://www. igismap. com/haversine-formula-calculate-geographicdistance-earth*, 2016.

WIKIMEDIA. *Haversine Formula*. 2004. <https://en.wikipedia.org/wiki/Haversine_formula>. Acesso em: 03 mar. 2022.

Apêndice

A Tabela da modelagem do banco de dados completo sem a presença de interação

Tabela 36: Coeficientes estimados, erro padrão, estatística do teste e p-valor para os modelos sem interação do banco de dados completo

Variáveis	Estimativa	Erro Padrão	Estatística do teste	P-valor
β_0	0.3975	0.1211	3.2813	0.0010
IRA	0.3805	0.0183	20.7574	0.0000
Sexo - M	0.1256	0.0598	2.1004	0.0357
Sistema de Cota - Sim	-0.0766	0.0453	-1.6905	0.0909
Cursou verão - Sim	0.1319	0.0575	2.2918	0.0219
Currículo - Velho	-0.1467	0.0410	-3.5829	0.0003
Média de créditos cursados p/ semestre	-0.0115	0.0059	-1.9669	0.0492
Soma de créditos reprovados	0.0164	0.0017	9.9088	0.0000
Soma de créditos trancados	0.0171	0.0029	5.8593	0.0000
Reprovou nos 2 primeiros anos - Sim	0.2253	0.0834	2.7027	0.0069
log *Scale*	-0.8486	0.0370	-22.9100	0.0000

B Análise bivariada da média de créditos por semestre por grupo de alunos que reprovaram nos dois primeiros anos ou não

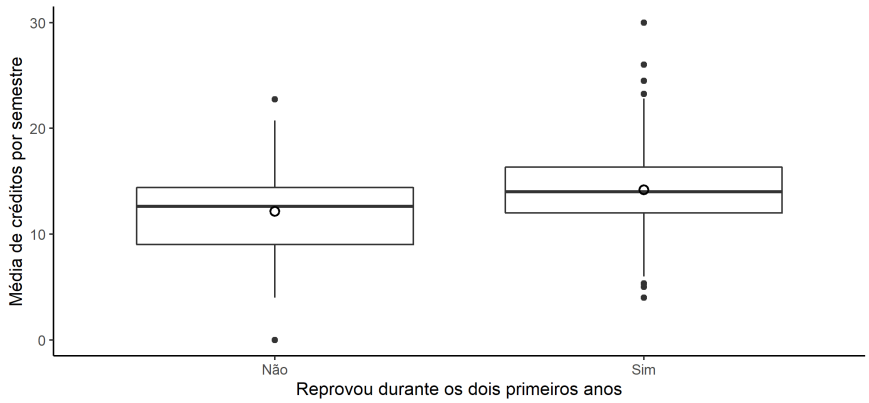

Figura 40: Boxplot da distribuição da média de créditos por semestre por grupo de alunos que reprovaram nos dois primeiros anos ou não

C Análise bivariada da soma de créditos reprovados por grupo de alunos que reprovaram nos dois primeiros anos ou não

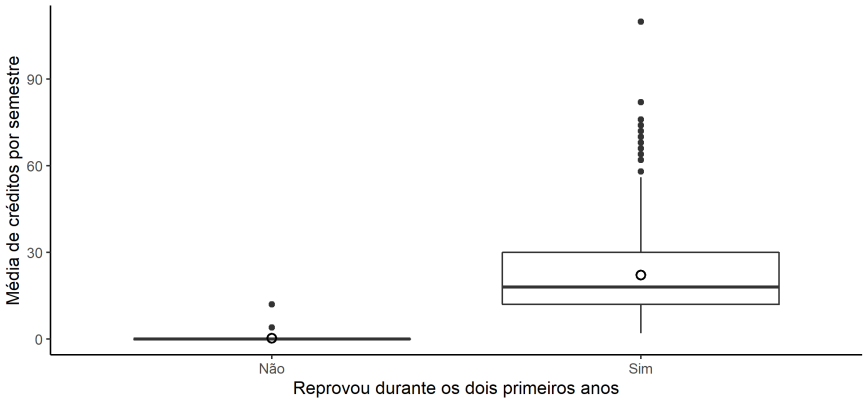

Figura 41: Boxplot da distribuição da soma de créditos reprovados por grupo de alunos que reprovaram nos dois primeiros anos ou não

D Modelo univariado da população de alunos do sexo masculino

Tabela 37: Coeficientes estimados, erro padrão, estatística do teste e p-valor dos modelos contendo apenas uma variável explicativa qualitativa para o banco de dados de alunos do sexo feminino

