



Universidade Federal de Uberlândia
Faculdade de Matemática

Bacharelado em Estatística

**ANÁLISE DE SOBREVIVÊNCIA
APLICADA À EVASÃO DE ALUNOS DE
GRADUAÇÃO**

Larissa Vitória Silva Araújo

Uberlândia-MG
2020

Larissa Vitória Silva Araújo

**ANÁLISE DE SOBREVIVÊNCIA
APLICADA À EVASÃO DE ALUNOS DE
GRADUAÇÃO**

Trabalho de conclusão de curso apresentado à Coordenação do Curso de Bacharelado em Estatística como requisito parcial para obtenção do grau de Bacharel em Estatística.

Orientador: Maria Imaculada de Sousa Silva

**Uberlândia-MG
2020**



**Universidade Federal de Uberlândia
Faculdade de Matemática**

Coordenação do Curso de Bacharelado em Estatística

A banca examinadora, conforme abaixo assinado, certifica a adequação deste trabalho de conclusão de curso para obtenção do grau de Bacharel em Estatística.

Uberlândia, _____ de _____ de 20____

BANCA EXAMINADORA

Maria Imaculada de Sousa Silva

Leandro Alves Pereira

Priscila Neves Faria

**Uberlândia-MG
2020**

AGRADECIMENTOS

Agradeço primeiramente à Deus que me permitiu viver tudo isso.

Agradeço à minha mãe Eliézer, ao meu pai Jaine e aos meus irmãos por estarem sempre presentes na minha vida, por todo o apoio, carinho, incentivo e por sempre acreditarem em mim.

Agradeço aos meus familiares e à todos que de alguma forma me apoiaram nessa trajetória.

À Universidade Federal de Uberlândia que tornou possível minha graduação. Sou grata à cada membro do corpo docente, à coordenação e à administração dessa instituição de ensino.

Agradeço à minha orientadora, Prof^a Maria Imaculada de Sousa Silva, pela confiança, dedicação e orientação. Ao Prof^o Leandro Alves Pereira e à Prof^a Priscila Neves Faria pela disponibilidade em participar da banca examinadora deste trabalho.

RESUMO

A evasão nos cursos de graduação é uma realidade em grande parte das Instituições de Ensino Superior, tanto públicas quanto privadas, com seus altos índices alarmantes. Nesse sentido, o presente trabalho buscou estudar o tempo até a ocorrência de evasão dos alunos de graduação da Universidade Federal de Uberlândia (UFU) por meio das técnicas de análise de sobrevivência. Os dados referem-se à situação dos alunos dos cursos situados nos campi Santa Mônica e Umuarama, na cidade de Uberlândia. As covariáveis inicialmente consideradas foram turno, área do conhecimento, forma de ingresso, naturalidade, sexo, idade e CRA. A princípio estimou-se a curva de sobrevivência por meio do estimador de Kaplan-Meier e comparou-se as curvas de cada covariável por meio do teste de *logrank*. O modelo de Cox foi ajustado para os dois campi, entretanto, a suposição de taxas de falha proporcionais não foi atendida para algumas covariáveis. Como alternativa ajustou-se os modelos de riscos aditivos que permite parte das covariáveis variem com o tempo e parte não. De maneira geral, os resultados foram concordantes para os dois campi, como por exemplo, a taxa de evasão para ingressantes do SISU é maior quando comparado ao vestibular, o sexo masculino apresentou maior taxa de desistência, o aumento da idade provoca aumento na taxa de evasão e a taxa de evasão diminui com o aumento do CRA. A covariável naturalidade foi a única que apresentou resultado divergente, no campus Umuarama a taxa de desistência é menor para os alunos naturais de Uberlândia, enquanto que no campus Santa Mônica a taxa é maior.

Palavras-chave: Vestibular; Kaplan-Meier; Modelo de Cox; Modelo de Aalen; Modelo semi-paramétrico.

ABSTRACT

Evasion in undergraduate courses is a reality in most Higher Education Institutions, both public and private, with their alarmingly high rates. In this sense, the present work to study the time until the evasion of undergraduate students at the Federal University of Uberlândia (UFU) through the techniques of survival analysis. The data set represents the situation of students in the courses located on the Santa Mônica and Umuarama campuses, in the city of Uberlândia. The covariates set consisted area of knowledge the course, from of entry in course, place of birth, sex, age of students and the academic performance coefficient of student, named CRA (coeficiente de rendimento acadêmico). Firstly, the survival curve was estimated using the Kaplan-Meier estimator and the curves of each covariate were compared using the *logrank* test. Cox's model was fitted for both campuses. Some covariates presented violation of the proportional hazards assumption, according to the test, but with graphical analysis the assumption was accepted. Alternatively, Aalen's additive model and semiparametric additive risk model were fitted, allowing time-dependent covariates and also time-independent. In general, the results were consistent for both campuses, for example, the dropout rate for SISU students is higher when compared to the entrance exam, the male showed a higher dropout rate, increasing age causes an increase in the dropout rate and the rate of permanent evasion with the increase in CRA. The place of birth covariate was the only one that presented a divergent result, on the Umuarama campus the dropout rate is lower for students from Uberlândia, while on the Santa Monica campus the rate is higher.

Keywords: University entrance exams; Kaplan-Meier; Cox model; Aalen's model; Semiparametric model.

SUMÁRIO

Lista de Figuras	I
Lista de Tabelas	III
1 Introdução	1
2 Fundamentação Teórica	3
2.1 Conceitos Básicos em Análise de Sobrevivência	3
2.2 Tempo de Sobrevivência	4
2.2.1 Função de Sobrevivência	4
2.2.2 Função de Taxa de Falha	4
2.3 Estimação Não Paramétrica da Função de Sobrevivência	5
2.3.1 Estimador de Kaplan-Meier	5
2.4 Modelos de Regressão em análise de sobrevivência	5
2.4.1 Modelo de Regressão de Cox	6
2.4.2 Modelo Aditivo de Aalen	8
2.4.3 Modelo de Riscos Aditivos Semiparamétrico	8
3 Metodologia	11
4 Resultados	15
4.1 Campus Santa Mônica	15
4.1.1 Análise Exploratória dos Dados	15
4.1.2 Resultados do Modelo de Regressão de Cox	18
4.1.3 Resultados dos Modelos Aditivos	22
4.2 Campus Umuarama	23
4.2.1 Análise Exploratória dos Dados	23
4.2.2 Resultados do Modelo de Regressão de Cox	25
4.2.3 Resultados dos Modelos Aditivos	29
5 Considerações Finais	33
Referências Bibliográficas	35

LISTA DE FIGURAS

4.1	Curva de sobrevivência estimada a partir do método de Kaplan-Meier para os dados do Campus Santa Mônica.	16
4.2	Curvas de sobrevivência estimadas a partir do método de Kaplan-Meier para cada covariável.	17
4.3	$\text{Log}(\hat{\Lambda}_{0j}(t))$ versus os tempos para as covariáveis do modelo de Cox.	20
4.4	Resíduos padronizados de Schoenfeld versus os tempos para as covariáveis do modelo de Cox.	20
4.5	Análise dos resíduos de Cox-Snell e <i>deviance</i> do modelo de Cox.	21
4.6	Estimativas das Funções de Regressão Acumulada (FRA) e seus respectivos intervalos de 95% de confiança associados ao modelo de aditivo semiparamétrico ajustado para os dados do Campus Santa Mônica.	23
4.7	Curva de sobrevivência estimada a partir do método de Kaplan-Meier para os dados do Campus Umuarama.	25
4.8	Curvas de sobrevivência estimadas a partir do método de Kaplan-Meier para cada covariável.	26
4.9	$\text{Log}(\hat{\Lambda}_{0j}(t))$ versus os tempos para as covariáveis do modelo de Cox.	27
4.10	Resíduos padronizados de Schoenfeld versus os tempos para as covariáveis do modelo de Cox.	28
4.11	Análise dos resíduos de Cox-Snell e <i>deviance</i> do modelo de Cox.	29
4.12	Estimativas das Funções de Regressão Acumulada (FRA) e seus respectivos intervalos de 95% de confiança associados ao modelo de aditivo semiparamétrico ajustado para os dados do Campus Umuarama.	32

LISTA DE TABELAS

3.1	Descrição das covariáveis do conjunto de dados do campus Santa Mônica.	12
3.2	Descrição das covariáveis do conjunto de dados do campus Umuarama.	12
4.1	Frequências absolutas e respectivos percentuais de alunos, de censuras e de falhas referente aos dados do Campus Santa Mônica.	16
4.2	Teste de <i>logrank</i>	17
4.3	Teste da proporcionalidade das taxas de falha no modelo de Cox ajustado para os dados do Campus Santa Mônica.	19
4.4	Resultados do modelo de regressão de Cox ajustado para os dados do Campus Santa Mônica.	21
4.5	Resultados do modelo aditivo de Aalen ajustado para os dados do Campus Santa Mônica.	22
4.6	Frequências absolutas e respectivos percentuais de alunos, de censuras e de falhas referente aos dados do Campus Umuarama.	24
4.7	Teste de <i>logrank</i>	25
4.8	Teste da proporcionalidade das taxas de falha no modelo de Cox ajustado para os dados do Campus Umuarama.	26
4.9	Resultados do modelo de regressão de Cox ajustado para os dados do Campus Umuarama.	29
4.10	Resultados do modelo aditivo de Aalen ajustado para os dados do Campus Umuarama.	30
4.11	Estimativa dos coeficientes com efeito constante no modelo de riscos aditivos semiparamétrico ajustado para os dados do Campus Umuarama.	30

1. INTRODUÇÃO

A evasão é, certamente, um dos maiores desafios que as Instituições de Ensino Superior (IES) vêm enfrentando. Segundo o estudo realizado por Silva Filho et al. [5] os estudantes que ingressam na universidade e não terminam seus cursos geram desperdícios sociais, acadêmicos e econômicos. No setor privado há uma perda de receita e no setor público não há o retorno dos recursos públicos investidos, e ambos são uma fonte de ociosidade de professores, funcionários, equipamentos e espaço físico. Para Lobo [6] o aluno que abandona o curso sem finalizar seus estudos representa uma perda social, de recursos e de tempo de todos os envolvidos no processo de ensino, pois perdeu o aluno, seus professores, a instituição de ensino, o sistema de educação e toda a sociedade.

No Brasil, as taxas de evasão vêm se mantendo em aproximadamente 22% com pequenas variações ao longo dos últimos 15 anos, sendo menor para o setor público e maior para o privado [4]. No trabalho de Lobo [6], a autora trata de questões gerais que envolvem a evasão no Ensino Superior brasileiro e, segundo ela, as causas mais frequente encontradas nas IES são: inadaptação do ingressante ao estilo do ensino superior e a falta de maturidade, formação básica deficiente, dificuldade financeira, irritação com a precariedade dos serviços oferecidos pelas IES, decepção com pouca motivação e atenção dos professores, dificuldades com transporte, alimentação e ambientação nas instituições, mudança de curso e mudança de residência.

Conforme Tigrinho [13] a evasão é resultado de um conjunto de fatores que pesam na decisão do aluno de permanecer ou não no curso. Segundo ele dentre as causas da evasão abordadas na literatura estão a repetência, após a reprovação em uma ou mais disciplinas os alunos são mais propensos a abandonarem seus cursos, a falta de informação sobre a profissão e sobre o curso, a mudança de curso e a desmotivação do aluno. Ainda segundo o autor, na visão dos dirigentes das IES, a falta de orientação vocacional/ profissional, a baixa qualidade da educação básica, busca de herança profissional, problemas financeiros e o horário de trabalho incompatível com o horário de estudo são fatores que contribuem para a evasão do discente.

De acordo com o relatório da Comissão Especial de Estudos sobre a Evasão nas Universidades Brasileiras [7] existem três tipos de evasão: a evasão de curso, quando o estudante desliga-se do curso superior em situações como abandono, desistência, transferência ou reopção, evasão da instituição, quando o aluno desliga-se da instituição na qual está matriculado e evasão do sistema, quando o estudante abandona de forma definitiva ou temporária o ensino superior.

Tendo em vista que o objetivo do estudo da evasão dos alunos de graduação é avaliar o tempo até a desistência do aluno e seus fatores associados, alguns estudos foram realizados

utilizando métodos estatísticos para análise de dados de sobrevivência. Queiroz [9] utilizou a análise de sobrevivência para caracterizar a evasão dos cursos do Centro de Ciências Exatas e Tecnológicas UFRB, especificamente dos cursos de bacharelado em Ciências Exatas e Tecnológicas (BCET) e de Engenharia Sanitária e Ambiental (ESA). Os resultados da análise de sobrevivência mostraram que alunos do curso do BCET possuem maior probabilidade de cancelar do que os do ESA. Os resultados apontaram também que a chance de ser cancelado aumenta em 24% quando o estudante é do sexo masculino e alunos que estudaram em escola pública as chances aumentam em 15% em relação ao discente que estudou em escola particular.

