Universidade de Brasília
Departamento de Estatística

Análise do tempo até a aquisição de um vínculo empregatício dos doutores titulados no Brasil via regressão de Cox com covariáveis dependentes no tempo

Stephany Lima de Oliveira

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

Brasília
2023

Stephany Lima de Oliveira

Análise do tempo até a aquisição de um vínculo empregatício dos doutores titulados no Brasil via regressão de Cox com covariáveis dependentes no tempo

Orientador: Prof. Eduardo Yoshio Nakano

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

Brasília
2023

Agradecimentos

Primeiramente, sou grata a Deus por permitir que meus sonhos se tornem realidade e por me fortalecer em todos os momentos desafiadores.

Agradeço à minha família, em especial à minha mãe, Alexandra, ao meu pai, Silvandir, e ao meu irmão, Maxwell, por todo amor, apoio constante e incentivo diário.

Aos amigos que estiveram comigo neste ciclo, Júlia Birbeire, Luana Feijão, Marcelo Batalha e Sabrina França, que se fizeram presentes em todos os momentos.

Ao meu namorado, Ítalo Luís, que me faz ser uma pessoa melhor todos os dias. Seu amor, sua paciência e seu companheirismo me potencializam.

Ao Centro de Gestão e Estudos Estratégicos e aos colaboradores, em especial à Rayany Santos, que tornaram possível a realização deste trabalho. Obrigada pelo apoio e confiança.

Ao meu orientador, Professor Eduardo Nakano, expresso minha profunda gratidão por sua orientação, pelos conhecimentos compartilhados e por sua disponibilidade.

Por fim, agradeço também a todos os professores que estiveram presentes ao longo da minha graduação, pelos quais tenho muita admiração.

Resumo

Este trabalho objetiva analisar o tempo até a aquisição de um vínculo empregatício formal de caráter acadêmico de doutores que titularam no Brasil por meio do modelo de regressão de Cox com covariáveis dependentes no tempo. Para a obtenção das informações de titulação e emprego foram utilizadas as bases de dados da Relação Anual de Informações Sociais (RAIS) dos anos 2012 a 2021, Plataforma Sucupira dos anos 2013 a 2021, Cadastro Geral de Empregados e Desempregados (CAGED) dos anos 2013 a 2021. Os resultados obtidos mostraram que não ter vínculo ativo fora da área acadêmica, ter obtido o título de doutor até os 32 anos, ter titulado em programa nota 5 da Capes e ter titulado na grande área de Linguística foram fatores que indicaram um menor tempo até a aquisição de um vínculo formal na área acadêmica.

Palavras-chaves: Análise de sobrevivência; Modelo de regressão; Riscos Proporcionais de Cox; Covariáveis Dependentes no Tempo; Doutores; Plataforma Sucupira; RAIS; CAGED.

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

A pós-graduação, em particular a formação de doutores, empenha-se na construção de um projeto de desenvolvimento sustentável para a produção e transformação de conhecimento, bem como na geração de inovação. Dessa forma, assume um papel estratégico e relevante no Sistema Nacional de Ciência, Tecnologia e Inovação (SNCTI), além de exercer forte influência nos processos de aumento de produtividade e na qualidade de vida dos cidadãos (CGEE, 2015).

Devido a esse papel desempenhado nos processos de produção e transmissão de conhecimento e tecnologia, é necessário estudar e acompanhar as peculiaridades e tendências desse seleto grupo composto por doutores.

Nesse contexto, o objetivo deste trabalho é analisar, por meio do modelo de riscos proporcionais de Cox com covariáveis dependentes no tempo, o tempo de duração desde a obtenção do título de doutorado até a aquisição de um vínculo empregatício formal no campo acadêmico, especificamente quando a atividade econômica principal do vínculo é Educação ou Atividades Profissionais, Científicas e Técnicas. Isso se deve ao fato de que, segundo CGEE (2019), é nessas atividades relacionadas à docência e pesquisa que se encontra a maioria dos doutores que possuem vínculo empregatício ativo. Além disso, busca-se identificar quais fatores podem influenciar o tempo de sobrevivência.

A análise de sobrevivência é um ramo da estatística que tem ganhado relevância nos últimos anos. Seu foco está na análise do tempo até a ocorrência de um evento específico, denominado tempo de falha. O modelo de regressão de Cox é uma importante técnica nessa área, uma vez que possibilita a análise de dados originários de tempo de vida com a presença de covariáveis. Uma das principais vantagens do modelo de Cox é sua flexibilidade e capacidade de incorporar o efeito de covariáveis dependentes do tempo (PARREIRA, 2007).

Para a obtenção dos conjuntos de dados, foi realizada uma solicitação de acesso ao Centro de Gestão e Estudos Estratégicos (CGEE), com o objetivo de obter as bases da Plataforma Sucupira, gerenciada pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes), bem como a Relação Anual de Informações Sociais (RAIS) e o Cadastro Geral de Empregados e Desempregados (CAGED), que são gerenciados pelo Ministério do Trabalho e Emprego (MTE).

2 Metodologia

2.1 Análise de sobrevivência

A Análise de sobrevivência é uma vertente da estatística que estuda o tempo, desde um instante inicial, até a ocorrência de um determinado evento de interesse. Este evento de interesse denota o tempo de falha e pode ser ampliado a diversas áreas para a sua aplicação, como por exemplo o tempo até a falha de um equipamento, o tempo que um indivíduo permanece desempregado ou empregado ou até mesmo o número de sessões de um tratamento até a cura (NAKANO, 2017).

2.1.1 Censura

A censura é uma das principais características de um conjunto de dados de sobrevivência e acontece quando não é possível observar a falha do indivíduo até a finalização do estudo. Ademais, suas causas podem estar relacionadas ao termino ou abandono do estudo, falha devida a outra causa ou número de falhas determinadas alcançadas.

Embora a presença de censura indique que a observação está incompleta, a sua inclusão é importante para a construção da análise, visto que, dados censurados apresentam informações sobre o tempo de vida e a sua omissão pode ocasionar conclusões viesadas. Com isso, a variável indicadora de censura é a adicionada, se o tempo de sobrevida é observado a variável é igual a um e caso o tempo de sobrevida seja censurado igual à zero.

$$\delta_i = \begin{cases} 0, \text{se o } i\text{-ésimo tempo foi censurado.} \\ 1, \text{se o } i\text{-ésimo tempo foi observado.} \end{cases}$$

em que $i = 1, 2, ..., n$.

Os tipos de censura são classificados como: (1) censura à esquerda, que tem por característica a falha ocorrida antes do tempo observado; (2) censura intervalar, é caracterizada pelo evento de interesse ocorrido dentro de um intervalo de tempo, contudo não é conhecido o tempo exato da falha; e, por fim, (3) censura à direita, que acontece quando o tempo da falha não é observado no tempo registrado.

Quanto a censura à direita é possível categoriza-la em três tipos sendo:

1. Censura do Tipo I: É caracterizada pela falha ocorrida após tempo de duração pré-determinado.

2. Censura do Tipo II: Ocorre quando o número de falhas é pré-estabelecido.

3. Censura Aleatória: Acontece quando o elemento não consegue ser acompanhado até o tempo registrado ou falhou por alguma razão diferente do fenômeno estudado.

2.1.2 Representação do tempo de sobrevivência

O tempo de sobrevivência de determinada observação configura uma variável aleatória não-negativa, $T \geq 0$, geralmente contínua que pode ser caracterizada a partir da função de densidade de probabilidades, $f(t)$, função de sobrevivência, $S(t)$, função de risco, $h(t)$ e função risco acumulado, $H(t)$ (NAKANO, 2017).

2.1.2.1 Função de densidade de probabilidades

A função densidade de probabilidades de T denotada por $f(t)$ satisfaz as seguintes condições (MEYER, 1983) :

- $f(t) \geq 0$ para todo $t \geq 0$;
- $\int_0^\infty f(t)\,dt = 1$;
- $P(a \leq T \leq b) = \int_a^b f(t)\,dt$, para todo $0 \leq a \leq b$.

2.1.2.2 Função de sobrevivência

A função de sobrevivência de T representa a probabilidade de um elemento não falhar até um determinado tempo t, ou seja, de sobreviver além desse tempo. Com isso, tem-se como função de sobrevivência (COLOSIMO; GIOLO, 2006):

$$S(t) = P(T > t) = \int_t^\infty f(t)\,dt,\ t \geq 0, \tag{2.1.1}$$

para qual, $S(t)$, é uma função não-crescente e absolutamente contínua, tal que $\lim_{t\to 0} S(t) = 1$ e $\lim_{t\to\infty} S(t) = 0$. A partir disso, a função de distribuição de T pode ser denotada por $F(t) = 1 - S(t)$, isto é, a probabilidade de um elemento não sobreviver ao tempo t.

2.1.2.3 Função de risco

A função de risco, também conhecida como função taxa de falha, caracteriza a taxa de falha instantânea em um terminado tempo$t$, ou seja, é o limite da probabilidade do evento de interesse ocorrer no intervalo $[t, t + \Delta t)$ dado que sobreviveu ao tempo t, sobre o intervalo de tempo Δt. Expressa por:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}, \; t \geq 0 \tag{2.1.2}$$

A função $h(t)$ tem a capacidade de descrever como a taxa de falha se modifica ao longo do tempo, podendo assumir comportamento constante, crescente ou decrescente e também descrever o comportamento do tempo de sobrevivência, apresentando forma unimodal ou a forma de curva da banheira.

