Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Dissertação de Mestrado

Formação de doutores para atividades de caráter acadêmico via Modelo de Riscos Proporcionais de Cox e Regressão Logística

por

Rayany de Oliveira Santos

Orientador: Prof. Eduardo Yoshio Nakano

Brasília

2017

Rayany de Oliveira Santos

Formação de doutores para atividades de caráter acadêmico via Modelo de Riscos Proporcionais de Cox e Regressão Logística

Dissertação apresentado ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília como requisito parcial à obtenção do título de Mestre em Estatística.

Orientador: Prof. Eduardo Yoshio Nakano

Universidade de Brasília

Brasília, 2017

Agradecimentos

Agradeço a Deus por permitir que meus maiores desejos se tornem realidade: a construção da minha família, a continuidade aos estudos e a formação de uma carreira sólida.

A minha família pelo amor incondicional, pelo apoio, suporte e incentivo que dispõem sempre que me proponho a superar um novo desafio, em especial a meus pais e irmã.

Ao meu esposo pelo companheirismo e dedicação no dia-a-dia e pelos conselhos que tanto me fizeram querer continuar quando a rotina se mostrou cansativa. Por compartilhar comigo a emoção de aumentar nossa pequena família.

Ao Professor Eduardo Nakano por ter me orientado mais uma vez em um trabalho de conclusão de curso, pela confiança, pela disposição em compartilhar seu conhecimento e por fazer das coisas o mais descomplicadas possível.

Ao Centro de Gestão e Estudos Estratégicos - CGEE, empresa onde trabalho, pelo apoio e suporte para que fosse possível a finalização do curso de mestrado.

Resumo

Formação de doutores para atividades de caráter acadêmico via Modelo de Riscos Proporcionais de Cox e Regressão Logística

Neste trabalho, o modelo de regressão de Cox e o modelo de regressão logística foram aplicados para analisar dados relacionados ao emprego de doutores titulados no Brasil. O objetivo foi estimar, através dos dados do Coleta Capes e da Plataforma Sucupira, disponibilizados pela Capes, e dos dados da Relação Anual de Informações Sociais - RAIS, disponibilizados pelo Ministério do Trabalho e Emprego, o tempo até os doutores, após a obtenção do título, obterem um vínculo formal de emprego cuja atividade principal possua natureza acadêmica. Ambos os modelos considerados apresentaram ajustes adequados para o conjunto de dados considerado, além de produzirem estimativas similares para um mesmo perfil de indivíduo em um certo tempo considerado.

Palavras-chave: Análise de sobrevivência; mestres e doutores; modelos de regressão; Coleta Capes; Plataforma Sucupira; RAIS.

Abstract

PhD training for academic activities via Cox proportional hazard and Logistic Regression

In this work, we propose to analyse survival data from formal labor market of PhD graduated in Brazil by the Cox proportional hazard and Logistic Regression models. The objective was to obtain a model that estimates the time to an individual with a PhD get an academic job. Both models presented an appropriate adjustment for Coleta Capes, Plataforma Sucupira and Relação Anual de Informações Sociais - RAIS data sets.

Keywords: Survival analysis; PhD; regression models ; Coleta Capes; Plataforma Sucupira; RAIS.

Sumário

Introdução **3**

1 Revisão de Literatura **5**

1.1 Conceitos Básicos de Análise de sobrevivência 5

1.1.1 Perda da Informação Temporal 5

1.1.2 Descrição do Tempo de Sobrevivência 8

1.1.3 Estimador de Kaplan-Meier 11

1.2 Modelo de Regressão de Cox 12

1.2.1 Estimação dos Parâmetros 13

1.2.2 Funções relacionadas a $h_0(t)$ 15

1.2.3 Interpretação dos Parâmetros no Modelo de Cox 16

1.2.4 Adequação do Modelo de Cox 16

1.3 Regressão Logística 19

1.3.1 Estimação dos Coeficientes 21

1.3.2 Interpretação dos Coeficientes 23

1.3.3 Avaliação do ajuste do Modelo 23

1.3.4 Avaliação do Modelo 24

1.4 RH para Ciência, Tecnologia e Inovação 25

1.4.1 Doutores 2010: estudos da demografia da base técnico-científica brasileira 25

1.4.2 Mestres 2012: estudos da demografia da base técnico-científica brasileira 26

1.4.3 Mestres e Doutores 2015: estudos da demografia da base técnico-científica brasileira 27

1.4.4 A Formação de Novos Quadros para CT&I: Avaliação do Programa Institucional de Bolsas de Iniciação Científica - Pibic . 27