Costa, Siqueira e Rocha [3] analisaram o comportamento da evasão nas universidades federais brasileiras no período de 2011 a 2016 por meio da análise de sobrevivência. O modelo de regressão de Cox foi ajustado para identificar as variáveis que estavam relacionadas ao fenômeno da evasão. Nesse estudo observou-se que 41,2% dos ingressantes de 2011 havia evadido do curso até 2016. Os resultados concluíram que no primeiro ano alunos do sexo masculino e feminino apresentaram a mesma taxa de evasão, entretanto, no decorrer do tempo houve aumento dessa diferença. Quanto ao grupo de beneficiário de programa de apoio social a taxa de sobrevivência foi de 70,3% enquanto para os demais a taxa foi de 53,2%. Concluiu-se ainda que evadem mais alunos da licenciatura quando comparados aos cursos de bacharelado e cursos noturno apresentam maior taxa de evasão comparado aos demais. A participação em atividades complementares mostrou forte influência sobre o risco de evadir, participar de atividades de pesquisa e extensão reduz drasticamente o risco do aluno abandonar o curso.

Outro estudo sobre os fatores explicativos para evasão no ensino superior através da análise de sobrevivência foi realizado por Silva Campos [1] para o caso da Universidade Federal de Pernambuco, no período de 2003 a 2014. O tempo mediano de permanência foi de 10 semestres até a evasão, ou seja, 5 anos de permanência. As variáveis significantes no modelo de Cox para explicar as chances de evasão foram o percentual da carga horária aprovada dos alunos e a área do conhecimento. Os resultados mostram alunos com percentual de aprovação de até 10% em disciplinas aumenta em 12 vezes a chance de evadir dos alunos evadirem. Os alunos das áreas de ciências exatas e tecnologia têm 1,3 vezes mais chances de evadir do que os alunos de outras áreas do conhecimento.

Nesse contexto, os dados de alunos dos cursos de graduação da Universidade Federal de Uberlândia foram analisados neste trabalho com o objetivo de estudar o tempo até a evasão dos estudantes desta universidade bem como possíveis fatores que possam estar relacionadas com o tempo de permanência. Com esse propósito foram utilizados métodos estatísticos de análise de sobrevivência.

O presente trabalho foi estruturado em quatro capítulos. No capítulo 2 é apresentado a fundamentação teórica onde é feita uma descrição dos conceitos básicos de análise de sobrevivência bem como os métodos estatísticos disponíveis na literatura para esse tipo de dados. No capítulo 3 são descritos os dados utilizados e a metodologia empregada. No capítulo 4 é apresentado os resultados dos modelos ajustados. No capítulo 5 são apresentadas as considerações finais das análises deste trabalho.

2. FUNDAMENTAÇÃO TEÓRICA

Nesta seção são apresentados os conceitos básicos de análise de sobrevivência bem como métodos e modelos de regressão para dados de sobrevivência.

2.1 CONCEITOS BÁSICOS EM ANÁLISE DE SOBREVIVÊNCIA

De acordo com Colosimo e Giolo [2] a análise de sobrevivência é uma técnica em que o interesse é avaliar o tempo até a ocorrência de um evento. Este tempo é denominado tempo de falha. Em geral, os estudos de análise de sobrevivência são prospectivos e de longa duração, portanto, são caracterizados pela presença de censura. A censura ocorre quando, por alguma razão, o acompanhamento do indivíduo é interrompido ou o evento não ocorre até o fim do estudo. As observações censuradas são incluídas nas análises, uma vez que, mesmo incompletas fornecem informações sobre o tempo de falha e a omissão poderá gerar vícios.

A censura é dividida em censura à direita, censura à esquerda e censura intervalar. A censura à direita refere-se à observação parcial quando o tempo de ocorrência do evento de interesse está à direita do tempo registrado. Esse tipo de censura é subdividida em outros três tipos de censura: censura do tipo I, censura do tipo II e censura do tipo aleatória. A censura do tipo I ocorre quando os estudos são finalizados após um período pré-estabelecido de tempo e alguns indivíduos ainda não apresentaram o evento de interesse. A censura do tipo II resultam de estudos os quais são finalizados após a ocorrência do evento de interesse em um número pré-estabelecido de indivíduos. A censura do tipo aleatória ocorre quando o indivíduo é retirado no decorrer do estudo sem que tenha ocorrido a falha.

Na censura à esquerda não se sabe exatamente quando o evento ocorreu, já que ocorreu antes do indivíduo ser observado no estudo. A censura intervalar sabe-se que o evento ocorreu, mas não se sabe exatamente quando, somente que aconteceu em um intervalo de tempo.

Em geral, os dados de sobrevivência são representados pelo par (t_i, δ_i) , para o indivíduo i , sendo t_i o tempo de falha ou de censura e δ_i a variável indicadora de falha ou censura, ou seja,

$$\delta_i = \begin{cases} 1 & \text{se } t_i \text{ é um tempo de falha} \\ 0 & \text{se } t_i \text{ é um tempo censurado.} \end{cases} \quad (2.1)$$

Desse modo, a variável resposta é composta por duas colunas no banco de dados. Quando

houver covariáveis, os dados ficam representados por $(t_i, \delta_i, \mathbf{x}_i)$, em que $\mathbf{x}_i$ é o vetor de covariáveis para o i -ésimo indivíduo [2].

2.2 TEMPO DE SOBREVIVÊNCIA

Em geral, os dados de sobrevivência são representados pela variável aleatória não-negativa T , usualmente contínua, que representa o tempo de falha. Em análise de sobrevivência, essa variável é geralmente caracterizada pela sua função de sobrevivência ou pela função de taxa de falha.

2.2.1 FUNÇÃO DE SOBREVIVÊNCIA

Segundo Colosimo e Giolo [2] a função de sobrevivência, denotada por $S(t)$, é definida como a probabilidade de um indivíduo não falhar até um determinado tempo t , ou seja, a probabilidade de o indivíduo sobreviver ao tempo t , expressa por:

$$S(t) = P(T > t) \quad (2.2)$$

Consequentemente a distribuição acumulada, que corresponde à probabilidade de um indivíduo não sobreviver ao tempo t , assume a seguinte forma :

$$F(t) = 1 - S(t) \quad (2.3)$$

Para descrever a função de sobrevivência é utilizada a curva de sobrevivência que é uma representação gráfica de $S(t)$, ou seja, o gráfico de $S(t)$ versus t . Uma curva íngreme representa razão de sobrevivência baixa e uma curva de sobrevivência gradual ou plana representa taxa de sobrevivência alta. A curva de sobrevivência pode ser usada para comparar distribuições de sobrevivência de dois ou mais grupos [8].

2.2.2 FUNÇÃO DE TAXA DE FALHA

A função de taxa de falha, ou de risco, do tempo de sobrevivência T fornece a taxa de falha condicional. Essa função é definida como a probabilidade de ocorrência do evento de interesse dentro de um intervalo pequeno de tempo $[t, t + \Delta t)$, supondo que o indivíduo sobreviveu até o início do intervalo. Essa função é muito útil para descrever a distribuição do tempo de vida, uma vez que descreve a forma com que a taxa instantânea de falha muda com o tempo. Dessa forma, possui a seguinte expressão:

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \quad (2.4)$$

2.3 ESTIMAÇÃO NÃO PARAMÉTRICA DA FUNÇÃO DE SOBREVIVÊNCIA

As utilização de técnicas usuais de estatística para uma análise descritiva dos dados pode gerar dificuldades com a presença de observações censuradas. Com isso, a análise descritiva para dados de sobrevivência é realizada por meio da função de sobrevivência. Desse modo a função de sobrevivência é estimada e, a partir dela, são estimadas as estatísticas de interesse. Para isso adota-se um estimador não-paramétrico, como o estimador de Kaplan-Meier.

2.3.1 ESTIMADOR DE KAPLAN-MEIER

O estimador de Kaplan-Meier é o mais utilizado na literatura para estimar a função de sobrevivência. Para utilizar este estimador não é necessário assumir nenhuma distribuição de probabilidade para a variável aleatória T , ou seja, é um estimador não paramétrico [12]. O estimador de Kaplan-Meier é definido como:

$$\hat{S}(t) = \prod_{j:t_j < t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j < t} \left(1 - \frac{d_j}{n_j} \right), \quad (2.5)$$

em que $t_1 < t_2 < \dots < t_k$ são os k tempos distintos e ordenados de falha, d_j o número de falhas em t_j e n_j é o número de indivíduos sob risco em t_j , para $j = 1, \dots, k$.

Utilizando este modelo é possível comparar curvas de sobrevivência associadas às categorias de cada covariável utilizada, através do teste de *logrank* [2]. Este teste é particularmente apropriado quando a razão das funções de taxa de falha dos grupos a serem comparados é aproximadamente constante, ou seja, as populações têm a propriedade de taxa de falhas proporcionais.

Apesar de as técnicas não paramétricas serem importantes para análise de dados de sobrevivência, existe a limitação quando há interesse em considerar um conjunto de covariáveis. Para isso adota-se modelos de regressão apropriados, descritos nas seções seguintes.

2.4 MODELOS DE REGRESSÃO EM ANÁLISE DE SOBREVIVÊNCIA

Em muitos estudos as covariáveis podem estar relacionadas com o tempo de sobrevivência, e para isso, alguns modelos de regressão apropriados para dados censurados têm sido propostos na literatura. Neste trabalho a ênfase é sobre o modelo de regressão de Cox e os modelos de riscos aditivos.

2.4.1 MODELO DE REGRESSÃO DE COX

O modelo de regressão de Cox permite a análise de tempo de vida em que a resposta é a ocorrência de um evento de interesse, ajustando por covariáveis. A expressão do modelo de Cox em termos da função de taxa de falha é dada, para o i -ésimo indivíduo em um dado tempo t , por:

$$\lambda(t|\mathbf{x}_i) = \lambda_0(t) \exp \left(\sum_{k=1}^p \beta_k x_{ik} \right), \quad (2.6)$$

em que p é o número de covariáveis, x_{ik} é o valor da k -ésima covariável observada para o i -ésimo indivíduo, β_k é o parâmetro que descreve o efeito da k -ésima covariável e $\lambda_0(t)$ é uma função do tempo não-negativa e não especificada, geralmente denominada função taxa de falha de base, que corresponde a taxa comum a todos os indivíduos [12].

AValiação da Proporcionalidade do Modelo

Também conhecido como modelo de taxas de falha proporcionais, a suposição básica para este modelo é que as taxas de falha sejam proporcionais, ou seja, a razão das funções de taxa de falha para os indivíduos i e j dada por:

$$\frac{\lambda(t|\mathbf{x}_i)}{\lambda(t|\mathbf{x}_j)} = \frac{\lambda_0(t) \exp\{\mathbf{x}_i' \beta\}}{\lambda_0(t) \exp\{\mathbf{x}_j' \beta\}} = \exp\{\mathbf{x}_i' \beta - \mathbf{x}_j' \beta\}, \quad (2.7)$$

não depende do tempo. Por exemplo, se um indivíduo no início do estudo tem uma taxa de falha igual a duas vezes a de um segundo indivíduo, então esta razão será a mesma para todo o período de acompanhamento [2].

Para verificar a suposição de taxas de falha proporcionais assumida pelo modelo de regressão de Cox utiliza-se métodos gráficos e testes estatísticos. O primeiro método gráfico descritivo utilizado neste trabalho consiste em dividir os dados em m estratos de acordo com cada covariável. Após isso, estima-se a função de taxa de falha acumulada de base $\hat{\Lambda}_{0j}(t)$ para cada estrato utilizando a expressão:

$$\hat{\Lambda}_{0j}(t) = \sum_{j:t_j < t} \frac{d_j}{\sum_{l \in R_j} \exp\{\mathbf{x}_l' \hat{\beta}\}}, \quad (2.8)$$

em que d_j é o número de falhas em t_j . Se a suposição de taxas de falha proporcionais for válida, as curvas de $\log(\hat{\Lambda}_{0j}(t))$ versus t , apresentarão diferenças aproximadamente constantes no tempo. Curvas não paralelas indicam desvios desta suposição [12].

Outro método gráfico utilizado para avaliar a suposição de taxas de falhas proporcionais no modelo de Cox é o gráfico dos resíduos padronizados de Schoenfeld versus os tempos. Os

resíduos padronizados de Schoenfeld é definido por:

$$\mathbf{s}_i^* = \left[I(\hat{\beta}) \right]^{-1} \times \mathbf{r}_i, \quad (2.9)$$

onde $I(\hat{\beta})$ é a matriz de informação observada e $\mathbf{r}_i = (r_{i1}, \dots, r_{ip})$ é o vetor de resíduos Schoenfeld, em que cada componente r_{iq} , para $q = 1, \dots, p$ definido por:

$$r_{iq} = x_{iq} - \frac{\sum_{j \in R(t_i)} x_{jq} \exp\{\mathbf{x}'_j \hat{\beta}\}}{\sum_{j \in R(t_i)} \exp\{\mathbf{x}'_j \hat{\beta}\}}. \quad (2.10)$$

O gráfico de $\mathbf{s}_{iq}^* + \hat{\beta}_q$ *versus* o tempo, ou uma função do tempo $g(t)$, mostra evidências a favor da proporcionalidade quando este apresenta inclinação zero [2].