2.1.2.4 Função de risco acumulado

A função de risco acumulado, também denotada por função taxa de falha acumulada, representa o risco acumulado do elemento no tempo t. Representada por:

$$H(t) = \int_0^t h(u)du, \; t \geq 0. \tag{2.1.3}$$

Como as funções $f(t)$, $S(t)$, $h(t)$ e $H(t)$ são matematicamente equivalentes, tem-se que:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} \log S(t), \tag{2.1.4}$$

$$H(t) = \int_0^t h(u)\, du = -\log S(t), \tag{2.1.5}$$

$$f(t) = \frac{d}{dt} \log S(t). \tag{2.1.6}$$

2.2 Estimador de Kaplan-Meier

O Estimador de Kaplan-Meier é uma técnica estatística não paramétrica utilizada para estimar a função de sobrevivência em dados censurados e é definido como (KAPLAN;

MEIER, 1958):

$$\hat{S}(t) = \prod_{j:t_j<t} \left[1 - \frac{d_j}{n_j}\right] \tag{2.2.1}$$

onde:

- $t_1 < t_2 < ... < t_k$, os k tempos distintos e ordenados de falha,
- n_j é o número de indivíduos que estão sob risco no tempo t_j,
- d_j número de indivíduos que experimentaram o evento de interesse no tempo t_j.

A função de sobrevivência em (2.2.1) é caracterizada como uma função escada, apresentando degraus nos momentos em que falhas são observadas. Entre esses pontos de falha, a função permanece constante. Contudo, quando ocorre uma falha, a função de sobrevivência sofre uma queda.

Além disso, essa função permite comparar os tempos de falha com base em diferentes variáveis qualitativas, com isso, é possível entender como fatores específicos podem influenciar a ocorrência de falhas em diferentes grupos ou categorias de dados.

2.3 Modelo de riscos porporcionais de Cox

O modelo de regressão de Cox possibilita a estimação do efeito das covariáveis de interesse nos tempos até a ocorrência do evento em estudo .

A expressão geral do modelo de Cox é dada por (COX, 1972):

$$h(t|\mathbf{x}) = h_0(t)g\{\mathbf{x}'\boldsymbol{\beta}\}, \tag{2.3.1}$$

em que $h(t|\mathbf{x})$ é a função de risco na presença das covariáveis $\mathbf{x}' = (x_1, x_2, ..., x_p)$, $h_0(t)$ é o risco base (risco de um indivíduo quando todas as covariáveis são nulas), $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)$ é o vetor de coeficientes associado a $\mathbf{x}$, e $g\{\mathbf{x}'\boldsymbol{\beta}\}$, uma função de ligação não negativa, tal que $g(0) = 1$, frequentemente apresentada por:

$$g\{\mathbf{x}'\boldsymbol{\beta}\} = \exp\{\mathbf{x}'\boldsymbol{\beta}\} = \exp\{\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p\}. \tag{2.3.2}$$

Este modelo é intitulado como o modelo de riscos proporcionais de Cox, uma vez que é possível estimação do efeito das covariáveis a partir da proporcionalidade dos riscos

ao longo de todo o tempo observado. Quando não é especificada uma forma paramétrica para $h_0(t)$, o modelo (2.3.1) é dito ser semi-paramétrico, visto que a função $g\{\mathbf{x}'\boldsymbol{\beta}\}$ é uma função paramétrica.

Ademais, a suposição básica deste modelo é a proporcionalidade das taxas de falha constante no tempo. De fato, a razão entre os riscos de dois indivíduos i e l,

$$\frac{h(t|\mathbf{x}_i)}{h(t|\mathbf{x}_l)} = \frac{h_0(t)\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{h_0(t)\exp\{\mathbf{x}_l'\boldsymbol{\beta}\}} = \exp\{\mathbf{x}_i'\boldsymbol{\beta} - \mathbf{x}_l'\boldsymbol{\beta}\}, \tag{2.3.3}$$

não depende de t.

2.3.1 Ajuste do Modelo

De acordo com Colosimo e Giolo (2006), os coeficientes $\boldsymbol{\beta}$ presentes no modelo de Cox precisam ser estimados a partir das observações amostrais contudo, o uso da função de verossimilhança usual para estimar $\boldsymbol{\beta}$ é inadequado, visto que há presença de um componente não paramétrico, $h_0(t)$.

Diante disso, o método de máxima verossimilhança parcial, proposto por Cox (1975), é o mais adequado para a estimação dos parâmetros desconhecidos $\boldsymbol{\beta}$ neste contexto. A construção deste método considera uma amostra de n indivíduos com $m \leq n$ falhas distintas nos tempos $t_1 < t_2 < ... < t_m$, e a probabilidade condicional da i-ésima observação vir a falhar no tempo t_i, conhecendo quais observações estão sob risco em t_i, é:

$$P[\text{indivíduo falhar em } t_i| \text{ uma falha em } t_i \text{ e história até } t_i] =$$

$$\frac{P[\text{indivíduo falhar em } t_i| \text{ sobreviveu a } t_i \text{ e história até } t_i]}{P[\text{ uma falha em } t_i| \text{ e história até } t_i]} =$$

$$= \frac{h(t_i|\mathbf{x}_i)}{\sum_{j\in R(t_i)} h(t_i|\mathbf{x}_j)} = \frac{h_0(t_i)\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{\sum_{j\in R(t_i)} h_0(t_i)\exp\{\mathbf{x}_j'\boldsymbol{\beta}\}} = \frac{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{\sum_{j\in R(t_i)}\exp\{\mathbf{x}_j'\boldsymbol{\beta}\}}, \tag{2.3.4}$$

no qual $R(t_i)$ é o conjunto de todos os indivíduos sob risco no tempo t_i. Com isso, é possível observar que a partir da probabilidade condicional na história de falhas e censuras até o tempo t_i, o componente não-paramétrico é eliminado. Dessa forma, a função de

verossimilhança parcial é expressa por:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{\sum_{j\in R(t_i)} \exp\{\mathbf{x}_j'\boldsymbol{\beta}\}} = \prod_{i=1}^{n} \left(\frac{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{\sum_{j\in R(t_i)} \exp\{\mathbf{x}_j'\boldsymbol{\beta}\}} \right)^{\delta_i}, \tag{2.3.5}$$

em que δ_i é o indicador de falha, que assume o valor 1 se t_i é um tempo observado ou 0 se t_i é um tempo censurado. Os parâmetros $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)$ são estimados por meio da maximização do logaritmo da função $L(\boldsymbol{\beta})$, que é obtida a partir da resolução do seguinte sistema de equações:

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_k} = \sum_{i=1}^{n} \delta_i \left[x_{ik} - \frac{\sum_{j\in R(t_i)} x_{jk} \exp\{\boldsymbol{\beta}'\mathbf{x}_j\}}{\sum_{j\in R(t_i)} \exp\{\boldsymbol{\beta}'\mathbf{x}_j\}} \right] = 0, \qquad k = 1, 2, ..., p \tag{2.3.6}$$

em (2.3.6), $\ell(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta})$ e x_{ik} é o valor da k-ésima covariável do indivíduo i.

A função de verossimilhança parcial em (2.3.6) não pressupõe a possibilidade de empates nos valores observados, contudo Breslow (1975) propôs uma aproximação em que $\mathbf{s}_i$ é um vetor formado pela soma das correspondentes p covariáveis para os indivíduos que falham no mesmo tempo t_i. Com isso, a aproximação possibilita que observações empatadas sejam consideradas e considera a seguinte função de verossimilhança parcial:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{\exp\{\mathbf{s}_i'\boldsymbol{\beta}\}}{\left[\sum_{j\in R(t_i)} \exp\{\mathbf{x}_j'\boldsymbol{\beta}\}\right]^{d_i}}, \tag{2.3.7}$$

em que d_i é o número de falhas no tempo t_i e $i = 1, 2, ..., m$.

2.3.2 Funções relacionadas com $h_0(t)$

Funções relacionadas com $h_0(t)$, principalmente $H_0(t)$ e $S_0(t)$, são de grande relevância. Com isso, tem-se como função de sobrevivência base:

$$S_0(t) = \exp\{-H_0(t)\}, \tag{2.3.8}$$

e como função de sobrevivência para um conjunto covariáveis $\mathbf{x}$:

$$S(t|\mathbf{x}) = [S_0(t)]^{\exp\{\mathbf{x}'\boldsymbol{\beta}\}}. \tag{2.3.9}$$

Como visto em (2.3.4), na construção da função de verossimilhança parcial, $h_0(t)$

é eliminado da função por ser um componente não-paramétrico, com isso para a estimação de $H_0(t)$, Breslow (1975) propôs o estimador de Nelson-Aalen-Breslow expresso por:

$$\hat{H}_0(t_i) = \sum_{i:t_i \leq t} \frac{d_i}{\sum_{j \in R(tj)} \exp\{\mathbf{x}'_j \hat{\boldsymbol{\beta}}\}} \tag{2.3.10}$$

em que d_i é o número de falhas em t_i e $\hat{\boldsymbol{\beta}}$ são os estimadores de $\boldsymbol{\beta}$ obtidos a partir da verossimilhança parcial.