2 Análise de dados **29**

2.1 Bases de dados 29

2.2 Tratamento da base de dados 31

2.2.1 Validação de CPF 31

2.2.2 Seleção de egressos titulados entre 2010 e 2013 32

2.2.3 Seleção do título mais recente 34

2.2.4 Seleção de vínculos relacionados a atividades acadêmicas com data de admissão mais próxima a da titulação 34

3 Resultados **37**

3.1 Análise descritiva dos dados 37

3.2 Modelo de regressão de Cox 40

3.3 Modelo de regressão logística 50

4 Conclusão **55**

Referências Bibliográficas **57**

Introdução

A pós-graduação brasileira, que é objeto de uma das políticas públicas mais consistentes e duradouras do país e passa por intenso processo de crescimento, diversificação e amadurecimento, já atingiu uma escala e um padrão de qualidade que a distingue entre as nações emergentes. A existência desses recursos humanos qualificados é essencial para o aumento das vantagens competitivas de base tecnológica, porque tais vantagens dependem de nossa capacidade de absorver, transformar e produzir novos conhecimentos e inovação. Essa parcela específica da população tem papel fundamental, em especial, na formação de doutores, que são profissionais com capacidade para realizar pesquisa e desenvolvimento (P&D) originais. (CGEE, 2010)

Apesar do alto e significativo patamar atingido pela formação de doutores nos últimos anos, é de fundamental importância a contínua expansão e desenvolvimento da qualidade do ensino dos doutores brasileiros de forma a melhor contribuir para o enfrentamento do desafio de produzir conhecimentos e inovações necessários ao avanço do processo de desenvolvimento sustentável brasileiro.

Nesse contexto, o presente trabalho objetiva analisar, primeiramente, por meio do modelo de riscos proporcionais de Cox, o tempo que indivíduos levam entre a obtenção do título de doutorado no Brasil e a obtenção de um vínculo empregatício formal em um estabelecimento cuja atividade econômica principal esteja classificada como "Educação"ou "Atividades profissionais, científicas e técnicas", que são os estabelecimentos que concentram grande parte dos mestres e doutores (CGEE, 2015). Pretende-se ainda identificar quais fatores podem ter influência sobre o tempo de sobrevivência. O foco principal é o percurso formativo e profissional para atividades aqui consideradas de caráter acadêmico, que são ligadas a docência e pesquisa. Outros trabalhos que apresentam aplicações do modelo de riscos proporcionais de Cox

podem ser vistos em Nakano e Cunha (2012), Santos e Nakano (2015), Maia e Nakano (2016) e Silva et al. (2017).

A análise de sobrevivência é uma das áreas da estatística que mais cresceu nas últimas duas décadas do século passado, segundo Colosimo e Giolo (2006), devido a combinação do desenvolvimento e aprimoramento de técnicas estatísticas com computadores cada vez mais velozes. A variável resposta do estudo é, geralmente, o tempo até a ocorrência de um evento de interesse, denominado tempo de falha e a principal característica dos dados é a presença de censura que se refere a situações em que o acompanhamento do indivíduo foi interrompido por alguma razão.

O modelo de Cox permite a análise de dados provenientes de tempo de vida com a presença de covariáveis em um contexto semi paramétrico. Ele é denominado modelo de taxas de falha proporcionais devido a razão das taxas de falha de dois indivíduos diferentes ser constante ao longo do tempo.

Ademais, outro objetivo do trabalho é estabelecer a chance de um doutor também estar empregado em um estabelecimento cuja atividade econômica principal esteja classificada como "Educação"ou "Atividades profissionais, científicas e técnicas", um ano após a obtenção do título de doutorado no Brasil, por meio de um modelo de regressão logística.

Essa técnica tem a finalidade de gerar uma função cuja resposta permita estabelecer a probabilidade de uma observação pertencer a um grupo previamente determinado, em razão do comportamento de um conjunto das variáveis independentes. Aqui, a função *logit* é utilizada como função de ligação e a medida de associação calculada a partir do modelo logístico recebe o nome de razão de chances (*odds ratio*), que é obtida por meio da comparação de indivíduos que diferem apenas na característica de interesse e que tenham os valores das outras variáveis constantes.