Uma medida também empregada com este propósito é o coeficiente de correlação de Pearson (ρ) entre os resíduos padronizados de Schoenfeld e os tempos t ou uma função dos tempos $g(t)$, para cada covariável. Valores de ρ próximos de zero mostram não haver evidências para rejeitar esta suposição [2]. Quando ocorre a violação desta suposição busca-se modelos alternativos como modelo aditivo de Aalen, o qual permite que os parâmetros e vetores de covariáveis variem com o tempo.

ADEQUAÇÃO DO MODELO

Alguns métodos foram propostos para verificar a qualidade de ajuste global deste modelo. Dentre eles está o método baseado nos resíduos de Cox-Snell, definidos por:

$$\hat{e}_i = \hat{\Lambda}(t_i | \mathbf{x}_i) = \hat{\Lambda}_0(t_i) \exp \left\{ \sum_{k=1}^p x_{ik} \hat{\beta}_k \right\} \quad (2.11)$$

em que $\hat{\Lambda}(\cdot)$ é a função de taxa de falha acumulada obtida a partir do modelo ajustado. Dessa forma, o modelo é bem ajustado quando os $\hat{e}'_i$ seguirem uma distribuição exponencial padrão. Da mesma maneira, o gráfico de $\hat{\Lambda}(\hat{e}_i)$ *versus* $\hat{e}_i$ deve ser aproximadamente uma reta com inclinação 1 para ser adequado [2] [12].

Além de avaliar o ajuste geral dos modelos, também é importante avaliar aspectos como a observação de pontos atípicos ou influentes. Com esse propósito, o gráfico dos resíduos *martingal* e *deviance* *versus* os tempos são utilizados. A ausência de observações atípicas ficará evidenciada se os resíduos $\hat{d}_i$ estiverem distribuídos aleatoriamente em torno do valor zero, geralmente dentro do intervalo $(-3, 3)$. [11]. Os resíduos *martingal* e *deviance* são definidos por:

$$\hat{m}_i = \delta_i - \hat{e}_i \quad (2.12)$$

e

$$\hat{d}_i = \text{sinal}(\hat{m}_i) [-2 (\hat{m}_i + \delta_i \log(\delta_i - \hat{m}_i))]^{1/2}, \quad (2.13)$$

para $i = 1, \dots, n$ em que δ_i é a variável indicadora de falha.

2.4.2 MODELO ADITIVO DE AALEN

O modelo de riscos aditivos proposto por Aalen é uma alternativa nos casos em que a suposição de taxas de falhas proporcionais não é satisfeita. Este modelo permite que o efeito de todas as covariáveis varie com o tempo. A função de taxa de falha para o tempo de sobrevivência t é dada por:

$$\lambda(t|\mathbf{x}_i(t)) = \beta_0(t) + \sum_{j=1}^p \beta_j(t)x_{ij}(t), \quad (2.14)$$

em que p é o número de covariáveis, x_{ij} é o valor da j -ésima covariável observada para o i -ésimo indivíduo, β_j é o parâmetro que descreve o efeito da j -ésima covariável, onde x_{ij} e β_j variam com o tempo. O componente β_0 corresponde à função taxa de falha base.

Este modelo é considerado não paramétrico e, conforme Santos e Werner [12], o modelo de Aalen permite que tanto o efeito quanto os valores das covariáveis variem com o tempo e assume que elas atuam de forma aditiva sobre a função taxa de falha. Também é denominado modelo dinâmico por permitir que as funções de regressão $\beta(t)$ e as covariáveis $\mathbf{x}_i(t)$ variem com o tempo.

É de interesse também testar se alguma covariável específica apresenta efeito não significativo ou efeito constante. Para isso é realizado os testes: (a) efeito não significativo das covariáveis, ou seja, $H_{0a} : \beta_k(t) = 0$ e (b) efeito constante das covariáveis, ou seja, $H_{0b} : \beta_k(t) = \beta_k$.

O principal foco do modelo são os gráficos de regressão acumulados $\hat{B}_k(t), k = 0, \dots, p$, *versus* os tempos. Por meio da inclinação observada nos gráficos é possível verificar se determinada covariável apresenta efeito constante ou tempo-dependente. Quando a inclinação é positiva em certo período indica que a covariável exerce efeito crescente sobre a taxa de falha e decrescente quando a inclinação é negativa. Inclinação próxima de zero é observada quando a covariável não tem efeito sobre a taxa de falha [12] [11].

2.4.3 MODELO DE RISCOS ADITIVOS SEMIPARAMÉTRICO

Em alguns casos pode ocorrer que algumas covariáveis apresentem efeito variando com o tempo e outras não. Para tal, encontra-se uma extensão do modelo aditivo de Aalen permitindo que parte dos coeficientes variem com o tempo e parte não, denominado modelo de riscos aditivos semiparamétrico [12] [11]. Esse modelo é expresso por:

$$\lambda(t|\mathbf{x}_i(t), \mathbf{z}_i(t)) = \beta_0(t) + \mathbf{x}_i^T(t)\beta(t) + \mathbf{z}_i^T(t)\gamma, \quad (2.15)$$

em que $\mathbf{x}_i(t) = (x_{i1}(t), \dots, x_{ip}(t))$ e $\mathbf{z}_i(t) = (z_{i1}(t), \dots, z_{il}(t))$ são vetores de covariáveis de dimensão p e l , assim como $\beta(t) = (\beta_1(t), \dots, \beta_p(t))$ e $\gamma = (\gamma_1, \dots, \gamma_l)$ são os vetores de coeficientes de regressão dependentes e independentes do tempo, respectivamente.

3. METODOLOGIA

Os dados utilizados neste trabalho foram disponibilizados pela Pró-Reitoria de Graduação (PROGRAD) da Universidade Federal de Uberlândia e referem-se à situação dos alunos dos cursos de graduação situados nos Campus Santa Mônica e Umuarama na cidade de Uberlândia. As informações coletadas são dos ingressantes do Campus Santa Mônica do ano de 2015, totalizando 3250 alunos, e dos ingressantes do Campus Umuarama dos anos de 2012 a 2014, totalizando 3893 alunos.

A técnica de análise de sobrevivência é utilizada quando o interesse é avaliar o tempo até a ocorrência de um evento, denominado tempo de falha. Neste estudo, a análise de sobrevivência foi aplicada com o objetivo de avaliar o tempo até a evasão dos alunos de graduação. Neste contexto, a evasão é considerada quando, por alguma razão, o aluno é desvinculado do curso de graduação, seja por abandono ou desistência oficial. O tempo de permanência foi contabilizado em semestres desde o ingresso do aluno até o semestre de evasão, formatura ou até o segundo semestre de 2019 para aqueles que ainda estavam matriculados. Em geral, estudos de análise de sobrevivência são caracterizados pela presença de censuras, neste caso, as observações censuradas foram aquelas em que os alunos se formaram até o segundo semestre de 2019 ou ainda estavam vinculados com a instituição.

As covariáveis utilizadas foram: turno de estudos e a área do conhecimento do curso no qual o aluno se matriculou, o sexo, a naturalidade e a idade do aluno, a forma de ingresso e o coeficiente de rendimento acadêmico (CRA) do aluno no curso. As categorias de cada covariável estão descritas nas Tabelas 3.1 e 3.2 para os campi Santa Mônica e Umuarama respectivamente.

Inicialmente foi realizada a análise exploratória dos dados com o objetivo de descrever e resumir os dados observados. O tempo até a ocorrência do evento foi avaliado por meio das curvas de Kaplan-Meier e a comparação entre elas foi feita com o teste de *logrank*. A princípio os modelos paramétricos foram utilizados ajustando-se os modelos de regressão exponencial, Weibull e lognormal, no entanto, não se mostraram adequados de acordo com teste de razão de verossimilhança. Dessa maneira, foi ajustado o modelo de regressão de Cox, em que a suposição básica para o uso desse modelo é que as taxas de falha sejam proporcionais em todo o período de tempo. Esta suposição pôde ser verificada por meio do coeficiente de correlação de Pearson ρ e por meio de técnicas gráficas. Quando há violação desta suposição os modelos de riscos aditivos são uma alternativa. O modelo aditivo de Aalen foi ajustado e, para os casos de covariáveis tempo-dependentes optou-se pelo modelo aditivo semiparamétrico. Este modelo permite que parte das covariáveis variem com o tempo e parte não.

Tabela 3.1: Descrição das covariáveis do conjunto de dados do campus Santa Mônica.

Covariável	Categoria
Turno	0 - Integral 1 - Noturno
Área do Conhecimento	0 - Multidisciplinar 1 - Ciências Humanas 2 - Ciências Exatas
Sexo	0 - Feminino 1 - Masculino
Naturalidade	0 - Uberlândia 1 - Outros
Forma de Ingresso	0 - Vestibular 1 - SISU 2 - Outros
CRA	0 a 100
Idade	Em anos

Tabela 3.2: Descrição das covariáveis do conjunto de dados do campus Umuarama.

Covariável	Categoria
Turno	0 - Integral 1 - Noturno
Área do Conhecimento	0 - Ciências da Saúde 1 - Ciências Biológicas 2 - Ciências Agrárias 3 - Multidisciplinar
Sexo	0 - Feminino 1 - Masculino
Naturalidade	0 - Uberlândia 1 - Outros
Forma de Ingresso	0 - SISU 1 - Vestibular 2 - PAAES 3 - Outros
CRA	0 a 100
Idade	Em anos

A adequação destes modelos foi avaliada por meio dos resíduos de Cox-Snell que permitem verificar a qualidade geral do modelo. Além disso, também verificou-se a presença de pontos atípicos por meio de técnicas gráficas baseadas nos resíduos *martingal* e *deviance*.

As análises foram realizadas com o *software* R [10], com auxílio dos pacotes *survival* e *timereg*.

4. RESULTADOS

4.1 CAMPUS SANTA MÔNICA

4.1.1 ANÁLISE EXPLORATÓRIA DOS DADOS

Inicialmente foi realizada a análise descritiva do conjunto de dados referente ao Campus Santa Mônica. A Tabela 4.1 apresenta as frequências absolutas de alunos, de censuras e de falhas para as categorias de cada covariável bem como seus respectivos percentuais. Vale lembrar que a falha corresponde à desistência do curso, e a censura corresponde à outras situações como matriculados, trancados e formados. As covariáveis CRA e idade foram categorizadas para facilitar a análise exploratória. Para a covariável turno a categoria integral apresentou maior frequência com 67,8% das observações. A área de ciências humanas corresponde à 57,9% enquanto que a área de ciências exatas corresponde à 40% do número total sobrando apenas 2,1% dos estudantes matriculados em cursos da área multidisciplinar. Estudantes naturais de Uberlândia correspondem a um percentual de 39,7% do total de matriculados, enquanto os outros 60,3% são naturais de outras cidades do país. Dentre as formas de ingresso, o SISU apresenta o maior número de matrículas. A covariável CRA apresentou média de 51,76 e desvio padrão de 30,22. Para a idade, a média foi de 26,44 e desvio padrão de 7,21.

Quanto ao percentual de falhas, a categoria noturno da covariável turno apresentou maior taxa de falha. Para a covariável área do conhecimento observou-se maior taxa de falha para a categoria multidisciplinar e menor taxa de falha para ciências humanas. Já para a covariável sexo, a categoria masculino apresentou maior taxa de falha. Para a covariável forma de ingresso observou-se menor taxa de falha para a categoria vestibular enquanto que a categoria SISU e outras formas de ingresso apresentaram taxas de falha próximas.

Em seguida, a curva de sobrevivência foi obtida por meio do estimador de Kaplan-Meier, como também os intervalos de 95% de confiança, para $0 \leq t \leq 10$, conforme a Figura 4.1. A sobrevivência final foi de 0,619, ou seja, a probabilidade de um estudante não desistir do curso até o tempo 10 semestres é de 61,9%. O tempo médio foi de 7,97 semestres, ou seja, a desistência ocorre em média no tempo 8 semestres, aproximadamente.

Também foi obtida a curva de sobrevivência para cada covariável na Figura 4.2 e o teste de *logrank* na Tabela 4.2 a fim de identificar diferenças significativas entre as curvas para as categorias de cada covariável. Por meio do teste todas as covariáveis apresentaram diferença significativa ao nível de significância 5%. Dessa maneira todas as covariáveis podem estar

Tabela 4.1: Frequências absolutas e respectivos percentuais de alunos, de censuras e de falhas referente aos dados do Campus Santa Mônica.