Com base no estimador de Nelson-Aelen-Breslow, tem-se como função de sobrevivência base estimada:

$$\hat{S}_0(t) = \exp\{-\hat{H}_0(t)\}, \tag{2.3.11}$$

e como função de sobrevivência estimada para um conjunto covariáveis $\mathbf{x}$:

$$\hat{S}(t|x) = \left[\hat{S}_0(t|x)\right]^{\exp\{\mathbf{x}'\hat{\boldsymbol{\beta}}\}}. \tag{2.3.12}$$

2.3.3 Interpretação dos parâmetros

Os coeficientes $\boldsymbol{\beta}$ no modelo de regressão de Cox, são capazes de mensurar os efeitos das covariáveis sobre a taxa de falha sendo que, uma covariável consegue desacelerar, ou acelerar, a função de risco.

Para a interpretação dos coeficientes deve-se recorrer a propriedade de proporcionalidade dos riscos. Considerando dois indivíduos (i e l) que possuem valores iguais paras as covariáveis, com exceção da $p-$ésima, a taxa de falha é expressa por:

$$\frac{h(t|\mathbf{x}_i)}{h(t|\mathbf{x}_l)} = \frac{\exp\{\beta_p x_{ip}\}}{\exp\{\beta_p x_{lp}\}} = exp\{\beta_p(x_{ip} - x_{lp})\} \tag{2.3.13}$$

em que x_{ip} é o valor da p-ésima covariável do indivíduo i e β_p é o p-ésimo coeficiente de regressão.

A razão apresentada em (2.3.13) é uma razão de riscos, dessa forma, por exemplo, supondo que x_p é a covariável dicotômica sexo em que $x_p = 1$ (masculino) e $x_p = 0$ (feminino), entende-se que o risco de falha dos indvíduos do sexo masculino é $\exp(\beta_p)$ vezes o risco de falha dos indivíduos do sexo feminino, mantendo-se fixas as demais covariáveis. (SANTOS, 2017)

2.3.4 Adequação do modelo

Como citado anteriormente, o modelo de regressão de Cox supõe proporcionalidade dos riscos, logo as funções de risco não podem se cruzar. Em vista disso, para se ter um modelo adequado, necessita-se que este pressuposto seja atendido e uma de suas verificações pode-se dar a partir de uma técnica gráfica, que consiste em particionar os dados em m estrados conforme as m categorias de certa covariável e estimar $\hat{H}_{0j}(t)$, para cada extrato da covariável.

Caso a suposição seja válida, a partir de um gráfico, as curvas log $\hat{H}_{0j}(t)$ versus t, ou $\log(t)$, irão apresentar diferenças contantes ao logo do tempo, isto é, serão aproximadamente paralelas.

Outra forma de verificar a suposição de riscos proporcionais no modelo de Cox é por meio da análise de resíduos de Schoenfeld. Quando o $i-$ésimo indivíduo, com vetor de covariáveis $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$, falhar, o vetor de resíduos de Schoenfeld $r_i = (r_{i1}, ..., r_{ip})'$ em que cada componente r_{iq}, para $q = 1, ..., p$, para esse indivíduo é definido por (COLOSIMO; GIOLO, 2006):

$$r_{iq} = x_{iq} - \frac{\sum_{j \in R(t_i)} x_{jq} \exp\{\mathbf{x}_j' \hat{\boldsymbol{\beta}}\}}{\sum_{j \in R(t_i)} \exp\{\mathbf{x}_j' \hat{\boldsymbol{\beta}}\}} \tag{2.3.14}$$

Os resíduos padronizados de Schoenfeld são definidos por:

$$\mathbf{s}_i^* = [I(\hat{\boldsymbol{\beta}})]^{-1} \mathbf{r}_i \tag{2.3.15}$$

em que $I(\hat{\boldsymbol{\beta}})$ é matriz de informação observada.

Caso os riscos sejam proporcionais, o gráfico de $\boldsymbol{\beta}_q(t)$ versus $\mathbf{s}_{iq}^*$ se aproximará de uma reta, indicando que os resíduos de Schoenfeld não apresentam tendências no tempo.

2.3.5 Avaliação do ajuste do modelo

Além de verificar a suposição dos riscos proporcionais, também é necessário realizar uma avaliação geral do modelo de Cox para certificar-se a sua adequabilidade aos dados. Sendo assim, os resíduos de Cox-Snell, propostos por Cox e Snell (1968), podem ser utilizados, tais resíduos são definidos por:

$$\hat{e}_i = \hat{H}_0(t_i) \exp\left\{\sum_{k=1}^{p} x_{ik}\beta_k\right\}, \tag{2.3.16}$$

em que o $\hat{H}_0(t_i)$ é o estimador de Breslow (1975) para a função de risco acumulado.

Para o modelo seja considerado adequado, os resíduos $\hat{e}_i$ devem seguir uma distribuição exponencial padrão e, portanto, o gráfico de $\hat{e}_i$ versus $\hat{H}(\hat{e}_i)$ deve ser aproximadamente uma reta.

2.4 Modelo de Cox com covariável dependente no tempo

Para a análise de sobrevivência, o modelo de Cox (1972) é fortemente utilizado devido a sua flexibilidade. A implementação do efeito da covariáveis dependentes no tempo, isto é, covariáveis que se modificam ao longo do tempo, pode ser facilmente realizada a partir de uma extensão.

De acordo com Kalbfleisch e Prentice (2011), as covariáveis dependentes no tempo possuem duas classificações, sendo covariáveis internas e externas. As covariáveis internas são medidas durante o estudo e obrigatoriamente devem ser medidas enquanto o objeto de estudo sobrevive, por outro lado, as covariáveis externas não necessitam da sobrevivência do objeto de estudo para existir.

Visto que faz-se necessário a extensão do modelo de Cox para a incorporação de covariáveis dependentes no tempo, Therneau e Grambsch (2000) propôs:

$$h(t|Z(t)) = h_0(t) \exp\left\{\theta'\mathbf{x} + \gamma X g(t)\right\}, \tag{2.4.1}$$

em que θ' e γ' são, respectivamente, os coeficientes das covariáveis fixas e das covariáveis dependentes no tempo. Ao considerar que $Z(t)$ representa o vetor das covariáveis no tempo t, tem-se:

$$Z(t) = [x_1, x_2, ..., x_{f_1}, X_1 g(t), X_2 g(t), ...X_{f_2}, g(t)], \tag{2.4.2}$$

possibilitando, assim a generalização do modelo:

$$h(t|Z(t)) = h_0(t) \exp\left\{\boldsymbol{\beta}' Z(t)\right\}, \tag{2.4.3}$$

em que $\boldsymbol{\beta} = (\theta_1, \theta_2, ..., \theta_{f_1}, \gamma_1, \gamma_2, ..., \gamma_{f_2})$ e $p = f_1 + f_2$, representa o número de coeficientes do modelo. Os valores das covariáveis $Z(t)$ dependem do tempo t.

Note que a razão de risco no tempo t para dois indivíduos i e l, expressa por:

$$\frac{h_i(t|Z(t))}{h_l(t|Z(t))} = \exp\left\{Z_i(t)\boldsymbol{\beta} - Z_l(t)\boldsymbol{\beta}\right\}, \tag{2.4.4}$$

em que $Z_i(t)$ é o vetor de covariáveis (2.4.4) do indivíduo i, depende de t. Assim, o modelo (2.4.3) não é mais um modelo de riscos proporcionais.

Ao estender o logaritmo da função de verossimilhança parcial, as estimativas de $\boldsymbol{\beta}$ é obtida a partir da resolução do seguinte sistema de equações:

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left[Z_i(t) - \frac{\sum_{j\in R(t_i)} Z_j(t_i)\exp\{Z_j(t_i)\boldsymbol{\beta}\}}{\sum_{j\in R(t_i)}\exp\{Z_j(t_i)\boldsymbol{\beta}\}}\right] = 0, \tag{2.4.5}$$

em que $Z_i(t)$ é o vetor de covariáveis (2.4.2) do indivíduo i, $R(t_i)$ é o número de indivíduos sob risco em t_i e $\boldsymbol{\beta} = (\theta_1, \theta_2, ..., \theta_{q1}, \gamma_1, \gamma_2, ..., \gamma_{q2})$ e $p = f_1 + f_2$ é o vetor de coeficientes.

2.4.1 Interpretação dos Coeficientes

Como o modelo não é mais considerado como de taxa de falhas proporcionais e a razão das funções de risco é dependente do tempo, então, para a interpretação dos coeficientes $\boldsymbol{\beta}$ do modelo deve-se considerar o tempo t e cada coeficiente $\boldsymbol{\beta}_l$, $l = 1, 2, ..., p$, que representa o logaritmo da razão de taxas de falhas o qual o valor da l-ésima covariável no tempo t difere de uma unidade, quando os valores das outras covariáveis são mantidos fixos neste tempo (COLOSIMO; GIOLO, 2006).

Portanto, os coeficientes podem ser interpretados como o logaritmo da razão de risco para dois indivíduos cujo valor da $l-$ésima cováriavel no tempo t difere de uma unidade quando as outras covariáveis assumem o mesmo valor neste tempo.

2.4.2 Avaliação do modelo

No contexto em que se tem covariáveis dependentes no tempo no modelo de Cox, a suposição de riscos proporcionais é violada, portanto, para uma covariável que foi medida no início do estudo, é importante que seu efeito no resultado não seja fixo ou constante ao longo do tempo de acompanhamento.

Quanto a avaliação do modelo, o modelo de Cox com covariáveis dependentes no tempo, também pode ser avaliado por meio dos resíduos de Cox-Snell definido, por:

$$\hat{e}_i = \hat{H}_0(t_i) \exp\left\{ \sum_{k=1}^{p} Z_k(t_i)\beta_k \right\}. \tag{2.4.6}$$

3 Conjunto de dados

Para a construção deste trabalho utilizou-se três bases de dados, sendo a Plataforma Sucupira para extração de informações sobre a titulação de doutores no Brasil, a RAIS e CAGED para informações de emprego. Essas bases de dados têm se mostrado apropriadas para a análise do impacto das informações relacionadas à pós-graduação nas políticas públicas brasileiras.