Os dados utilizados no estudo são provenientes das bases do Coleta Capes e da Plataforma Sucupira, fornecidas pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) do Ministério da Educação (MEC), e das bases da Relação Anual de Informações Sociais (RAIS) do Ministério do Trabalho e Emprego (MTE). Todas as análises serão realizadas por meio do *software* livre R (R Core Team, 2016).

Capítulo 1

Revisão de Literatura

1.1 Conceitos Básicos de Análise de sobrevivência

A ciência estatística possui uma área designada análise de sobrevivência que compreende modelos e técnicas destinados à análise de dados de sobrevivência, que são resultado da observação do tempo transcorrido até a ocorrência de um evento de interesse, geralmente a morte de um indivíduo ou a falha de um equipamento. Esse tempo é denominado tempo de falha. Por possuir a flexibilidade de ser aplicada em diversas áreas de estudo, como a medicina, engenharia e demografia, a análise de sobrevivência vem tomando posição de destaque nas últimas décadas em todo o mundo.

A resposta desse tipo de estudo é caracterizada pelas censuras e pelos tempos de falha. O instante em que os indivíduos começam a fazer parte do estudo varia quando as coortes são abertas. (Colosimo e Giolo, 2006)

Neste capítulo, alguns conceitos básicos e técnicas para analisar dados de sobrevivência serão abordados.

1.1.1 Perda da Informação Temporal

Geralmente, em estudos de longa duração, é comum a perda do acompanhamento de alguns indivíduos durante o passar do tempo, visto que estes podem não vir a falhar devido, por exemplo, ao óbito por causas não relacionadas ao estudo, ou não é possível saber se o evento de interesse ocorreu, devido o término do estudo, desistência por parte do indivíduo, entre outras causas. Outra situação frequentemente observada

é a exclusão de certos indivíduos do estudo.

Truncamento

O truncamento é caracterizado pela exclusão de alguns indivíduos que pertenciam naturalmente à população estudada por motivo relacionado a ocorrência do evento de interesse. Eles não são acompanhados a partir do tempo inicial, apenas a partir do momento que experimentam um certo evento. Um exemplo dessa situação acontece quando apenas uma amostra de indivíduos de uma população é utilizada para a realização do estudo por possuírem um certa característica derivada de um evento, como quando apenas os aposentados de uma comunidade são observados para se estimar a distribuição do tempo de vida dos moradores.

Censura

A presença de censura é a principal característica de dados de sobrevivência e ocorre quando o evento de interesse não é observado para algum indivíduo durante o período de realização do estudo, decorrendo em observações incompletas. Ainda assim, os dados censurados devem ser incluídos na análise pois eles fornecem informações sobre o tempo de vida de indivíduos e a omissão deles pode acarretar conclusões viciadas.

Alguns mecanismos de censura podem ser considerados, visto que são diversos os motivos para que ela aconteça. São eles: censura à esquerda, censura intervalar e censura à direita.

A censura à esquerda é caracterizada pelo evento de interesse já ter ocorrido quando o indivíduo começou a fazer parte do estudo, ou seja, o tempo registrado é maior que o tempo de falha. Um exemplo de situação que envolve censura à esquerda é um estudo que tem a finalidade de determinar a idade em que certas crianças aprenderam a ler. As observações censuradas são caracterizadas pelas crianças que já sabiam ler e não lembravam com que idade isto tinha acontecido. (Colosimo e Giolo, 2006)

A censura intervalar ocorre quando os indivíduos são acompanhados periodicamente e o evento de interesse acontece em um intervalo de tempo. Logo, tempo de

falha não é conhecido exatamente mas pertence a esse intervalo.

A censura à direita ocorre quando o tempo de ocorrência do evento de interesse está à direita do tempo registrado. Ela pode ser classificada como:

1. Censura Tipo I: É caracterizada pela presença de uma ou mais observações que não apresentaram o evento de interesse após um período pré-estabelecido de tempo.

   A Figura 1.1 ilustra a situação em que alguns indivíduos não experimentaram o evento até o final do estudo. A falha é representada por • e a censura por ∘. É importante observar que o tempo $t = 20$ é fixo.