Covariável	Categoria	N	N%	Censura	Censura%	Falha	Falha%
Turno	Integral	2205	67,8%	1446	65,6%	759	34,4%
	Noturno	1045	32,2%	595	56,9%	450	43,1%
Área do Conhecimento	Multidisciplinar	68	02,1%	32	47,1%	36	52,9%
	Ciências Humanas	1881	57,9%	1249	66,4%	632	33,6%
	Ciências Exatas	1301	40,0%	760	58,4%	541	41,6%
Sexo	Feminino	1364	42,0%	907	66,5%	457	33,5%
	Masculino	1886	58,0%	1134	60,1%	752	39,9%
Naturalidade	Uberlândia	1290	39,7%	780	60,5%	510	39,5%
	Outros	1960	60,3%	1261	64,3%	699	35,7%
Forma de Ingresso	Vestibular	852	26,2%	641	75,2%	211	24,8%
	SISU	1791	55,1%	1047	58,5%	744	41,5%
	Outros	607	18,7%	353	58,1%	254	41,9%
CRA	< 60	1650	50,8%	644	39,0%	1006	61,0%
	≥ 60	1600	49,2%	1397	87,3%	203	12,7%
Idade	19 a 22	1084	33,3%	744	68,6%	340	31,4%
	23 a 24	833	25,6%	568	68,2%	265	31,8%
	25 a 28	599	18,4%	363	60,6%	236	39,4%
	>28	734	22,6%	366	49,8%	369	50,2%

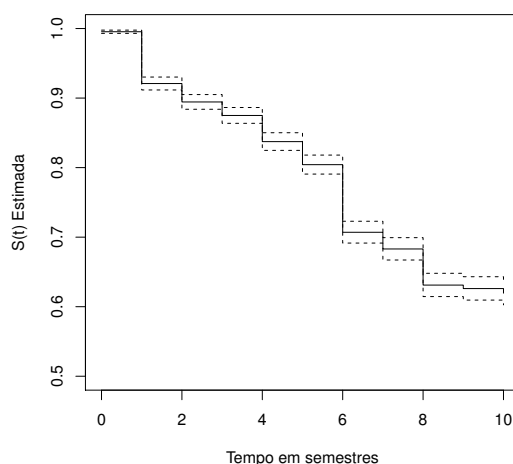


Figura 4.1: Curva de sobrevivência estimada a partir do método de Kaplan-Meier para os dados do Campus Santa Mônica.

influenciando no tempo de desistência dos alunos.

Tabela 4.2: Teste de *logrank*.

Covariável	Qui-quadrado	p-valor
Turno	18,5	0,0001
Área do Conhecimento	26,5	0,0000
Sexo	9,3	0,0010
Naturalidade	2,9	0,0400
Forma de Ingresso	53,4	0,0000

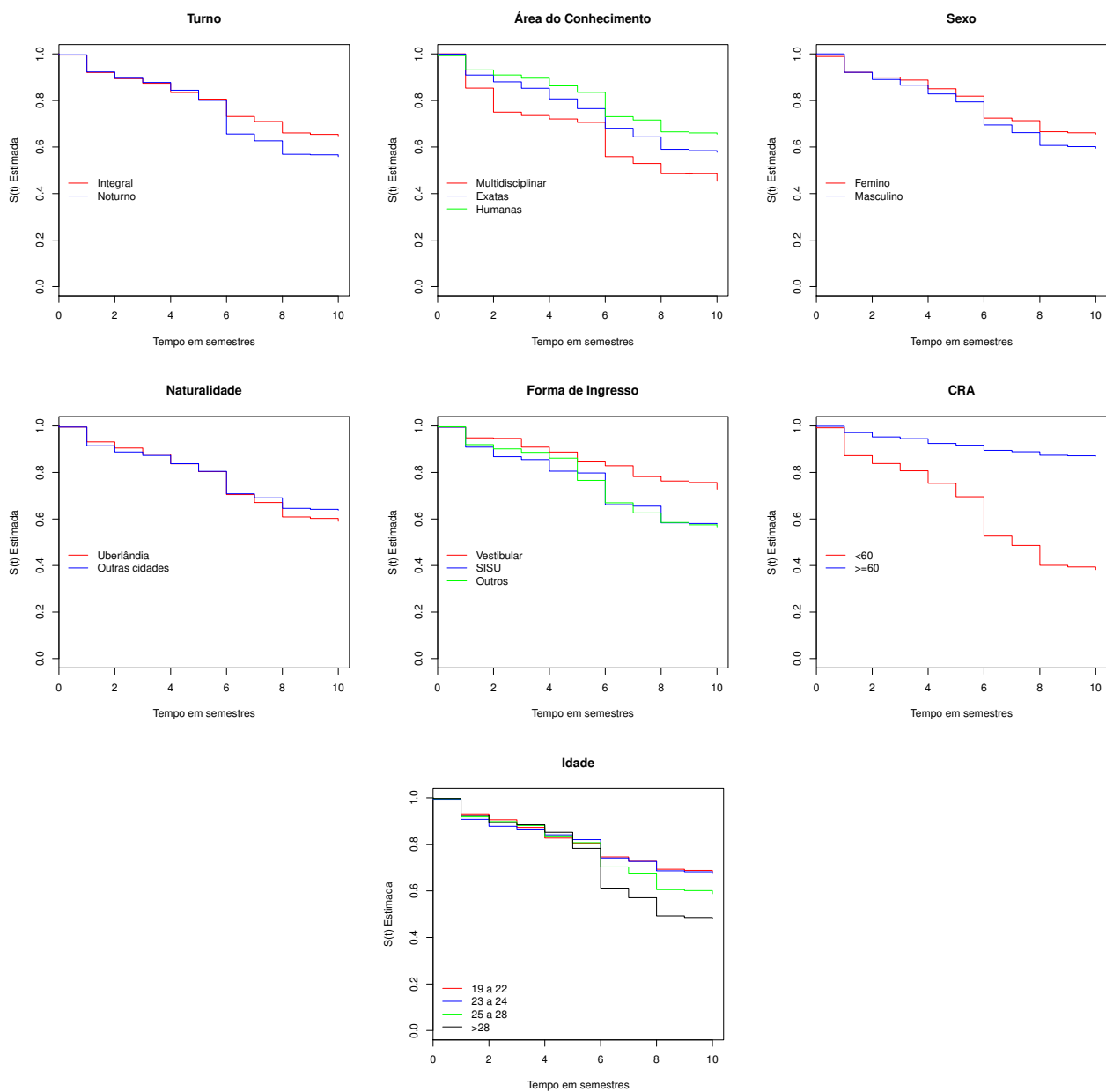


Figura 4.2: Curvas de sobrevivência estimadas a partir do método de Kaplan-Meier para cada covariável.

4.1.2 RESULTADOS DO MODELO DE REGRESSÃO DE COX

Após a análise das possíveis covariáveis a serem consideradas ajustou-se o modelo de regressão de Cox expresso na equação (2.6). Para a seleção das covariáveis ajustou-se primeiramente um modelo com todas as covariáveis, retirando-se uma por vez, aquelas não significativas ao nível 5% de significância, utilizando o teste de razão de verossimilhança para comparar os modelos. O modelo final foi composto pelas covariáveis turno, área do conhecimento, sexo, naturalidade, forma de ingresso e idade. A covariável CRA apresentou efeito significativo na maioria dos modelos testados, no entanto, a presença da covariável no modelo provocou interferência nos resultados das demais covariáveis, mudando a significância de alguns parâmetros, e até mesmo invertendo de positivo para negativo o efeito das covariáveis. Esse fato pode ocorrer na presença de multicolinearidade entre covariáveis, gerando um confundimento dos efeitos dessas variáveis. Apesar de não ser considerada no modelo, a covariável CRA possui um forte efeito sobre a taxa de falha ou sobre a probabilidade de sobrevivência do aluno no curso, como pode ser visto na curva de Kaplan-Meier para esta covariável (Figura 4.2). Fica claro que a probabilidade de sobreviver no curso para alunos com CRA abaixo de 60 é bem menor que para alunos com CRA acima de 60.

Inicialmente a suposição de taxas de falha proporcionais do modelo foi verificada pelo coeficiente de correlação de Pearson ρ . Os resultados na Tabela 4.3 indicam a violação da suposição para as covariáveis naturalidade, idade e turno ao nível de significância de 5%. Para grandes amostras, como neste caso, maior cuidado deve ser dado ao coeficiente de correlação ρ , pois p-valores muito pequenos podem ser obtidos em decorrência do tamanho amostral.

Observando os gráficos do logaritmo da função de taxa de falha acumulada ($\log(\hat{\Lambda}_{0j}(t))$) contra os tempos para cada covariável do modelo na Figura 4.3, percebe-se um cruzamento entre as curvas para a covariável naturalidade. Para as demais covariáveis, embora as mesmas não sejam perfeitamente paralelas ao longo do tempo, não existem afastamentos marcantes entre as curvas para apontar séria violação dessa suposição. Nos gráficos dos resíduos padronizados de Schoenfeld na Figura 4.4 é observada uma tendência nos períodos finais da covariável naturalidade, sendo que para as demais covariáveis do modelo não foram observadas tendências acentuadas que pudessem indicar violação da suposição.

Tabela 4.3: Teste da proporcionalidade das taxas de falha no modelo de Cox ajustado para os dados do Campus Santa Mônica.

Covariável	ρ	χ^2	p-valor
Turno (Noturno)	0,0223	7,47	0,006
Área do Conhecimento (C. Humanas)	-0,0065	0,04	0,833
Área do Conhecimento (C. Exatas)	0,0061	0,23	0,631
Sexo (Masculino)	0,0416	2,99	0,084
Naturalidade (Outros)	-0,0086	12,08	0,001
Forma de Ingresso (SISU)	-0,0501	2,68	0,102
Forma de Ingresso (Outros)	0,0004	3,25	0,071
Idade	-0,0057	18,30	0,000

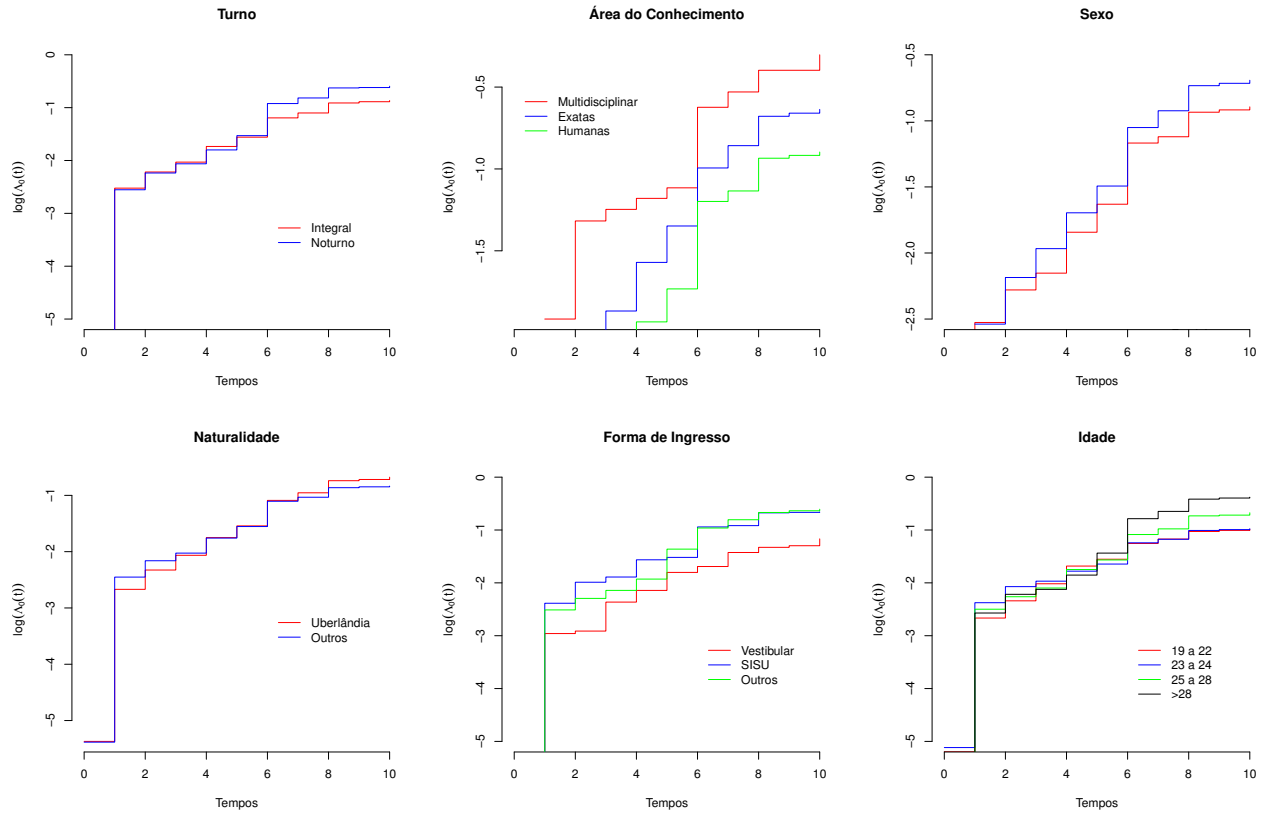


Figura 4.3: $\text{Log}(\hat{\Lambda}_{0j}(t))$ versus os tempos para as covariáveis do modelo de Cox.