Além disso, para alcançar o objetivo principal deste trabalho, foram adotadas metodologias específicas, realizados tratamentos adequados nos dados e também realizada a cruzamento das diferentes bases de dados utilizadas. Esses processos são apresentados e explicados detalhadamente a seguir.

3.1 Plataforma Sucupira (CAPES)

A Plataforma Sucupira tem como objetivo principal coletar, em tempo real, dados dos programas de pós-graduação em um único sistema. A idealização e a construção foram realizadas pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) em parceria com a Universidade Federal do Rio Grande do Norte (UFRN).

A partir da Plataforma é possível ter maior transparência para a comunidade acadêmica quanto a geração de dados, avaliação e procedimento executados pela CAPES. Além disso, essas informações computadas são utilizada como base de referência no Sistema Nacional de Pós-Graduação (SNPG). (ARCANJO, 2012)

Através da base de dados, foram coletadas informações abrangentes de indivíduos, incluindo CPF, nome, sexo, idade de titulação e data de obtenção do título, além de informações relevantes sobre os programas de pós-graduação, como a nota atribuída na avaliação da Capes e a área do conhecimento à qual o programa pertence.

Para este estudo, foram considerados exclusivamente os doutores que obtiveram seu primeiro título de doutorado realizado no Brasil no período de 2013 a 2021. Essa seleção se baseia no fato de que, uma vez que o objetivo deste trabalho envolve o primeiro vínculo empregatício em que a atividade econômica principal está relacionada às atividades acadêmicas, a titulação de doutorado aumenta as chances de inserção no mercado de trabalho nesse âmbito acadêmico. Ao considerar apenas a titulação de doutorado mais recente do indivíduo, corre-se o risco de excluir aqueles que já estavam vinculados a essa atividade durante o período da titulação mais recente, o que poderia distorcer os

resultados finais do estudo.

3.2 RAIS

Estabelecida pelo Decreto nº 76.900, de 23/12/75, Relação Anual de Informações Sociais (RAIS) tem como objetivos amparar as necessidades de controle da atividade trabalhista no País, providenciar dados para a e elaboração de estatísticas do trabalho e a disponibilizar informações do mercado de trabalho às entidades governamentais. (MTE, 2021)

Conforme a revogação presente no Decreto Nº 10.854, de 10/11/2021, a RAIS identifica o empregador pelo número de inscrição no Cadastro de Pessoas Jurídicas (CNPJ), no Cadastro Nacional de Obras e no Cadastro de Atividade Econômica da Pessoa Física, já o empregado reconhecido pelo número de inscrição no Cadastro de Pessoas Físicas (CPF). (BRASIL, 2021)

Além disso, todo estabelecimento deve fornecer os dados solicitados de cada um de seus empregados para o Ministério do Trabalho e Previdência através da RAIS.

Para este estudo, foram utilizadas as bases de dados da RAIS, no período de 2012 a 2021, filtradas a partir dos CPFs extraídos na Plataforma Sucupira. Em um primeiro momento, verificou-se a presença de vínculo ativo na data de titulação, informação importante para a construção do estudo, e posteriormente examinado os vínculos que foram adicionados após a data de titulação.

3.3 CAGED

O Cadastro Geral de Empregados e Desempregados, criado pela Lei 4.923 de 23/12/1965, foi desenvolvido com a finalidade de acompanhar e fiscalizar a admissão e demissão de trabalhadores regidos pela legislação trabalhista, e sua atualização ocorre mensalmente. (MTE, 2016)

O CAGED difere da RAIS em sua metodologia de construção. Enquanto a RAIS é baseada em informações anuais que incluem todos os vínculos empregatícios ativos até 31 de dezembro do ano correspondente, o CAGED é elaborado com base nas movimentações de admissão e desligamento registradas durante o mês de referência. Por essa razão, o CAGED não tem a capacidade de identificar vínculos empregatícios já existentes.

Dessa forma, para este estudo, o CAGED foi utilizado, assim como a RAIS,

para verificar se o vínculo empregatício que estava ativo no ano anterior à titulação de doutorado permaneceu ativo na data de titulação. Além disso, analisar se durante esse período, o indivíduo adquiriu algum novo vínculo empregatício e se manteve nele mesmo após a obtenção da titulação.

Ademais, o CAGED também permite a obtenção das movimentações de vínculo dos indivíduos presentes na extração da Plataforma Sucupira após a data de titulação, sendo assim possível a análise do tempo até a aquisição de um vínculo empregatício com enfoque na área acadêmica.

As bases de dados do CAGED que foram utilizadas são referentes aos anos de 2013 e 2021.

3.4 Seleção de titulados entre 2013 e 2021

Como dito na Seção 3.1, os indivíduos selecionados para este estudo concluíram seu doutorado entre os anos de 2013 e 2021. No entanto, para alcançar o objetivo deste trabalho, que é analisar do tempo até a obtenção de um vínculo empregatício na área acadêmica, é fundamental verificar se, na data de titulação, esses indivíduos estavam empregados em um estabelecimento cuja atividade econômica principal estava relacionada às atividades acadêmicas. Essas atividades são categorizadas nas seções M ("Atividades profissionais, científicas e técnicas") e P ("Educação") da Classificação Nacional de Atividades Econômicas (CNAE 2.0).

Para obter essa informação, foi realizado um processo de análise das bases de dados da RAIS e do CAGED, conforme descrito nas Seções 3.2 e 3.3. O tratamento das duas bases de dados foi semelhante, envolvendo a extração das datas de admissão e desligamento de cada vínculo de emprego, quando aplicável. Essas informações foram usadas para verificar se a data de titulação de doutorado estava dentro do período de cada vínculo.

No caso em que a data de obtenção do título de doutorado estava dentro do período do vínculo, uma análise adicional foi realizada com base na CNAE associada a cada vínculo. Se o vínculo estivesse relacionado às categorias M ou P da CNAE, os indivíduos eram excluídos da análise, uma vez que já contemplavam o objetivo deste estudo na data de titulação.

Além de identificar os indivíduos que foram excluídos da análise, esse processo também permitiu identificar os doutores que estavam com vínculo de emprego ativo na

data de titulação.

3.5 Construção do conjunto de dados

As informações sobre empregos após a data de titulação foram obtidas através do CAGED e da RAIS. No entanto, essas informações precisaram passar por diversos tratamentos de dados, feitos a partir do *software* R, para serem aplicadas no modelo de riscos proporcionais de Cox, que utilizou covariáveis dependentes do tempo.

Após o primeiro tratamento dos dados, foi obtida uma base de dados em que cada linha referia-se a um determinado vínculo empregatício. As informações coletadas incluíam a CNAE do estabelecimento, a data de admissão e a data de desligamento. Para os indivíduos que deixaram aquele vínculo durante o período entre a data de titulação e 31 de dezembro de 2021, a data real de desligamento foi registrada na base de dados. Se o indivíduo permaneceu no vínculo até a data final do estudo, foi registrado o dia 31/12/2021 como data de desligamento, visto que é o período que compreende o estudo, portanto data de censura.

Foram desconsiderados os vínculos empregatícios em que o período entre a admissão e o desligamento fosse inferior a 1 mês. Além disso, também foram excluídos os vínculos que foram obtidos após a aquisição do primeiro vínculo cujo a CNAE estava classificada como M ou P.

Para a construção da base de dados para o modelo de regressão de cox, considerou-se as seguintes variáveis:

- **id** - Número de identificação do doutor;
- **Tempo** - Variável que indica o tempo, em meses, em que adquiriu-se o vínculo empregatício nas CNAEs M e P ou tempo em que o indivíduo foi censurado;
- **Censura** - Variável que assume o valor 0 para indíviduos censurados, ou seja, casos em que o indivíduo ainda não adquiriu vínculo ativo nas CNAEs M e P, e assume o valor 1 no tempo em que ocorre a falha, ou seja, no momento em que o indivíduo adquire um vínculo ativo nas CNAEs M e P.
- **Vínculo Ativo** - Covariável que indica o status do vinculo empregaticio do individuo na data de titulação, codificada como 1 para representar a presença de um vínculo ativo e como 0 para indicar a ausência de um vínculo ativo;
- **sexo** - Covariável fixa que representa o sexo do indivíduo

- **Classificação Doutor** - Covariável fixa que indica se o indivíduo é jovem doutor, que titulou com idade menor ou igual a 32 anos, ou doutor, que titulou com idade acima de 32 anos;
- **Área** - Covariável fixa que indica a grande área do conhecimento à qual o programa da titulação pertence;
- **conceito** - Covariável fixa que indica a nota atribuída na avaliação da Capes ao programa da titulação pertence;

Já para a construção da base de dados para o modelo de regressão de Cox com covariáveis dependentes no tempo, foram agrupados os períodos em comum nos quais o indivíduo estava com vínculo ativo. Para mais, foram acrescidos também, quando necessário, períodos em que o indivíduo não possuía nenhum vínculo ativo. As variáveis construídas e utilizadas para a realização desta análise foram:

- **id** - Número de identificação do doutor;
- **Início** - Variável que indica o tempo inicial da movimentação, medida em meses;
- **Final** - Variável que indica o tempo final da movimentação, sendo o último tempo de falha ou censura, medida em meses;
- **Censura** - Variável que assume o valor 0 nos tempos em que o evento ainda não ocorreu, ou seja, nos períodos em que o indivíduo ainda não possui vínculo ativo nas CNAEs M e P, e assume o valor 1 no tempo em que ocorre a falha, ou seja, no momento em que o indivíduo adquire um vínculo ativo nas CNAEs M e P. Para indivíduos censurados o valor é sempre igual a 0;
- **Vínculo Ativo** - Covariável dependente no tempo que indica o status do vínculo empregatício do individuo, codificada como 1 para representar a presença de um vínculo ativo e como 0 para indicar a ausência de um vínculo ativo;

A Figura 3.5.1 apresenta a visualização das primeiras linhas do conjunto de dados.

id	Início	Final	Censura	Vínculo Ativo	Sexo	Classificação Doutor	Área	Nota Capes
1	0	42	1	0	F	Jovem doutor	Ciências Agrárias	5
2	0	26	0	0	M	Doutor	Ciências Biológicas	4
2	26	35	1	0	M	Doutor	Ciências Biológicas	4
2	35	36	0	0	M	Doutor	Ciências Biológicas	4
2	36	38	1	0	M	Doutor	Ciências Biológicas	4
2	38	39	0	1	M	Doutor	Ciências Biológicas	4
3	0	20	0	0	M	Jovem doutor	Ciências Humanas	7
3	20	41	1	1	M	Jovem doutor	Ciências Humanas	7
4	0	25	0	1	M	Doutor	Ciências Humanas	4
5	0	16	0	0	F	Jovem doutor	Ciências Biológicas	7
5	16	32	1	0	F	Jovem doutor	Ciências Biológicas	7
5	32	93	0	0	F	Jovem doutor	Ciências Biológicas	7
6	0	70	1	1	M	Jovem doutor	Ciências da Saúde	4
7	0	13	1	1	F	Doutor	Multidisciplinar	4

Figura 3.5.1: Visualização das 16 primeiras linhas do banco de dados.

A partir da Figura 3.5.1 é possível perceber que a cada mudança na covariável vínculo, que é a dependente no tempo, uma nova linha é acrecida no banco de dados, como é o caso do doutor de $id = 2$, que modificou seu status de vínculo empregatício 4 vezes até a falha.

No caso em que $id = 4$, verificou-se que o doutor não possuía nenhum vínculo empregatício na data de titulação. No entanto, após 25 meses da data de titulação, ele adquiriu um vínculo empregatício cuja atividade econômica estava classificada com M ou P.

Já no caso em que $id = 6$, constatou-se que o doutor possuía vínculo empregatício ativo na data de titulação. Posteriormente, após 70 meses da data de titulação, ele adquiriu um novo vínculo empregatício cuja atividade econômica estava classificada com M ou P.

4 Resultados

4.1 Análise descritiva

O número de indivíduos que obtiveram seu primeiro título de doutorado no Brasil entre 2013 e 2021 foi de 182.161. No entanto, após aplicar os filtros descritos na Seção 3.4, esse número foi reduzido para 100.401.

Além da exclusão dos doutores com vínculo empregatício ativo associados às categorias M ou P da CNAE na data de titulação, também foi necessário remover os doutores cujos programas de pós-graduação receberam a nota A, 2 ou 3 na avaliação da Capes. Isso ocorre porque a nota mínima exigida para o funcionamento dos programas de doutorado é 4, e programas com nota inferior à 3 são descontinuados.

Também foi necessário excluir os doutores que não possuíam a data de nascimento registrada na base de dados da Plataforma Sucupira. Isso ocorre porque a informação de data de nascimento é fundamental para determinar se um indivíduo é considerado um jovem doutor ou apenas um doutor.

Após aplicar todas essas exclusões, restaram 98.581 doutores para análise.

A partir da Tabela 4.1.1 nota-se que a maior parte, 56,77%, dos doutores do estudo pertence ao sexo feminino, enquanto 43,23% ao sexo masculino.

Tabela 4.1.1: Tabela de frequência da variável sexo.

Sexo	Frequência absoluta	Frequência Relativa %
Feminino	55.959	56,77%
Masculino	42.612	43,23%

Observa-se, com a Tabela 4.1.2, que 45,61% dos titulados são considerados como jovem doutor, ou seja, adquiriram o título de doutor antes dos 32 anos. Por outro lado, a maioria, 54,39%, adquiriram o título de doutor após os 32 anos.

Tabela 4.1.2: Tabela de frequência da variável classe doutor.

Classe	Frequência absoluta	Frequência Relativa %
Doutor	53.610	54,39%
Jovem doutor	44.961	45,61%

Ao analisar as informações dos programas aos quais os doutores pertenciam observa-se, a partir da Tabela 4.1.3, que as grandes áreas do conhecimento com as maiores

frequências são as ciências da saúde, ciências agrárias e ciências humanas. Essas três áreas juntas representam aproximadamente 46% da frequência relativa. Em contrapartida as grandes áreas com as menores frequências são multidisciplinar, ciências sociais aplicadas, linguística letras e artes.

Tabela 4.1.3: Tabela de frequência da variável grande área.

Grande Área	Frequência absoluta	Frequência Relativa %
Ciências da Saúde	17.861	18,12%
Ciências Agrárias	15.437	15,66%
Ciências Humanas	14.041	14,24%
Ciências Biológicas	12.748	12,93%
Ciências Exatas e da Terra	10.602	10,76%
Engenharias	8.531	8,65%
Multidisciplinar	7.402	7,51%
Ciências Sociais Aplicadas	7.020	7,12%
Linguística Letras e Artes	4.929	5,00%

Quanto a nota atribuída pela Capes a Tabela 4.1.4 mostra que as menores notas mais frequentes, tendo as notas 4 e 5 frequências relativas de 27,06% e 33,25% e a maiores notas são menos frequentes, tendo as notas 6 e 7 frequências relativas de 22,30% e 17,39%.

Tabela 4.1.4: Tabela de frequência da variável nota Capes.

Nota Capes	Frequência absoluta	Frequência Relativa %
4	26.677	27,06%
5	32.772	33,25%
6	21.981	22,30%
7	17.141	17,39%

Nota-se na Tabela 4.1.5 que a maior parte, 70,71%, dos doutores não possuía vínculo ativo na data de titulação, enquanto 29,29% possuía vínculo ativo na data de titulação.

Tabela 4.1.5: Tabela de frequência da variável vínculo ativo na data de titulação.

Vínculo Ativo	Frequência absoluta	Frequência Relativa %
Não	69.701	70,71%
Sim	28.870	29,29%

A análise de sobrevivência é caracterizada pela presença de censura, e neste trabalho em particular, 68,26% dos dados foram censurados, o que corresponde a 67.290 observações. Existem duas razões que justificam esses valores. A primeira está relacionada ao curto período de acompanhamento dos doutores que se formaram nos últimos anos. Como resultado, eles tiveram menos tempo de acompanhamento, o que resulta em

um maior número de casos censurados, conforme observado na Tabela 4.1.7.

A segunda razão está relacionada ao impacto no mercado de trabalho. A pandemia do Coronavírus desencadeou uma série de consequências, incluindo a suspensão de muitas oportunidades de emprego durante os anos de 2020 e 2021. Isso afetou diversas áreas, inclusive os concursos públicos, que foram interrompidos até 31 de dezembro de 2021, conforme estipulado na Lei Complementar Nº 173, de 27 de Maio de 2020.

Tabela 4.1.6: Tabela de censura e falha.

	Frequência absoluta	Frequência Relativa %
Censura	67.290	68,26%
Falha	31.291	31,74%

Tabela 4.1.7: Tabela de censura e falha por ano de titulação.

	Ano de Titulação								
	2013	2014	2015	2016	2017	2018	2019	2020	2021
Censura	3.743	4.528	5.415	6.484	7.519	8.823	10.275	9.451	11.052
Falha	4.395	4.514	4.689	4.694	4.295	3.798	2.800	1.502	604

A Figura 4.1.1 representa a função de sobrevivência estimada por Kaplan-Meier que indica a probabilidade de um indivíduo sobreviver até um determinado tempo (t), entre a obtenção do título de doutorado ao ingresso em atividade de caráter acadêmico.

No tempo igual a zero, 100% dos titulados estão sob risco (100% de sobreviventes) e a função de sobrevivência decai ao longo do tempo até um valor diferente de zero (aproximadamente 0,37) devido a presença das censuras resultantes do tempo final do estudo.