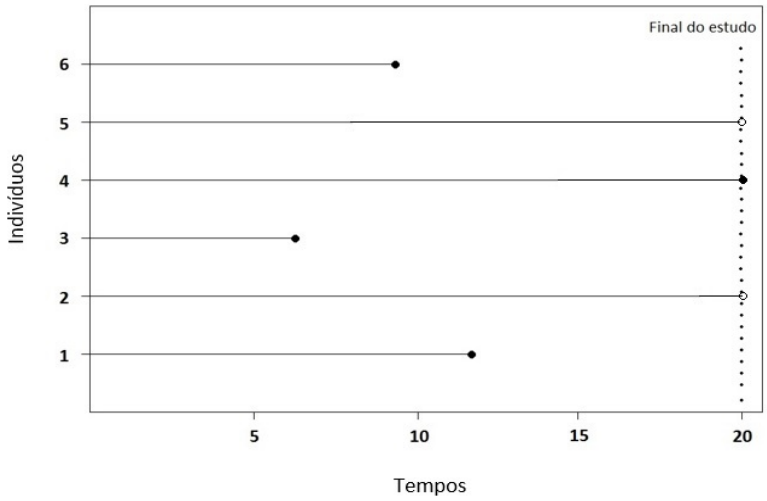

Figura 1.1: Dados com censura tipo I.

2. Censura Tipo II: É resultado de estudos que são finalizados após a ocorrência do evento de interesse em um número pré-estabelecido de indivíduos.

   A Figura 1.2 ilustra o mecanismo de censura à direita do tipo II. Nesse caso, o número de falhas é fixo, ou seja, o estudo foi finalizado após a ocorrência de 4 falhas, já estabelecidas anteriormente. A falha é representada por • e a censura por ∘.

3. Censura aleatória: Ocorre quando um indivíduo é retirado durante a realização do estudo sem que a falha tenha acontecido, quando ele morre por uma razão qualquer, diferente da estudada ou quando o evento de interesse não foi observado até o fim do estudo.

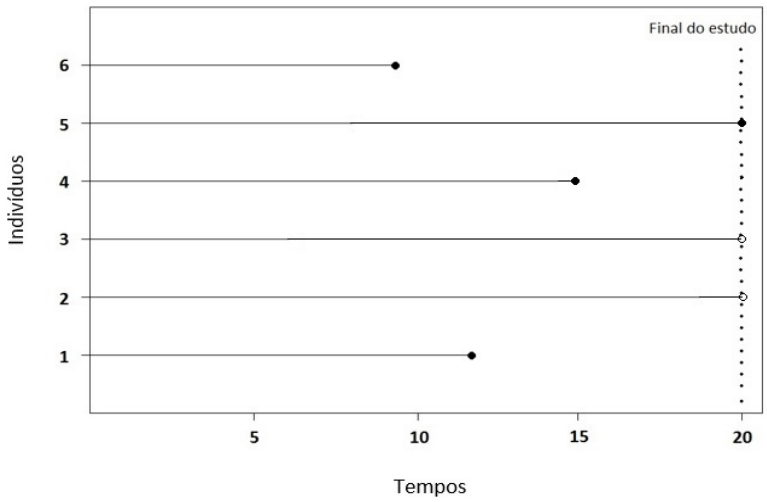

Figura 1.2: Dados com censura tipo II.

A Figura 1.3 ilustra a censura a direita aleatória. A falha é representada por • e a censura por ∘.

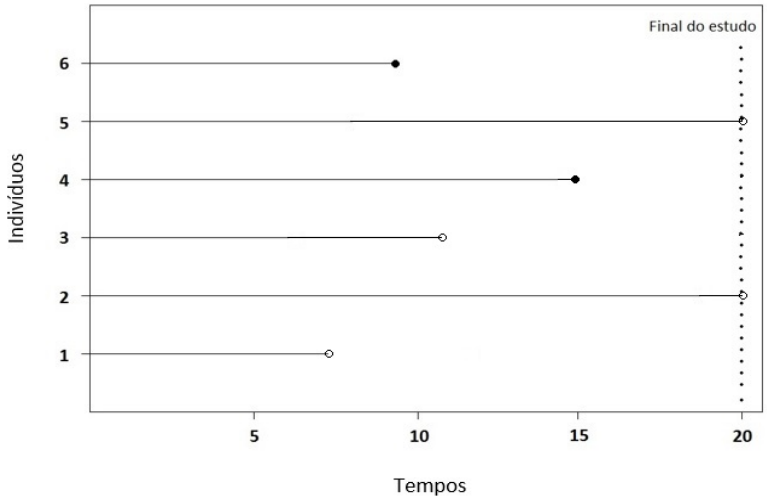

Figura 1.3: Dados com censura aleatória.