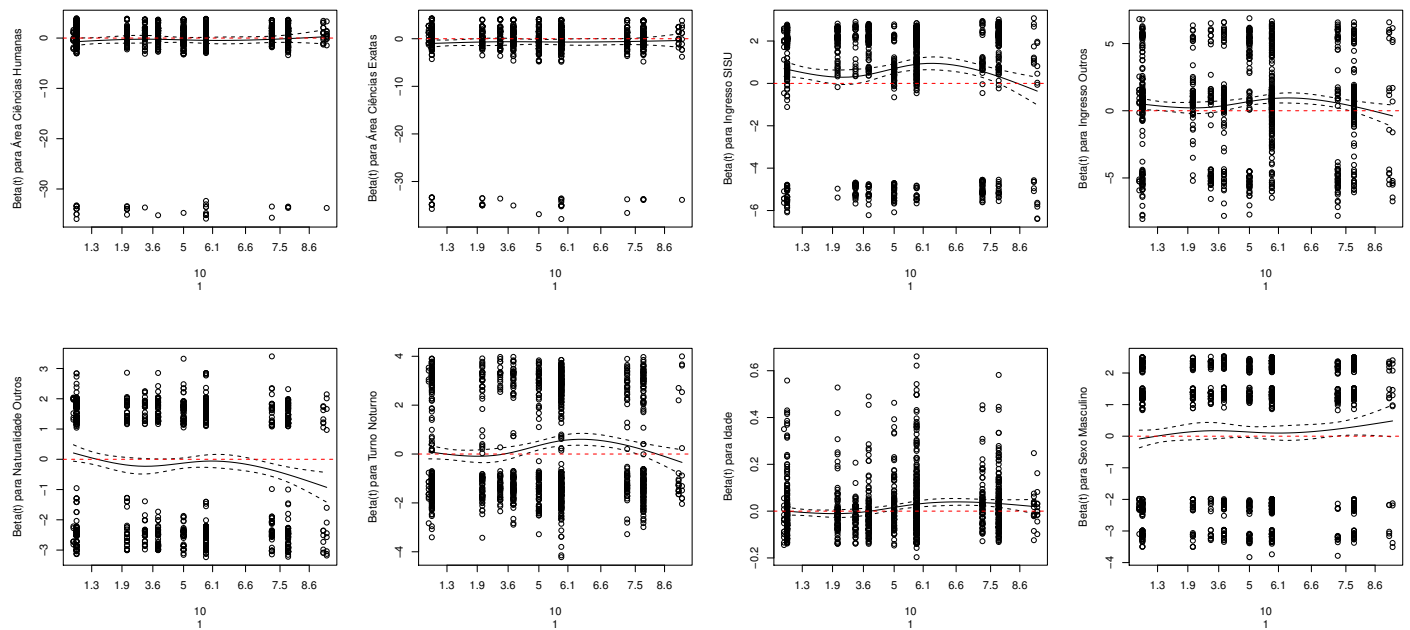


Figura 4.4: Resíduos padronizados de Schoenfeld versus os tempos para as covariáveis do modelo de Cox.

Para avaliar o ajuste global do modelo de Cox foram obtidos os resíduos de Cox-Snell e os gráficos correspondentes apresentados na Figura 4.5. A sobrevivência estimada para os resíduos de Cox-Snell por Kaplan-Meier se aproxima da sobrevivência estimada por uma distribuição exponencial padrão, havendo um afastamento entre as duas curvas a partir da metade do gráfico. No gráfico de $\hat{\Lambda}(\hat{e}_i)$ contra os resíduos de Cox-Snell $\hat{e}_i$ não se percebe grandes desvios de uma reta com inclinação 1. Logo, pode-se concluir que o modelo apresentou bom ajuste, apesar de ter apresentado covariáveis que não satisfazem a suposição de taxas de falha proporcionais. Na mesma figura, com o propósito de encontrar possíveis *outliers*, foram feitos os gráficos dos resíduos *deviance*. Estes gráficos sugerem que não existem pontos atípicos tendo em vista que a maior parte dos resíduos estão distribuídos entre -3 e 3.

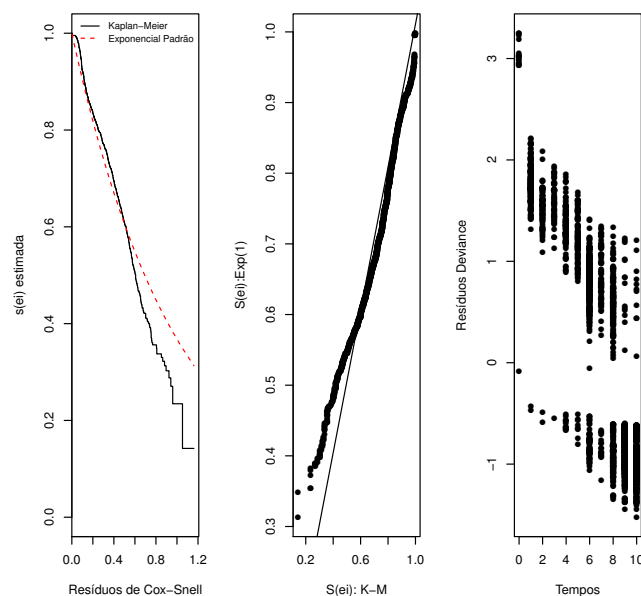


Figura 4.5: Análise dos resíduos de Cox-Snell e *deviance* do modelo de Cox.

Foram obtidas as estimativas dos parâmetros para o modelo de Cox, conforme a Tabela 4.4. É possível concluir a partir desses resultados que taxa de evasão para estudantes dos cursos noturnos é aproximadamente 25,4% maior do que os alunos dos cursos integrais. Os estudantes das áreas ciências humanas e ciências exatas apresentaram taxa de evasão 31,9% e 50,7% menor que a área do conhecimento multidisciplinar, nesta ordem. A taxa de evasão para estudantes naturais de outras cidades do Brasil é de aproximadamente 13,4% menor do que estudantes naturais da cidade de Uberlândia. Concluiu-se ainda que estudantes do sexo masculino possuem taxa de falha 13% maior do que estudantes do sexo feminino. Os resultados mostraram também que alunos que ingressam na universidade através do SISU possuem taxa 80,1% maior de abandonar o curso do que alunos ingressantes do vestibular, enquanto que os ingressantes de outras formas de ingresso possuem taxa de falha 71,3% maior do que alunos do vestibular. Por fim, a taxa de evasão aumenta cerca de 1,5% com o aumento de um ano na idade.

Tabela 4.4: Resultados do modelo de regressão de Cox ajustado para os dados do Campus Santa Mônica.

Covariável	coef	exp(coef)	p-valor
Turno			
Integral			
Noturno	0,2267	1,2544	0,0003
Área do Conhecimento			
Multidisciplinar			
Ciências Humanas	-0,3836	0,6813	0,0286
Ciências Exatas	-0,7082	0,4925	0,0000
Naturalidade			
Uberlândia			
Outros	-0,1438	0,8661	0,0153
Sexo			
Feminino			
Masculino	0,1219	1,1297	0,0456
Forma de Ingresso			
Vestibular			
SISU	0,5887	1,8017	0,0000
Outros	0,5385	1,7135	0,0000
Idade	0,0149	1,0150	0,0001

4.1.3 RESULTADOS DOS MODELOS ADITIVOS

Após a análise com o modelo de regressão de Cox, foi também ajustado o modelo aditivo de Aalen expresso em (2.14), já que houve violação da suposição de proporcionalidade das taxas de falha. O modelo de Aalen foi ajustado inicialmente com todas as covariáveis e foram retiradas do modelo aquelas que apresentaram efeito não significativo ($p > 0,05$). O modelo final permaneceu com as covariáveis da Tabela 4.5. Nesta Tabela, são apresentados o teste para o efeito não significativo das covariáveis, isto é, $H_{0a} : \beta_k(t) = 0$ e o teste para efeito constante das covariáveis, isto é, $H_{0b} : \beta_k(t) = \beta_k$, ou seja, efeito constante da variável sobre a taxa de falha, em todo o período de tempo analisado. De acordo com este último teste, para as covariáveis forma de ingresso, turno e idade rejeitou-se a hipótese nula de efeito tempo-independente (efeito constante), ou seja, o efeito dessas covariáveis sobre a taxa de falha depende do tempo.

Desse modo, ajustou-se o modelo de riscos aditivos semiparamétrico expresso na equação (2.15) que permite parte das covariáveis com efeito tempo-dependente e parte não. Nesse modelo a única covariável que apresentou efeito constante ao longo do tempo foi área do conhecimento, apresentando coeficientes estimados em -0,0301 e -0,0475 para as áreas ciências humanas e exatas respectivamente. A partir das estimativas desses coeficientes conclui-se que alunos das áreas do conhecimento ciências humanas e exatas apresentaram menor taxa de evasão quando comparado com a área multidisciplinar.

Tabela 4.5: Resultados do modelo aditivo de Aalen ajustado para os dados do Campus Santa Mônica.

Covariável	$H_{01}: \beta_q(t) = 0$		$H_{02}: \beta_q(t) = \beta_q$	
	Estatística	p-valor	Estatística	p-valor
Turno				
Integral				
Noturno	3,98	0,000	0,068	0,004
Área do Conhecimento				
Multidisciplinar				
Ciências Humanas	2,65	0,132	0,102	0,625
Ciências Exatas	3,37	0,014	0,105	0,561
Forma de Ingresso				
Vestibular				
SISU	5,68	0,000	0,085	0,003
Outros	3,54	0,008	0,070	0,173
Idade	3,05	0,036	0,007	0,000

Para as covariáveis com efeito tempo-dependente, têm-se os gráficos na Figura 4.6 evidenciando o efeito das covariáveis. O gráfico do intercepto corresponde à taxa de falha acumulada de um aluno que pertence às categorias de referência e com a covariável idade em seu valor médio. O gráfico da estimativa da função de regressão acumulada para a forma de ingresso SISU está acima do eixo no ponto zero, ou seja, a taxa de evasão para ingressantes pelo SISU é maior que a taxa de evasão para ingressantes pelo vestibular, que é a categoria de referência. Além disso, o gráfico apresentou leve aumento da taxa de falha até o 6^o período apresentando um acréscimo expressivo a partir deste, indicando que a partir do sexto período, o risco de falha aumenta consideravelmente. De outro modo, a categoria outras formas de ingresso apresentou uma elevação da taxa de falha no sexto período e se manteve aproximadamente constante a partir dele. Para a covariável idade a taxa possui inclinação levemente negativa até o 6^o período e após este há uma inclinação positiva, ou seja, até o sexto período, com o aumento da idade a taxa de ocorrência de evasão diminui, e a partir do sexto período, a taxa aumenta com o aumento da idade. Pode-se notar também que não houve diferença significativa na taxa de falha para estudantes dos cursos do turno noturno até o sexto período quando comparado aos alunos dos cursos integrais, entretanto, a partir desse, o risco de um aluno do noturno abandonar o curso se torna maior que o do curso integral.

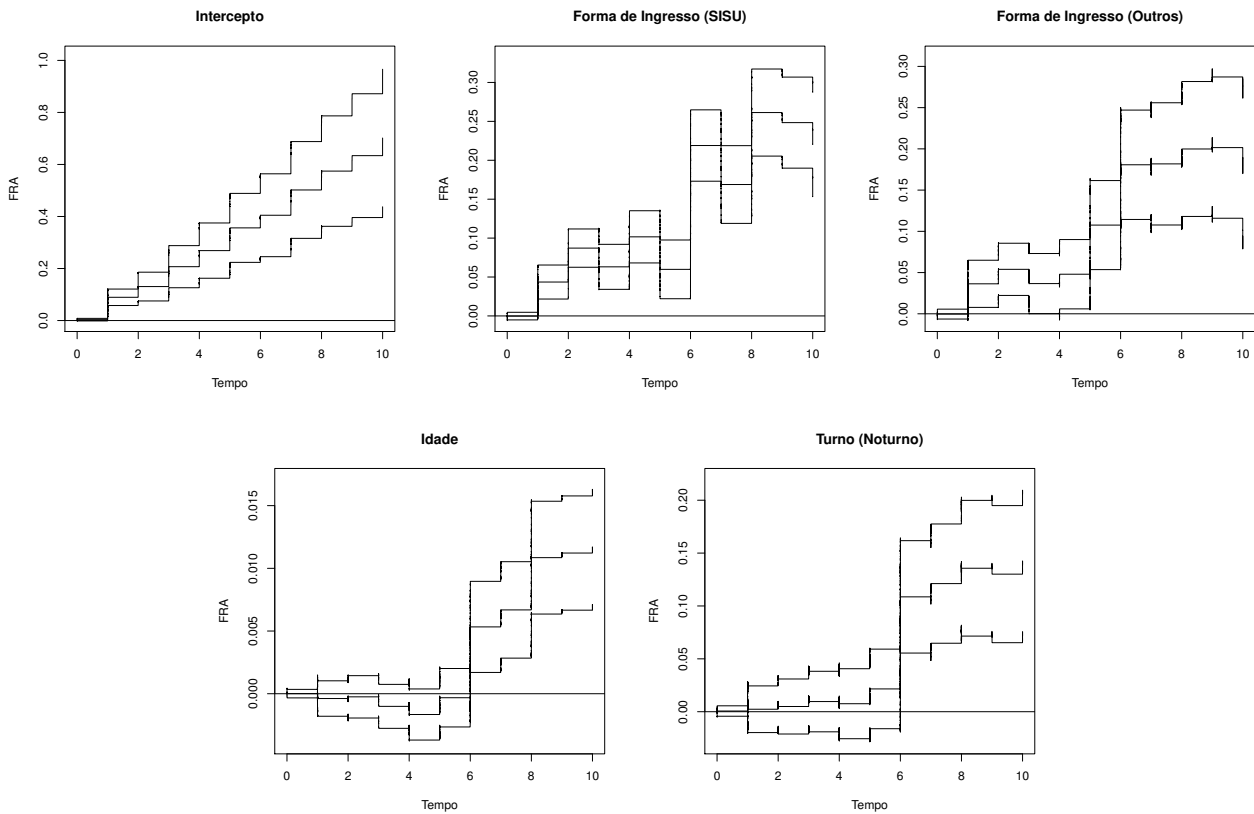


Figura 4.6: Estimativas das Funções de Regressão Acumulada (FRA) e seus respectivos intervalos de 95% de confiança associados ao modelo de aditivo semiparamétrico ajustado para os dados do Campus Santa Mônica.