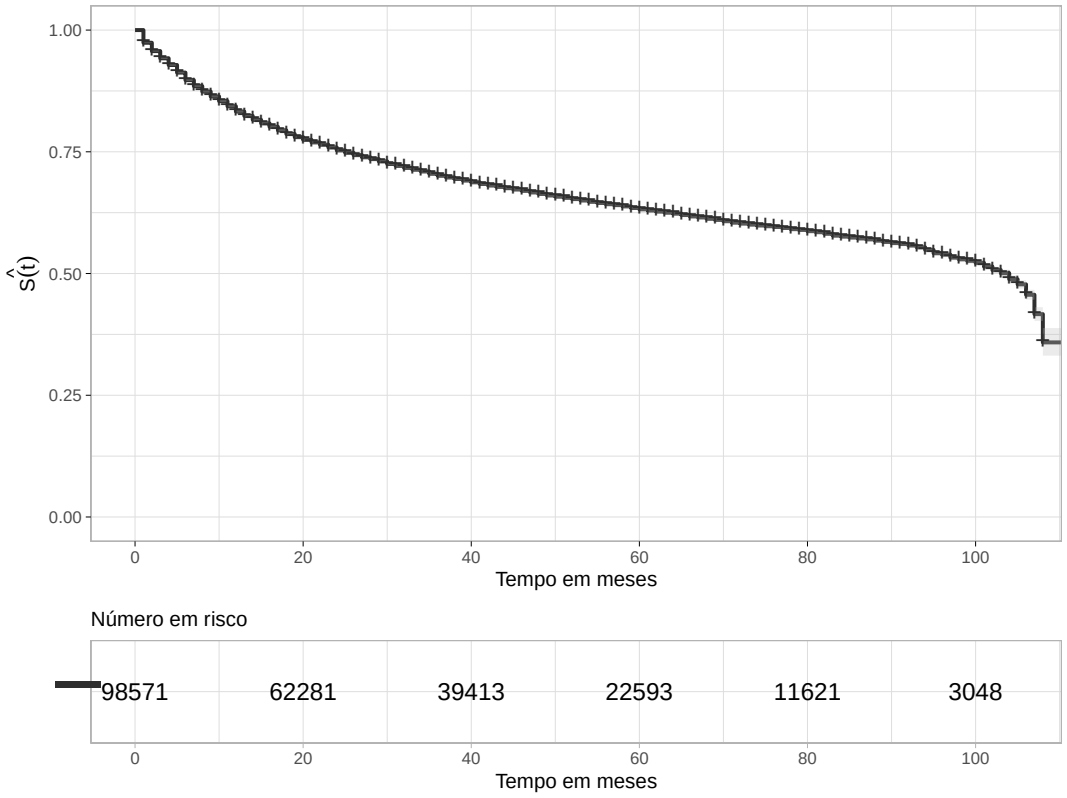

Figura 4.1.1: Curva estimada pelo método não-paramétrico de Kaplan-Meier para os tempos de sobrevivência dos doutores titulados no Brasil

Com a Figura 4.1.2, nota-se que não existem diferenças significativas entre as curvas de sobrevivências das categorias das covariáveis sexo e nota Capes. Já para a grande área do conhecimento, observa-se que as curvas se cruzam, mas é possível perceber que a área Multidisciplinar possui os maiores valores da função estimada ao longo do tempo, dessa forma, indivíduos que titulam nesta área, aparentemente, demoram mais tempo para adquirirem o vínculo nas CNAEs selecionadas.

Ademais, para a covariável vínculo ativo na data de titulação, tem-se que as curvas não se cruzam e a partir disso é possível entender que os não possuem vínculo ativo no momento da obtenção do título, adquirirem o vínculo nas CNAEs M ou P em menos tempo. Já para a classificação doutor, as curvas não se cruzam também e indivíduos classificados com jovem doutor, adquirirem o vínculo nas CNAEs M ou P em menos tempo.

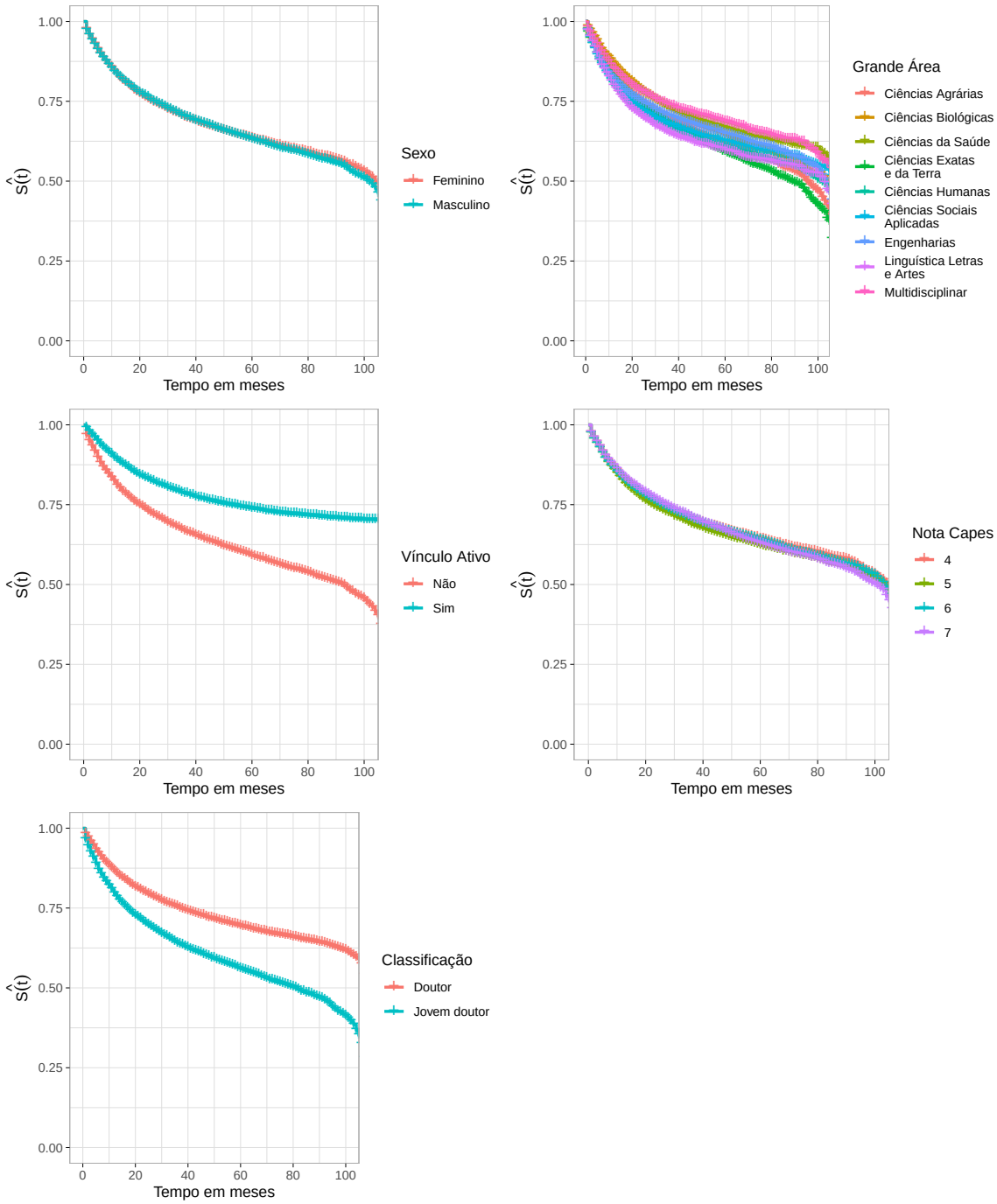

Figura 4.1.2: Curvas de sobrevivência das categorias das covariáveis estimada por Kaplan-Meier.

4.2 Modelo

Inicialmente, o modelo de regressão de Cox foi ajustado ao conjunto de dados com as covárivais fixas sexo, classificação doutor, grande área e nota Capes. Com isso, é necessário verificar se a suposição básica de proporcionalidade dos riscos é violada. Para tal verificação, utilizou-se os resíduos de Schoenfeld.

Observando a Figura 4.2.1, percebe-se, para todas as covariáveis fixas, que os gráficos apresentam uma reta sem uma tendência acentuada, indicando assim a não violação da suposição básica do modelo de Cox.

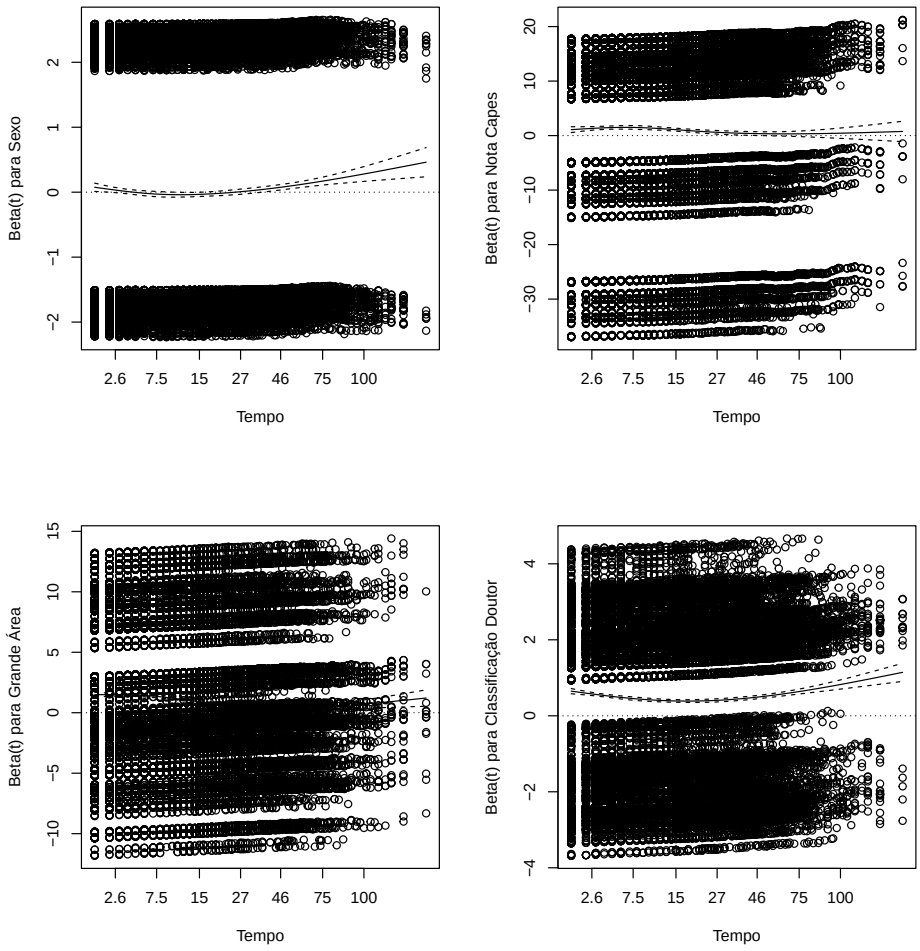

Figura 4.2.1: Resíduos de Schoenfeld para as covariáveis sexo, classificação doutor, grande área e nota Capes.