Variáveis	Estimativa	Erro Padrão	Estatística do teste	P-valor
IRA	0.4061	0.0218	18.6484	<0,0001
Qtd. de reprovações	0.0527	0.0094	5.5822	<0,0001
Qtd. de disciplinas cursadas	0.0638	0.0027	23.5076	<0,0001
Qtd. de trancamentos	0.1867	0.0175	10.6958	<0,0001
Média de créditos cursados p/ semestre	0.0371	0.0093	3.9899	<0,0001
Dif. entre a entrada na UnB e no curso (semestres)	0.1627	0.1190	1.3664	0.1718
Idade	-0.0159	0.0039	-4.0995	<0,0001
Distância da residência até a UnB (metros)	-0.0000	0.0000	-0.0143	0.9886
Soma de créditos reprovados	0.0129	0.0021	6.0885	<0,0001
Soma de créditos cursados	0.0152	0.0006	23.4305	<0,0001
Soma de créditos trancados	0.0416	0.0040	10.2754	<0,0001
Proporção de créditos reprovados	-1.5656	0.0880	-17.7935	<0,0001

Tabela 38: Coeficientes estimados, erro padrão, estatística do teste e p-valor dos modelos contendo apenas uma variável explicativa qualitativa para o banco de dados de alunos do sexo masculino

Variáveis	Estimativa	Erro Padrão	Estatística do teste	P-valor
Sistema de cota				
Sim	-0.0608	0.0777	-0.7831	0.4336
Escola				
Pública	-0.0047	0.0676	-0.0699	0.9442
Forma de ingresso				
Programa de Avaliação Seriada	0.3369	0.1112	3.0302	0.0024
Sisu-Sistema de Seleção Unificada	0.0868	0.1122	0.7737	0.4391
Vestibular	0.2958	0.0963	3.0722	0.0021
Cursou verão				
Sim	0.787	0.0903	8.7167	<0,0001
Currículo				
Velho	-0.2475	0.0714	-3.4677	5e-04
Distância da residência até a UnB (metros)				
De 7089.43 até 14044.27	-0.0465	0.0946	-0.4914	0.6231
De 14044.27 até 23457.23	0.0493	0.096	0.5132	0.6078
Maior que 23457.23	0.0681	0.0963	0.7065	0.4799
Reprovou durante os 2 primeiros anos				
Sim	-0.0705	0.1318	-0.5348	0.5928

E Modelo univariado da população de alunos do sexo feminino

Tabela 39: Coeficientes estimados, erro padrão, estatística do teste e p-valor dos modelos contendo apenas uma variável explicativa quantitativa para o banco de dados de alunos do sexo feminino

Variáveis	Estimativa	Erro Padrão	Estatística do teste	P-valor
IRA	0.2936	0.0698	4.2069	<0,0001
Qtd. de reprovações	0.0457	0.0220	2.0807	0.0375
Qtd. de disciplinas cursadas	0.0492	0.0060	8.1930	<0,0001
Qtd. de trancamentos	0.1890	0.0372	5.0815	<0,0001
Média de créditos cursados p/ semestre	0.0210	0.0227	0.9236	0.3557
Dif. entre a entrada na UnB e no curso (semestres)	0.0924	0.0925	0.9980	0.3183
Idade	-0.0140	0.0109	-1.2823	0.1998
Distância da residência até a UnB (metros)	-0.0000	0.0000	-0.2020	0.8399
Soma de créditos reprovados	0.0097	0.0048	2.0195	0.0434
Soma de créditos cursados	0.0116	0.0014	8.1935	<0,0001
Soma de créditos trancados	0.0433	0.0081	5.3204	<0,0001
Proporção de créditos reprovados	-1.2087	0.2797	-4.3216	<0,0001

Tabela 40: Coeficientes estimados, erro padrão, estatística do teste e p-valor dos modelos contendo apenas uma variável explicativa qualitativa para o banco de dados de alunos do sexo feminino

Variáveis	Estimativa	Erro Padrão	Estatística do teste	P-valor
Sistema de cota				
Sim	0.0632	0.2176	0.2907	0.7713
Escola				
Pública	-0.3301	0.1682	-1.9625	0.0497
Forma de ingresso				
Programa de Avaliação Seriada	0.4226	0.2559	1.651	0.0987
Sisu-Sistema de Seleção Unificada	-0.0701	0.3044	-0.2303	0.8178
Vestibular	0.2062	0.2468	0.8355	0.4035
Cursou verão				
Sim	0.6883	0.1805	3.8126	1e-04
Currículo				
Velho	-0.2754	0.1827	-1.5072	0.1318
Distância da residência até a UnB (metros)				
De 7089.43 até 14044.27	0.3823	0.2731	1.3999	0.1615
De 14044.27 até 23457.23	-2e-04	0.2235	-7e-04	0.9994
Maior que 23457.23	0.0755	0.2352	0.3211	0.7481
Reprovou durante os 2 primeiros anos				
Sim	0.0679	0.2744	0.2476	0.8044