4.2 CAMPUS UMUARAMA

4.2.1 ANÁLISE EXPLORATÓRIA DOS DADOS

A princípio foi realizada a análise descritiva do conjunto de dados referente ao Campus Umuarama. As frequências absolutas de alunos, de censuras e de falhas para as categorias de cada covariável bem como seus respectivos percentuais estão apresentados na Tabela 4.6. Neste caso, as covariáveis CRA e idade também foram categorizadas para facilitar a análise exploratória.

Considerando inicialmente a covariável turno, observa-se que o maior percentual está na categoria integral com 96% do número total de alunos. Com relação a área do conhecimento o maior número de estudantes pertencem a área de ciências da saúde com 52%, seguido da área de ciências agrárias com 31%. Quanto à naturalidade, 47% dos alunos deste campus são naturais da cidade de Uberlândia, enquanto os outros 53% são naturais de outras cidades do país. A forma de ingresso vestibular possui o maior número de ingressantes com 40%, seguido do SISU com 27%. A média do CRA dos alunos foi igual a 62,28 e apresentou desvio padrão igual a 26,89. Para a covariável idade, a média foi de 25,95 e desvio padrão de 4,44.

Quanto ao percentual de falha, que corresponde à evasão, o turno noturno apresentou o maior percentual. Quanto a área do conhecimento a categoria ciências da saúde obteve a menor

Tabela 4.6: Frequências absolutas e respectivos percentuais de alunos, de censuras e de falhas referente aos dados do Campus Umuarama.

Covariável	Categoria	N	N%	Censura	Censura%	Falha	Falha%
Turno	Integral	3721	96%	2755	74,0%	966	26,0%
	Noturno	172	04%	88	51,2%	84	48,8%
Área do Conhecimento	Ciências da Saúde	2018	52%	1548	76,7%	470	23,3%
	Ciências Biológicas	540	14%	350	64,8%	190	35,2%
	Ciências Agrárias	1193	31%	847	71,0%	346	29,0%
	Multidisciplinar	142	03%	98	69,0%	44	31,0%
Sexo	Feminino	2538	65%	1912	75,3%	626	24,7%
	Masculino	1355	35%	931	68,7%	424	31,3%
Naturalidade	Uberlândia	1818	47%	1393	76,6%	425	23,4%
	Outros	2075	53%	1450	69,9%	625	30,1%
Forma de Ingresso	SISU	1057	27%	625	59,1%	432	40,9%
	Vestibular	1567	40%	1225	79,2%	342	21,8%
	PAAES	804	21%	661	82,2%	143	17,8%
	Outros	465	12%	333	71,6%	132	28,4%
CRA	< 60	1248	32,1%	426	34,1%	822	65,9%
	≥ 60	2645	67,9%	2417	91,4%	228	8,6%
Idade	20 a 24	1736	44,6%	1345	77,5%	391	22,5%
	25	818	21,0%	603	73,7%	215	26,3%
	26	366	9,4%	261	71,3%	105	28,7%
	> 27	973	25,0%	634	65,2%	339	34,8%

taxa de falha. Para o sexo a maior taxa de falha foi observada para a categoria masculino. Dentre as formas de ingresso, o SISU obteve a maior taxa de falha.

A curva de sobrevivência foi obtida por meio do estimador de Kaplan-Meier, como também os intervalos de 95% de confiança, para $0 \leq t \leq 16$, conforme a Figura 4.7. A sobrevivência final foi de 0,678, ou seja, a probabilidade de sobreviver ao tempo 16 semestres sem desistir do curso é de 67,8%. O tempo médio de permanência dos estudantes até a evasão foi de 12,9 semestres.

Também foram obtidas as curvas de sobrevivência para cada covariável na Figura 4.8 e o teste de *logrank* na Tabela 4.7 a fim de identificar diferenças significativas entre as curvas para as categorias de cada covariável. De acordo com o teste todas as covariáveis apresentaram diferença significativa, dessa maneira o resultado sugere que todas as covariáveis podem estar influenciando no tempo de desistência dos alunos.

Tabela 4.7: Teste de *logrank*.

Covariável	Qui-quadrado	p-valor
Turno	37,3	0
Área do Conhecimento	78	0
Sexo	16,9	0,00004
Naturalidade	29,8	0
Forma de Ingresso	206	0

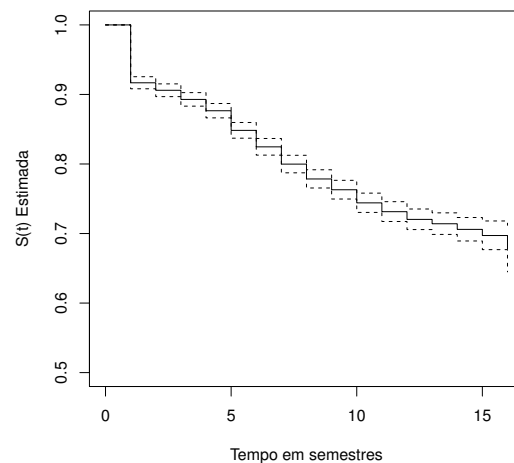


Figura 4.7: Curva de sobrevivência estimada a partir do método de Kaplan-Meier para os dados do Campus Umuarama.

4.2.2 RESULTADOS DO MODELO DE REGRESSÃO DE COX

Após a análise exploratória dos dados, ajustou-se o modelo de regressão de Cox expresso em (2.6). Para a seleção das covariáveis ajustou-se um modelo com todas as covariáveis, eliminando-se uma por vez, aquelas não significativas ao nível 5% de significância. Apesar de não ser considerada no modelo, a covariável CRA possui forte efeito na probabilidade de sobrevivência do aluno no curso, como visto na curva de sobrevivência para essa covariável (Figura 4.8). O modelo final permaneceu com as covariáveis idade, naturalidade, área do conhecimento, forma de ingresso, turno e sexo.

A suposição de taxas de falha proporcionais foi verificada pelo coeficiente de correlação de Pearson, conforme a Tabela 4.8, onde observou-se que apenas a categoria ciências agrárias da covariável área do conhecimento e a categoria outros da covariável forma de ingresso não violaram a suposição de taxas de falhas proporcionais. Contudo, para amostras grandes como dito anteriormente, os p-valores muito pequenos podem ser obtidos em consequência do tamanho amostral.

Observando os gráficos dos resíduos de Schoenfeld na Figura 4.10 pode ser observado que existem tendências ao longo do tempo que possam sugerir violação dessa suposição para as categorias vestibular e PAAES da covariável forma de ingresso. Os gráficos do logaritmo da função de taxa de falha acumulada ($\log(\hat{\Lambda}_{0j}(t))$) contra os tempos na Figura 4.9 sugerem violação para a covariável turno visto que suas curvas se cruzam. Para as demais covariáveis não há indicativos de violação visto que não há afastamentos marcantes entre as curvas.

A partir das estimativas dos coeficientes do modelo de Cox apresentadas na Figura 4.9 conclui-se que a taxa de evasão para alunos dos cursos noturnos é de aproximadamente 63,9% maior do que para os alunos dos cursos integrais. Os estudantes das áreas ciências biológicas e agrárias apresentaram taxa de evasão de 37,6% e 37,5%, respectivamente, maior do que os alunos da área ciências da saúde, enquanto que a área multidisciplinar apresentou taxa de evasão

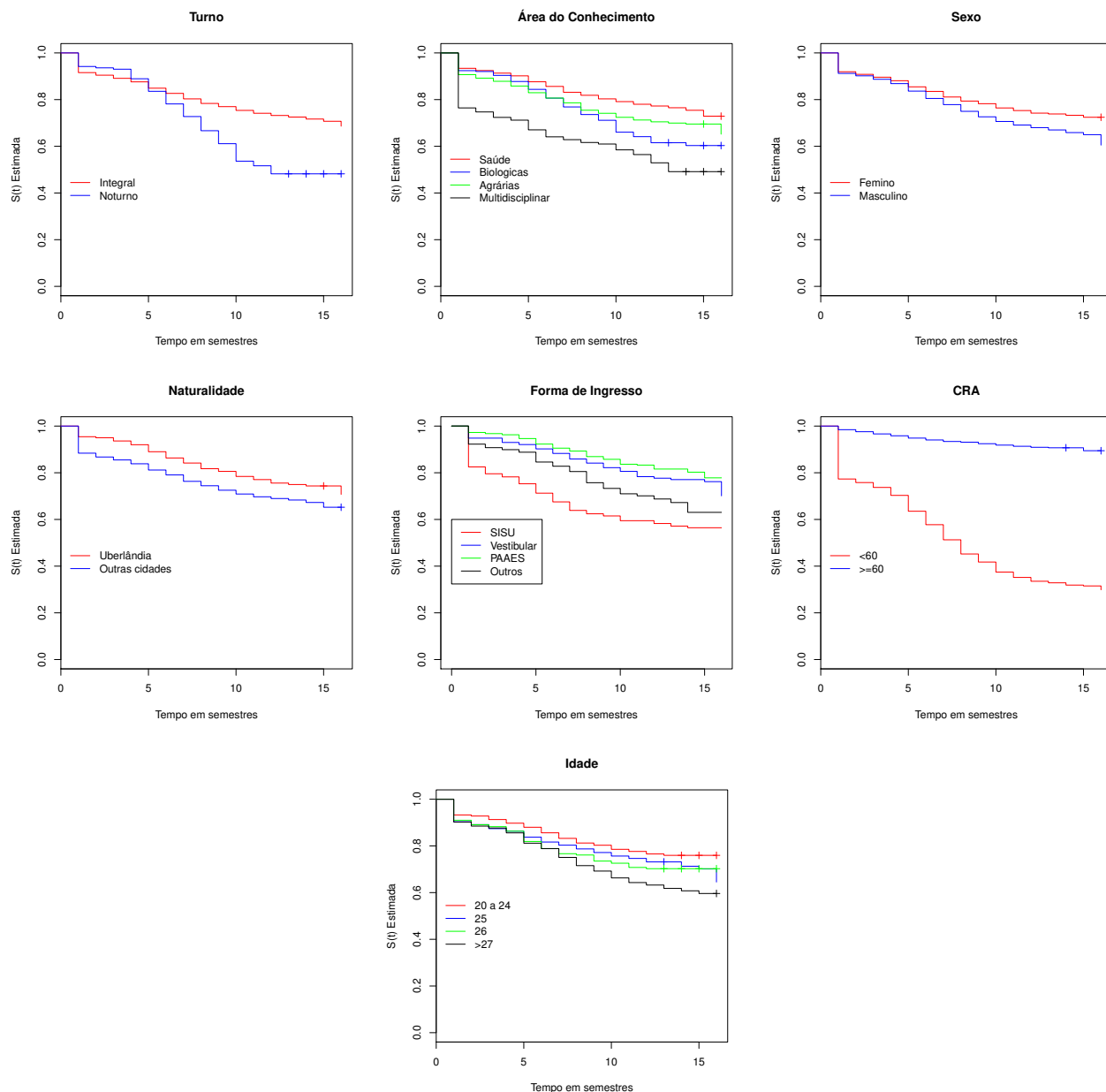


Figura 4.8: Curvas de sobrevivência estimadas a partir do método de Kaplan-Meier para cada covariável.

Tabela 4.8: Teste da proporcionalidade das taxas de falha no modelo de Cox ajustado para os dados do Campus Umuarama.