Para o modelo de Cox com covariáveis dependentes no tempo, a covariável que é dependente no tempo é o vínculo ativo que indica o status do vinculo empregatício do doutor em determinado período, sendo sim para representar a presença de um vínculo ativo e não para indicar a ausência de um vínculo ativo.

A partir da Tabela 4.2.1, percebe-se que a maioria, 84,60%, dos doutores modificaram seu status do vinculo empregatício pelo menos uma vez ao longo do tempo. Por outro lado, cerca de 15,40% modificaram seu status do vinculo empregatício pelo menos uma vez ao longo do tempo.

Tabela 4.2.1: Tabela de frequência da mudança de status do vinculo empregatício.

	Frequência absoluta	Frequência Relativa %
Mudou de status	83.388	84,60%
Não mudou de status	15.183	15,40%

O modelo de regressão de Cox com covariáveis dependentes no tempo 2.4.1, foi ajustado para os dados desse estudo considerando as covariáveis fixas: sexo, nota Capes, grande área, e classificação doutor; e a covariável vinculo ativo dependente no tempo. As estimativas dos parâmetros do modelo foram obtidas por meio do comando "coxph"do pacote "survival"do *software* R e são apresentadas na Tabela 4.2.2.

Tabela 4.2.2: Estimativas dos coeficientes β no modelo de regressão de Cox com covariáveis dependentes no tempo.

Covariáveis	β	$e^{\beta}(I.C.)95\%$.	Erro padrão	P-valor
Sexo				
Feminino	0	1	-	-
Masculino	0,02	1,02 (1,00;1,04)	0,01	0,09
Nota Capes				
Nota 4	0	1	-	-
Nota 5	0,01	1,01 (0,98;1,04)	0,01	0,55
Nota 6	-0,07	0,94 (0,91;0,97)	0,02	<0,001
Nota 7	-0,14	0,87 (0,84;0,90)	0,02	<0,001
Grande Área				
Ciências Agrárias	0	1	-	-
Ciências Biológicas	-0,15	0,86 (0,82;0,90)	0,02	<0,001
Ciências da Saúde	-0,01	0,99 (0,96;1,03)	0,02	0,80
Ciências Exatas e da Terra	0,11	1,11 (1,07;1,16)	0,02	<0,001
Ciências Humanas	0,32	1,38 (1,33;1,44)	0,02	<0,001
Ciências Sociais Aplicadas	0,28	1,32 (1,25;1,39)	0,03	<0,001
Engenharias	0,07	1,07 (1,02;1,12)	0,02	0,01
Linguística Letras e Artes	0,36	1,43 (1,35;1,51)	0,03	<0,001
Multidisciplinar	-0,06	0,94 (0,89;0,99)	0,03	0,02
Vínculo Ativo				
Não	0	1	-	-
Sim	-0,34	0,71 (0,69;0,73)	0,01	<0,001
Classificação doutor				
Doutor	0	1	-	-
Jovem doutor	0,54	1,71 (1,67;1,75)	0,01	<0,001

Notas: A categoria com β= 0 é o nível de referência.

O número de indivíduos no estudo é de 98.581, portanto é necessário que tenha-se cuidado ao analisar os resultados dos testes de hipótese contidos neste estudo, visto que devido ao expressivo número de observações o poder do teste é potencializado.

Com base na Tabela 4.2.2, observa-se que todas as covariáveis selecionadas, com exceção do sexo, demonstraram pelo menos uma categoria que se mostrou estatisticamente significante no modelo, considerando um nível de significância de 5%. Apesar disso, optou-se por manter a variável sexo, mesmo que não tenha sido significativa, e os titulados que pertencem ao sexo masculino possui risco 2% maior que os titulados do sexo feminino.

Quanto à nota Capes, observou-se que os riscos das Notas 4 e 5 são semelhantes. Além disso, verificou-se que quanto maior a nota, menor é o risco. Portanto, indivíduos que titularam em programas de doutorado com notas mais baixas tendem a obter vínculos empregatícios em atividades acadêmicas mais rapidamente do que aqueles titularam em programas com notas mais altas.

Em relação a grande área do conhecimento, notou-se que a área com menor risco foi a Ciências Biológicas um risco 14% menor (RR=0,86) em relação à Ciências Agrárias (nível de referência). Em contrapartida, as áreas com os maiores riscos foram Linguística, Letras e Artes e Ciências Humanas apresentando, respectivamente, riscos 43% e 38% maiores do que a área de referência.

O risco relativo de um doutor que possui vínculo é de 71% em relação ao risco de um doutor que não possui vínculo. Portanto, quando não se tem vínculo, a obtenção do vínculo empregatício em atividades acadêmicas ocorre mais rapidamente.

Os titulados considerados com Jovem doutor possuem o risco relativo 71% maior em comparação aos doutores com idade superior a 32 anos.

A adequação do modelo de riscos proporcionais de Cox foi avaliada através da análise dos resíduos de Cox-Snell, conforme ilustrado nas Figuras 4.2.2 e 4.2.3. Observa-se que, embora haja um desvio na cauda da distribuição dos resíduos, a função de sobrevivência dos resíduos se ajustou satisfatoriamente à função de sobrevivência da distribuição exponencial padrão. Cabe destacar que apenas 3% dos valores dos resíduos de Cox-Snell são maiores que 0,7, ponto em que se torna evidente um desvio significativo.

Portanto, assim como visto em Santos e Nakano (2015), pode-se considerar que para todas as covariáveis, a função de sobrevivência dos resíduos de Cox-Snell se ajustou bem à função de sobrevivência da exponencial padrão, bem como para o modelo de cox múltiplo com covariável dependente no tempo. Dessa forma, indicando bom ajuste.

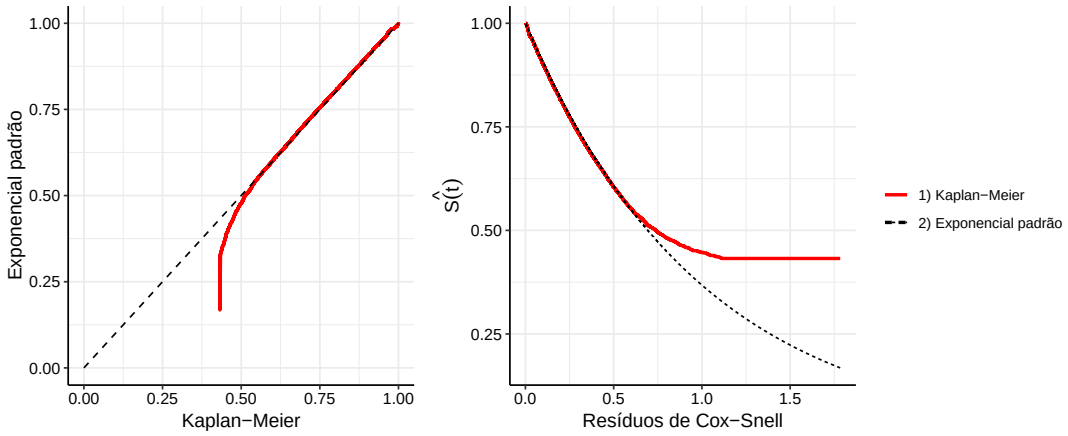

Figura 4.2.2: Comparação das funções de sobrevivência dos resíduos de Cox-Snell do modelo de Cox com covariáveis dependentes no tempo e da distribuição Exponencial padrão.