Covariável	ρ	χ^2	p-valor
Turno (Noturno)	0,0092	30,80	0,0000
Área do Conhecimento (C. Biológicas)	0,0062	21,07	0,0000
Área do Conhecimento (C. Agrárias)	0,0292	2,70	0,1006
Área do Conhecimento (Multidisciplinar)	-0,0394	7,69	0,0055
Sexo (Masculino)	-0,0459	5,09	0,0241
Naturalidade (Outros)	0,0216	27,52	0,0000
Forma de Ingresso (Vestibular)	-0,0487	11,82	0,0006
Forma de Ingresso (PAAES)	-0,0254	11,86	0,0006
Forma de Ingresso (Outros)	-0,0166	3,12	0,0772
Idade	-0,0068	12,67	0,0004

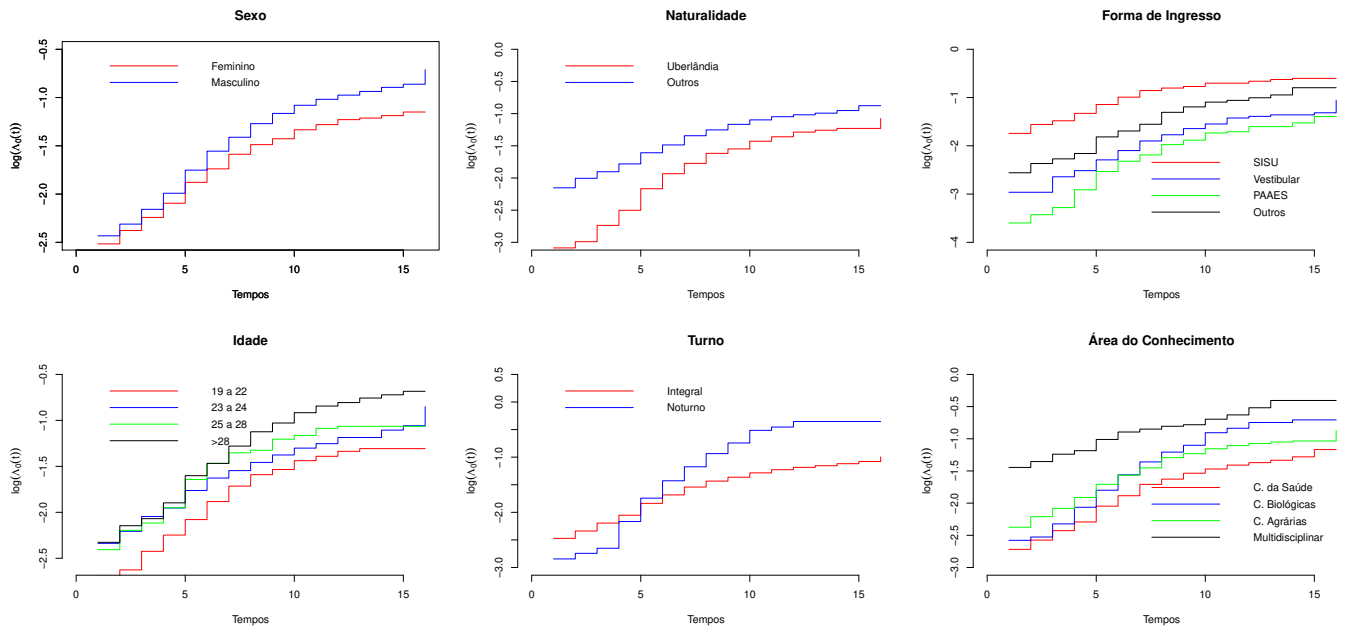


Figura 4.9: $\text{Log}(\hat{\Lambda}_{0j}(t))$ versus os tempos para as covariáveis do modelo de Cox.

2,46 vezes a taxa da área ciências da saúde. Os estudantes naturais de outras cidades possuem taxa 2,19% maior quando comparado aos estudantes naturais de Uberlândia. Em relação ao sexo, alunos possuem taxa de evasão 15,2% maior do que alunas. A taxa de evasão aumenta 3,21% quando aumenta um ano na idade. Por fim, os ingressantes do vestibular apresentaram taxa 54% menor do que os ingressantes do SISU, para os para os ingressantes do PAAES essa taxa foi de 59% e para ingressantes de outras formas de ingresso essa taxa foi de 43,3%.

A qualidade do ajuste global do modelo foi verificada pelos gráficos dos resíduos de Cox-Snell, exibidos na Figura 4.11. Os gráficos mostram um ótimo ajuste, já que os resíduos de Cox-Snell se aproximam bastante de uma distribuição exponencial padrão e o gráfico de $\hat{\Lambda}(\hat{e}_i)$ contra $\hat{e}_i$ se aproxima de uma reta. Verificou-se também a ausência de observações atípicas no gráfico dos resíduos *deviance*.

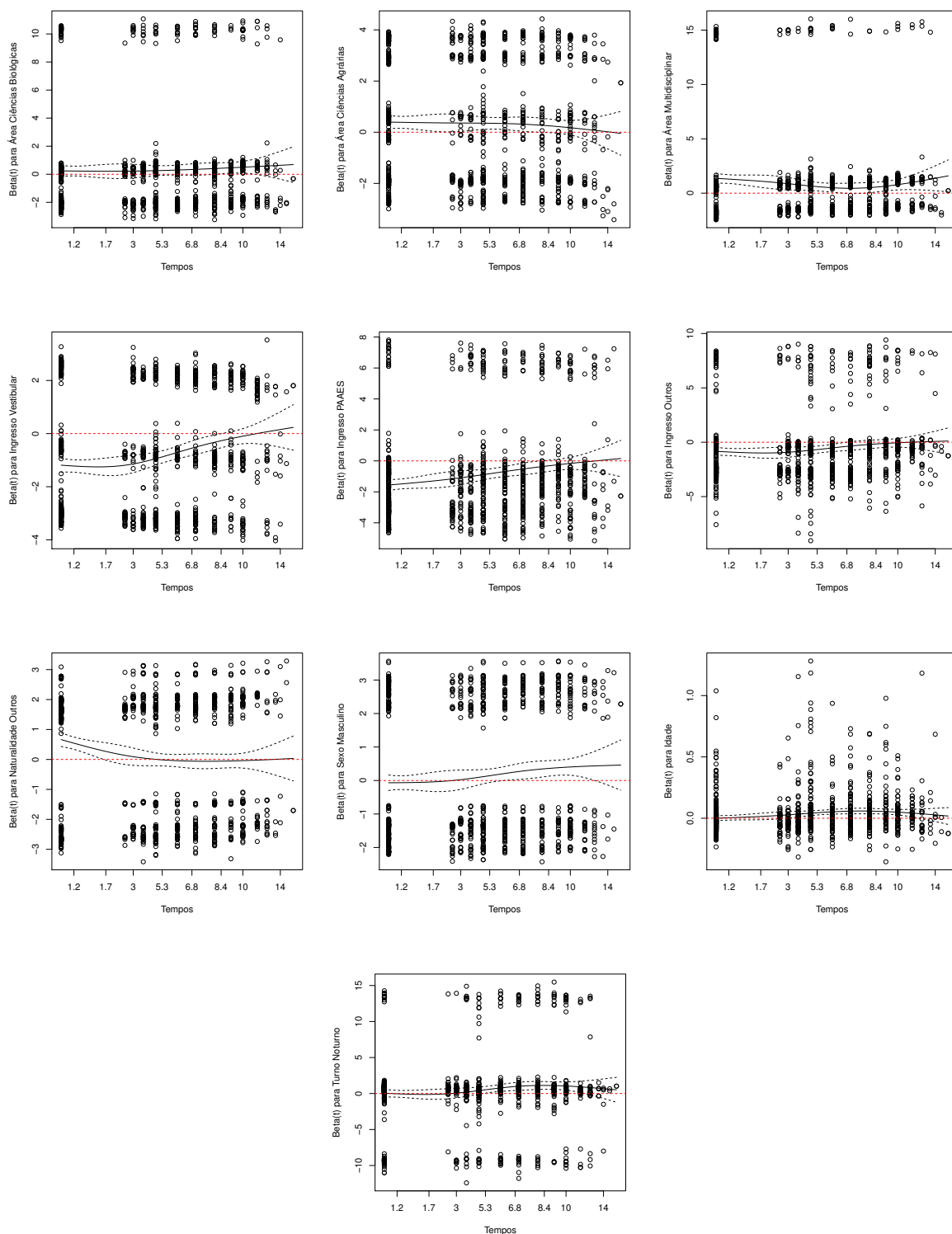


Figura 4.10: Resíduos padronizados de Schoenfeld *versus* os tempos para as covariáveis do modelo de Cox.

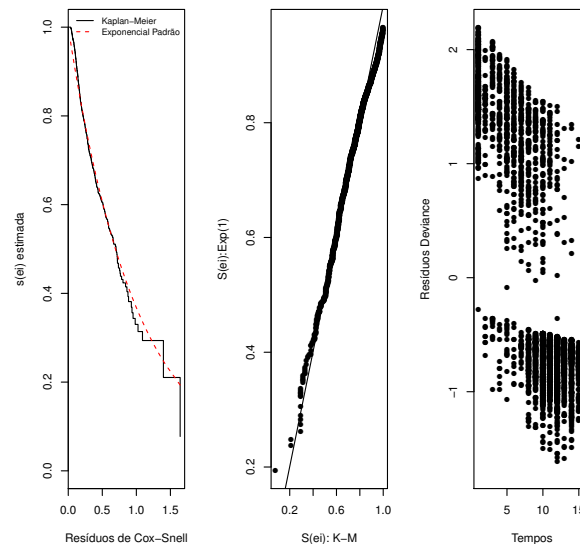


Figura 4.11: Análise dos resíduos de Cox-Snell e *deviance* do modelo de Cox.

Tabela 4.9: Resultados do modelo de regressão de Cox ajustado para os dados do Campus Umuarama.

Covariável	coef	exp(coef)	p-valor
Turno			
Integral			
Noturno	0,4939	1,6387	0,0009
Área do Conhecimento			
Ciências da Saúde			
Ciências Biológicas	0,3191	1,3759	0,0034
Ciências Agrárias	0,3186	1,3752	0,0000
Multidisciplinar	0,8985	2,4560	0,0000
Naturalidade			
Uberlândia			
Outros	0,1978	1,2187	0,0021
Sexo			
Feminino			
Masculino	0,1422	1,1528	0,0264
Forma de Ingresso			
SISU			
Vestibular	-0,7773	0,4596	0,0000
PAAES	-0,8922	0,4097	0,0000
Outros	-0,5677	0,5668	0,0000
Idade	0,0315	1,0321	0,0000

4.2.3 RESULTADOS DOS MODELOS ADITIVOS

O modelo aditivo de Aalen expresso na equação (2.14) foi ajustado inicialmente com todas as covariáveis e considerando um nível 5% de significância para a retirada de covariáveis. O modelo final permaneceu com as covariáveis forma de ingresso, área do conhecimento, idade e turno. Na Tabela 4.10 são apresentados os testes para efeito não significativo das covariáveis ($H_{01}: \beta_q(t)$)

$= 0$) e efeito constante das covariáveis ($H_{02}: \beta_q(t) = \beta_q$). A partir desses resultados é possível concluir que as covariáveis área do conhecimento e forma de ingresso possuem efeito tempo-dependente. Dessa forma, ajustou-se o modelo de riscos aditivos semiparamétrico expresso em (2.15), que permite parte das covariáveis com efeito tempo-dependente e parte não.

Tabela 4.10: Resultados do modelo aditivo de Aalen ajustado para os dados do Campus Umuarama.

Covariável	$H_{01}: \beta_q(t) = 0$		$H_{02}: \beta_q(t) = \beta_q$	
	Estatística	p-valor	Estatística	p-valor
Turno				
Integral				
Noturno	3,87	0,003	0,129	0,182
Área do Conhecimento				
Ciências da Saúde				
Ciências Biológicas	3,05	0,049	0,110	0,224
Ciências Agrárias	4,71	0,000	0,085	0,113
Multidisciplinar	5,09	0,000	0,233	0,046
Naturalidade				
Uberlândia				
Outros	6,07	0,000	0,060	0,248
Forma de Ingresso				
SISU				
Vestibular	8,78	0,000	0,179	0,001
PAAES	9,59	0,000	0,180	0,004
Outros	7,33	0,000	0,204	0,031
Idade	4,52	0,000	0,007	0,221

Os parâmetros das covariáveis que possuem efeito constante foram estimados, conforme apresentado na Tabela 4.11, e evidenciam que alunos dos cursos noturnos possuem maior taxa de evasão do que alunos dos cursos noturnos. Além disso, estudantes naturais de outras cidades do país apresentaram taxa de falha maior do que alunos naturais de Uberlândia. A taxa de falha cresce com o aumento da idade do aluno.

Tabela 4.11: Estimativa dos coeficientes com efeito constante no modelo de riscos aditivos semiparamétrico ajustado para os dados do Campus Umuarama.

Covariável	Coeficiente	Erro padrão	Z	p-valor
Turno				
Integral				
Noturno	0,019	0,0060	3,29	0,001
Naturalidade				
Uberlândia				
Outros	0,006	0,0020	3,05	0,002
Idade	0,001	0,0003	4,35	0,000

Para as covariáveis com efeito tempo-dependente tem-se os gráficos em 4.12 das funções de regressão acumulada evidenciando que o efeito das covariáveis mudam com o tempo. A função de regressão acumulada no gráfico do intercepto corresponde à taxa de falha acumulada de um

aluno que pertence às categorias de referência e com a covariável idade em seu valor médio. O gráfico da função de regressão acumulada para a forma de ingresso vestibular está abaixo do eixo no ponto zero, ou seja, a taxa de evasão para a forma de ingresso vestibular é menor que a taxa de evasão para a forma de ingresso SISU. Além disso, a forma de ingresso vestibular apresentou inclinação negativa até o décimo período, ou seja, a taxa de falha dos ingressantes pelo vestibular decresce nesse período, à medida que o tempo passa. Do mesmo modo, na forma de ingresso PAAES observou-se um decrescimento da taxa de evasão em todo o período em relação ao SISU. Para a área do conhecimento ciências biológicas a função de regressão acumulada está acima do eixo zero, ou seja, a taxa de evasão da área ciências biológicas é maior que a taxa de evasão da área ciências da saúde. Além disso, percebe-se crescimento na taxa de evasão ao longo do tempo para estudantes da área de ciências biológicas. O mesmo ocorre para as áreas ciências agrárias e multidisciplinar, ou seja, a taxa de evasão aumenta ao longo do tempo, além disso, é maior quando comparada com a área de ciências da saúde.

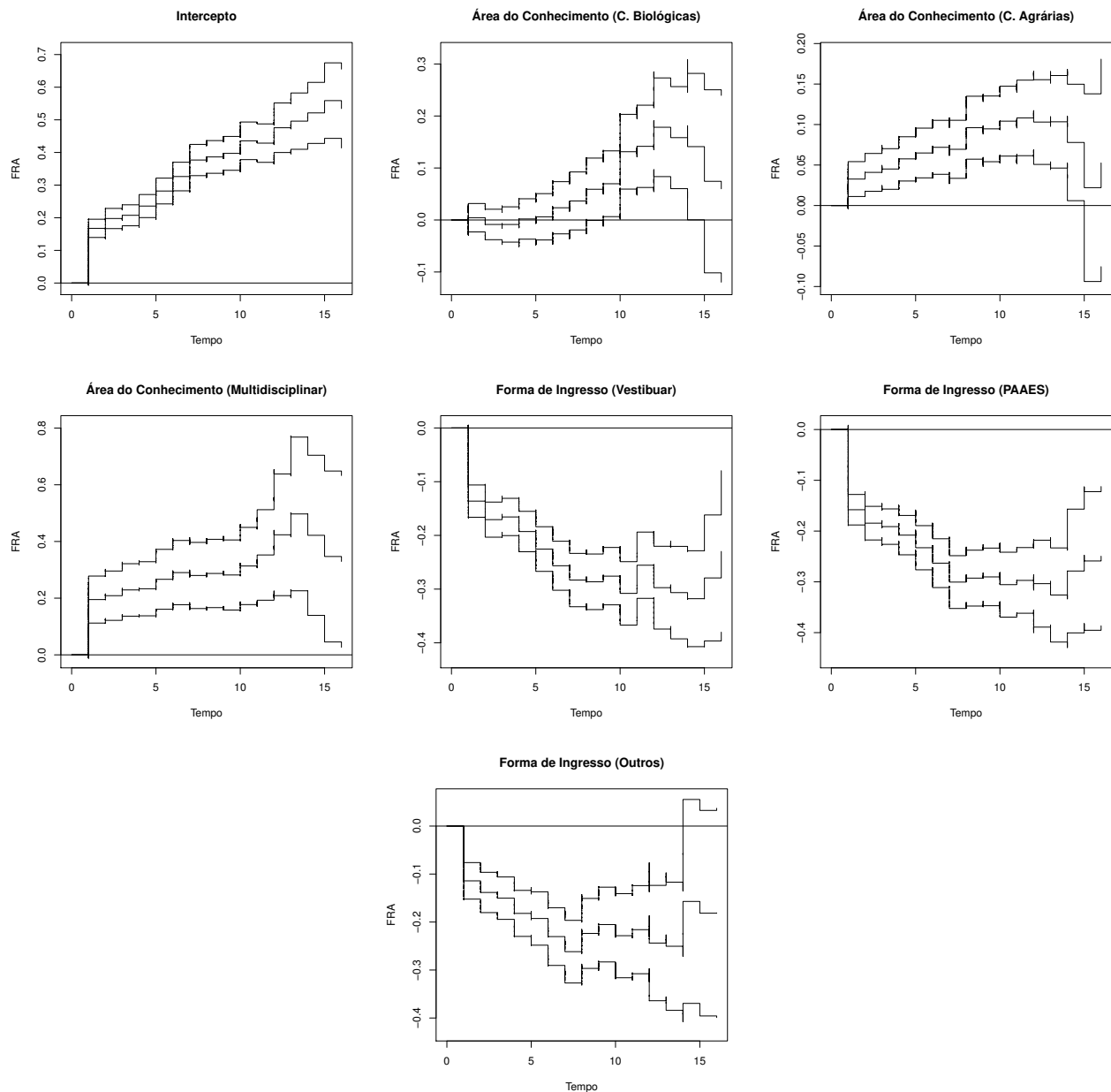


Figura 4.12: Estimativas das Funções de Regressão Acumulada (FRA) e seus respectivos intervalos de 95% de confiança associados ao modelo de aditivo semiparamétrico ajustado para os dados do Campus Umuarama.

5. CONSIDERAÇÕES FINAIS

A evasão universitária é um fenômeno que atinge as instituições de ensino superior (IES), tanto públicas quanto privadas, em todo o mundo. O abandono dos estudos por um estudante gera prejuízos sociais, acadêmicos e econômicos. Em razão disso, estudos sobre o tempo até a evasão do aluno e os fatores associados são fundamentais. O modelo de regressão de Cox é bastante utilizado em estudos de sobrevivência, entretanto, alternativas como os modelos de riscos aditivos são utilizados quando há presença de covariáveis tempo-dependentes. Neste trabalho, esses modelos foram aplicados aos dados referentes aos ingressantes nos cursos de graduação do Campus Santa Mônica e Campus Umuarama da Universidade Federal de Uberlândia.

A partir dos resultados do modelo de Cox ajustado aos dados do campus Santa Mônica concluiu-se que a taxa de evasão dos estudantes que ingressam na universidade por meio do SISU e de outras formas de ingresso é maior do que para os ingressantes do vestibular. Com resultado semelhante, o modelo de riscos aditivos semiparamétrico mostrou que a taxa de evasão desses ingressantes apresentou maior aumento a partir do sexto período. Observou-se também no modelo de Cox que estudantes do sexo masculino possuem maior risco de evadir do que estudantes do sexo feminino, no entanto, no modelo de riscos aditivos essa covariável não apresentou efeito significativo. De acordo com o modelo de Cox, a taxa de evasão aumenta com o aumento da idade e, segundo o modelo de riscos aditivos, esse aumento é mais expressivo após o sexto período. Quanto ao turno, no modelo de Cox, os alunos do período noturno apresentaram maior risco de abandonar o curso comparado aos do período integral e, conforme o modelo de riscos aditivos, após o sexto período o risco é mais significativo. Os resultados mostraram ainda que nos dois modelos os alunos das áreas ciências humanas e exatas possuem menor risco de evadir do curso comparado à área multidisciplinar. Por fim, considerando os cursos e os dados do Campus Santa Mônica, concluiu-se por meio do modelo de Cox que alunos de outras cidades possuem menor risco de evadir do que pessoas naturais de Uberlândia, não apresentando efeito significativo no modelo semiparamétrico.

Os resultados do modelo de Cox ajustado aos dados do campus Umuarama mostram que alunos que ingressam na universidade por meio do vestibular, PAAES e outras formas de ingresso apresentaram risco de evasão menor quando comparado à forma de ingresso SISU. No modelo de riscos aditivos semiparamétrico, para o vestibular, até o tempo dez semestres o risco é decrescente, enquanto que para o PAAES e outras formas de ingresso as taxas de falha o risco é decrescente até o sétimo período. Concluiu-se também que com o aumento da idade em anos a taxa de evasão aumenta em ambos os modelos. Também nos dois modelos os alunos

que pertencem aos cursos noturnos apresentaram maior risco de abandonar o curso do que os alunos dos cursos integrais. Segundo o modelo de Cox, alunos das áreas do conhecimento ciências biológicas, ciências agrárias e multidisciplinar possuem taxa de evasão maior do que a área ciências da saúde e, segundo o modelo aditivo semiparamétrico, essa taxa é maior até o tempo 14 aproximadamente. Quanto aos alunos naturais de Uberlândia, a taxa de evasão se mostrou menor que a dos alunos naturais de outras cidades do país nos dois modelos. Por fim, os resultados mostram que no modelo de Cox os alunos do sexo masculino possuem maior risco de evadir do que estudantes do sexo feminino, porém, no modelo de riscos aditivos não apresentou efeito significativo.

Comparando-se os modelos ajustados, no caso dos dados do Santa Monica, houve concordância nas conclusões do modelo de Cox com os modelos aditivos. Apesar de o teste indicar violação da suposição de taxas de falhas proporcionais, as verificações gráficas permitem supor que não houve violação da suposição, e que o ajuste foi satisfatório.

Considerando os dados do Umuarama, também houve razoável concordância entre o Modelo de Cox e os modelos aditivos. Neste caso, algumas verificações gráficas também sinalizam taxas de falhas não proporcionais, embora o ajuste tenha sido razoavelmente satisfatório, de acordo com os gráficos.

Fazendo um comparativo entre os dados dos dois Campi, verificou-se que para as covariáveis que apresentam as mesmas categorias, nos dois locais, os resultados em geral foram concordantes, como por exemplo, maior taxa de evasão para ingressantes do SISU comparado ao vestibular, maior taxa de desistência para o sexo masculino, aumento na taxa de evasão com aumento da idade, diminuição na taxa de evasão com o aumento do CRA. A covariável área do conhecimento não apresenta as mesmas categorias para os dois locais, por esse motivo não pode ser comparada. A covariável naturalidade apresentou resultado divergente nos dois Campi.

REFERÊNCIAS BIBLIOGRÁFICAS

- [1] CAMPOS, J.D.D.S.: *Fatores Explicativos para a Evasão no Ensino Superior através da Análise de Sobrevivência: O caso da UFPE*. Dissertação de Mestrado, Universidade Federal de Pernambuco, 2016.
- [2] COLOSIMO, E. A. e GIOLO, S.R.: *Análise de Sobrevivência Aplicada*. Edgard Blucher, 2006.
- [3] COSTA, F.P., SIQUEIRA, K.J.D.S.P.D. e ROCHA, E.C.D.: *Sobrevivendo ao Ensino Superior: uma análise da evasão nas universidades federais brasileiras*. <https://coreconpe.gov.br/enpecon/vii/enpecon/artigos/sessao1/>, acesso em 16/07/2020.
- [4] FILHO, R.L.L.E.S.: *A Evasão no Ensino Superior Brasileiro - Novos Dados*. <https://www.institutolobo.org.br/artigos/?pag=3>, acesso em 20/07/2020.
- [5] FILHO, R.L.L.E.S., MOTEJUNAS, P.R., HIPÓLITO, O. e LOBO, M.B.D.C.M.: *A Evasão no Ensino Superior Brasileiro*. Instituto Lobo para o Desenvolvimento da Educação, da Ciência e da Tecnologia, 37(132):641–659, 2007.
- [6] LOBO, M.B.D.C.M.: *Panorama da Evasão no Ensino Superior Brasileiro: Aspectos Gerais das Causas e Soluções*. Em *Evasão no Ensino Superior*. ABMES Cadernos, 2012.
- [7] MEC/SESU: *Comissão Especial de Estudos sobre a Evasão nas Universidades Públicas Brasileiras*. 1996.
- [8] PEREIRA, T.L.: *Modelos de Riscos Proporcionais e Aditivos para o Tratamento de Covariáveis Dependentes do Tempo*. Tese de Mestrado, Universidade Federal de Pernambuco, 2004.
- [9] QUEIROZ, C.S.: *Um Estudo sobre o Tipo de Evasão no Centro de Ciências Exatas e Tecnológicas/UFRB: Uma Aplicação na Análise de Sobrevivência*. Trabalho de Conclusão de Curso, Universidade Federal do Recôncavo da Bahia, 2018.
- [10] R Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. <https://www.R-project.org/>.
- [11] RAMINELLI, J.A.: *Métodos de Adequação e Diagnóstico em Modelos de Sobrevivência Dinâmicos*. Tese de Doutorado, Escola Superior de Agricultura "Luiz de Queiroz", 2015.

- [12] SANTOS, A.F.D.C. e WERNER, F.: *Análise Estatística de Sobrevivência: um estudo de pacientes com câncer de mama tratados no município de Curitiba*. Trabalho de Conclusão de Curso, Universidade Federal do Paraná, 2016.
- [13] TIGRINHO, L.M.V.: *Evasão Escolar nas Instituições de Ensino Superior*. <http://www.gestaouniversitaria.com.br/artigos/evasao-escolar-nas-instituicoes-de-ensino-superior>, acesso em 21/07/2020.