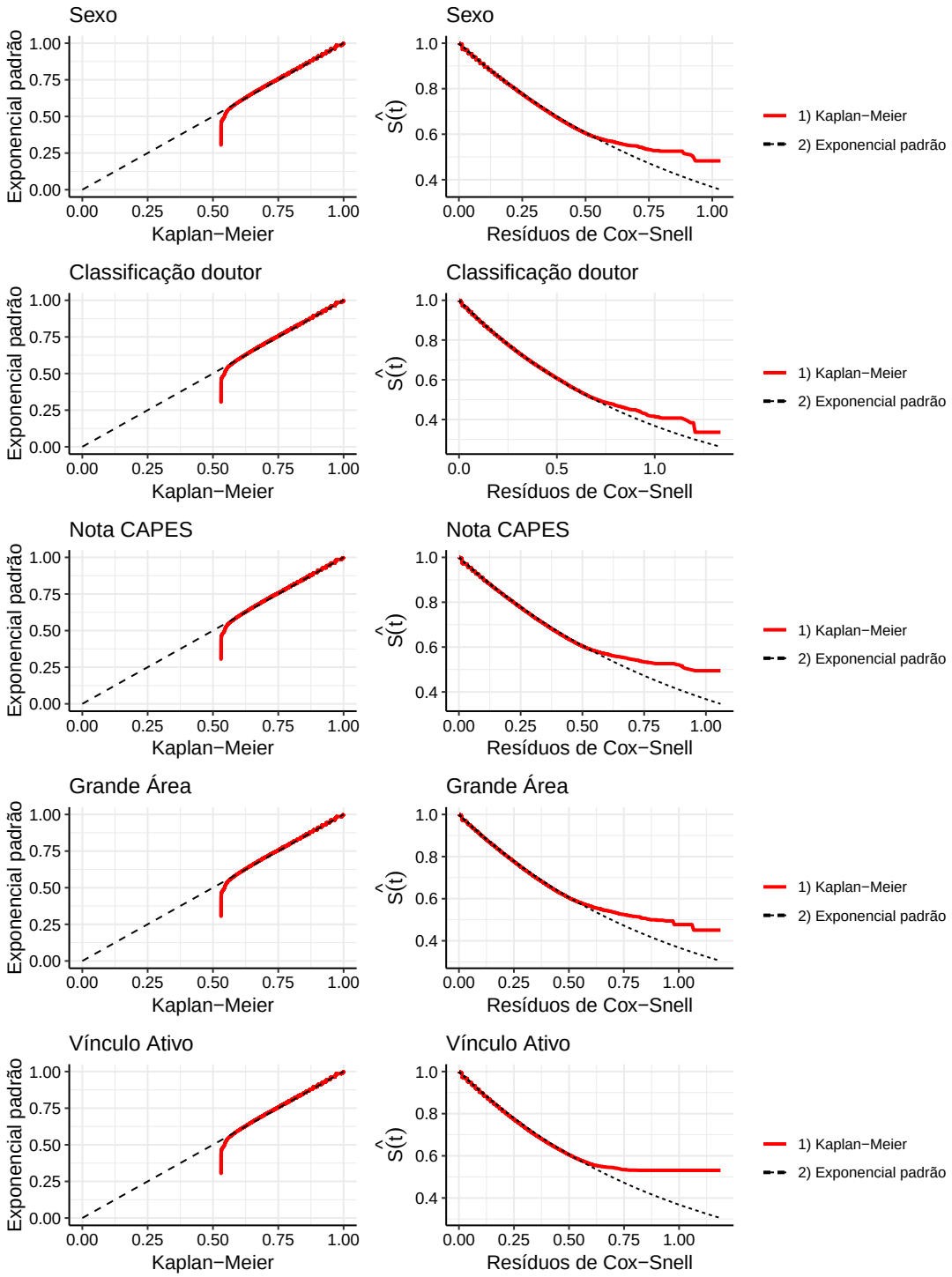

Figura 4.2.3: Comparação das funções de sobrevivência dos resíduos de Cox-Snell do modelo de Cox simples com covariável dependente no tempo para a variável vínculo ativo e Modelo de riscos proporcionais de Cox para as demais covariáveis fixas.

5 Considerações Finais

Este trabalho teve como objetivo avaliar o tempo até a aquisição de um vínculo empregatício formal de caráter acadêmico de doutores que titularam no Brasil por meio do modelo de regressão de Cox com covariáveis dependentes no tempo. Com base nos resultados obtidos, pode-se concluir que o modelo apresentou bom ajuste aos dados, possibilitando a indicação de fatores que podem influenciar o tempo até a aquisição do vínculo empregatício.

Diferente do esperado, os riscos foram menores conforme maior fosse a nota Capes do programa que que o indivíduo obteve o título de doutor, indicando que quanto maior a nota do programa, maior o tempo até a aquisição do vínculo empregatício. Uma hipótese desse fenômeno é a maior seletividade em relação ao local onde o individuo obteve o vínculo. É razoável supor que aqueles que titularam em programas com maiores notas também buscassem um vínculo em instituições de maior prestígio, dificultando (e consequentemente aumentando o seu tempo) a obtenção do vínculo.

Neste trabalho, o vínculo ativo (vínculo fora da área acadêmica) é uma covariável que se modifica ao longo do tempo, justificando o uso do modelo com covariáveis dependentes no tempo. Os resultados mostraram que doutores que não possuíam vínculo ativo apresentaram maior risco (isto é, menor tempo) para a aquisição do vínculo ligado à atividades acadêmicas. Este foi um resultado esperado, visto que é natural entender que àqueles que já possuem um vínculo tem uma fonte de renda e, portanto, tem uma urgência menor em obter um vinculo empregatício na área acadêmica, podendo assim ser mais seletivo em relação ao local onde irá obter o vínculo. Os resultados também mostraram que aqueles que são considerados jovem doutor (obtiveram o título de doutor até os 32 anos) apresentaram maior risco (menor tempo), quando comparados àqueles que titularam após os 32 anos, que a grande área de Linguística foi a área que apresentou o maior risco (menor tempo) e a área de Ciências Biológicas foi a que apresentou o menor risco (maior tempo) de aquisição do vínculo empregatício na área acadêmica. Ademais, o sexo aparenta não tem relação com o tempo até a aquisição do vinculo formal na área acadêmica, tendo os titulados do sexo masculino apresentado um risco de apenas 2% maior que os titulados do sexo feminino.

Cabe destacar que o grande número de observações na amostra ($n = 98.581$) concedeu poder suficiente para rejeitar facilmente os testes de significância realizados, independente da sua significância prática. Por esse motivo, todas as decisões de julgamento da adequação do modelo e significância das covariáveis foram realizadas conside-

rando técnicas gráficas e o tamanho do efeito das estimativas, ao invés da significância estatística.

Como propostas futuras sugere-se realizar o estudo por meio de um modelo de regressão paramétrico discreto com covariáveis dependentes no tempo. Visto que o tempo medido é em meses, a adoção de um modelo contínuo pode não ser adequado (Nakano e Carrasco (2006) e Biazatti e Nakano (2020)). Ademais, um modelo de sobrevivência com fração de cura também pode ser uma alternativa para modelar esse tipo de dados, uma vez que existe uma proporção significativa de doutores que nunca irão atuar na área acadêmica.

Referências

ARCANJO, P. *Sistema de pós-graduação colhe informação com nova ferramenta*. 2012. Disponível em: ⟨http://portal.mec.gov.br/component/tags/tag/35995⟩. Acesso em 21 de dezembro de 2022.

BIAZATTI, E. C.; NAKANO, E. Y. Uma proposta de orientação para o uso de modelos contínuos em dados de sobrevivência discretos. *REMAT: Revista Eletrônica da Matemática*, v. 6, n. 2, p. e4002–e4002, 2020.

BRASIL. Decreto nº 10.854, de 10 de novembro de 2021. *Diário Oficial [da] República Federativa do Brasil*, Brasília, DF, 2021. Disponível em: ⟨https://www.planalto.gov.br/ccivil_03/_Ato2019-2022/2021/Decreto/D10854.htm⟩.

BRESLOW, N. E. Analysis of survival data under the proportional hazards model. *International Statistical Review/Revue Internationale de Statistique*, JSTOR, p. 45–57, 1975.

CGEE. *Estudo sobre os Doutores Titulados no Exterior: expansão da base de doutores no exterior e novas análises (1970 – 2014)*. 2015. Disponível em: ⟨https://www.cgee.org.br/documents/10195/734063/doutores_no_exterior_relatorio_final.pdf⟩. Acesso em 21 de dezembro de 2022.

CGEE. *BRASIL: Mestres e Doutores 2019*. 2019. Disponível em: ⟨https://mestresdoutores2019.cgee.org.br/web/guest/inicio⟩. Acesso em 21 de dezembro de 2022.

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

COX, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 34, n. 2, p. 187–202, 1972.

COX, D. R. Partial likelihood. *Biometrika*, Oxford University Press, v. 62, n. 2, p. 269–276, 1975.

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.

KALBFLEISCH, J. D.; PRENTICE, R. L. *The statistical analysis of failure time data*. [S.l.]: John Wiley & Sons, 2011.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, Taylor & Francis, v. 53, n. 282, p. 457–481, 1958.

MEYER, P. L. *Probabilidade: aplicações à estatística*. [S.l.]: Livros Técnicos e Científicos Rio de Janeiro, 1983.

MTE. *O que é CAGED?* 2016. Disponível em: ⟨http://pdet.mte.gov.br/o-que-e-caged⟩. Acesso em 03 de Julho de 2023.

MTE. *O QUE É RAIS?* 2021. Disponível em: ⟨http://www.rais.gov.br/sitio/sobre.jsf⟩. Acesso em 21 de dezembro de 2022.

NAKANO, E. Y. Um curso de análise de sobrevivência. Departamento de Estatística, Universidade de Brasília, 2017.

NAKANO, E. Y.; CARRASCO, C. G. Uma avaliação do uso de um modelo contínuo na análise de dados discretos de sobrevivência. *Trends in Computational and Applied Mathematics*, v. 7, n. 1, p. 91–100, 2006.

PARREIRA, D. R. M. Um modelo de risco proporcional dependente do tempo. Universidade Federal de São Carlos, 2007.

SANTOS, R. d. O. Formação de doutores para atividades de caráter acadêmico via modelo de riscos proporcionais de cox e regressão logística. 2017.

SANTOS, R. de O.; NAKANO, E. Y. Análise do tempo de permanência de trabalhadores no mercado de trabalho do distrito federal via modelo de riscos proporcionais de cox e log-normal. *Rev. Bras. Biom*, v. 33, n. 4, p. 570–584, 2015.

THERNEAU, T. M.; GRAMBSCH, P. M. *The cox model*. [S.l.]: Springer, 2000.