1.1.2 Descrição do Tempo de Sobrevivência

O tempo de vida do indivíduo, conhecido como tempo de sobrevivência é representado pela variável aleatória não-negativa T, geralmente contínua. Ela pode ser especificada pela função densidade de probabilidade, $f(t)$; pela função de sobrevivência, $S(t)$; pela função de falha, $h(t)$; e por relações existentes entre essas funções. Estudos que consideram o tempo de sobrevivência discreto podem ser vistos em Na-

kano e Carrasco (2006), Carrasco et al. (2012) e Brunello e Nakano (2015).

O tempo de sobrevivência, T, é dado pela expressão 1.1 a seguir:

$$T = T_F - T_I, \tag{1.1}$$

em que T_F é o momento em que o indivíduo experimentou o evento de interesse ou foi censurado e T_I é o momento em que o indivíduo deu entrada no estudo.

A variável indicadora de falha ou censura deve ser incluída no estudo para fins da análise e é expressa por:

$$\delta_i = \begin{cases} 1, & \text{se o } i\text{-ésimo indivíduo falhou} \\ 0, & \text{se o } i\text{-ésimo indivíduo foi censurado} \end{cases}$$

A variável δ_i representa, juntamente com o tempo de falha t_i, os dados de sobrevivência para o indivíduo i $(i = 1, ..., n)$. Na presença de um vetor de covariáveis $\mathbf{x}_i$, $i = 1, ..., n$, os dados de sobrevivência são representados por $(t_i, \delta_i, \mathbf{x}_i)$.

Função de Densidade de Probabilidade

A variável aleatória T será considerada contínua se existir uma função f, denominada *função densidade* que satisfaz as seguintes condições (Magalhães, 2006):

(C1) $f(t) \geq 0, \forall t \in \mathbb{R}$;

(C2) $\int_{-\infty}^{\infty} f(w)dw = 1$.

Função Distribuição

O conhecimento da função de distribuição de uma variável aleatória permite que qualquer informação sobre esta seja obtida. Ela também é conhecida como função de distribuição acumulada por acumular as probabilidades dos valores inferiores ou iguais a t (Magalhães, 2006).

A função de distribuição da variável aleatória T é definida por:

$$F(t) = P(T \in (-\infty, t]) = P(T \leq t), \tag{1.2}$$

com t percorrendo todos os reais. $F_T(t)$ possui as seguintes propriedades:

(P1) $\lim_{t\to-\infty} F(t) = 0$ e $\lim_{t\to\infty} F(t) = 1$;

(P2) F é contínua à direita;

(P3) F é não decrescente, isto é, $F(t) \leq F(y)$ sempre que $t \leq y$, $\forall\ t, y \in \mathbb{R}$.

Para uma variável aleatória T não negativa, a função distribuição acumulada representa a probabilidade de uma observação não sobreviver ao tempo t, ou seja, $F(t) = 1 - S(t)$, em que $S(t)$ representa a função de sobrevivência, descrita a seguir.

Função de Sobrevivência

A função de sobrevivência é a probabilidade de uma observação sobreviver ao tempo t, ou seja, a probabilidade de um indivíduo não falhar até um certo tempo t. Ela é definida na equação 1.3 (Colosimo e Giolo, 2006):

$$S(t) = P(T \geq t). \tag{1.3}$$

Função Taxa de Falha

A função taxa de falha é também chamada função de risco e representa a taxa de falha instantânea no tempo t condicional à sobrevivência até o tempo t. (Colosimo e Giolo, 2006)

Considerando-se o intervalo $[t, t+\Delta t)$ e assumindo Δt pequeno, a função é definida como:

$$h(t) = \lim_{\Delta t\to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{1.4}$$

A função $h(t)$ pode assumir a forma crescente, constante ou decrescente quando a taxa de falha de um indivíduo aumenta, não se altera ou diminui com o passar do tempo, respectivamente. Pode também assumir a forma unimodal ou a forma de curva da banheira.

A função Taxa de Falha Acumulada é útil na avaliação da função taxa de falha quando esta é difícil de ser estimada através da estimação não paramétrica. Ela é dada por:

$$H(t) = \int_0^t h(u)du. \tag{1.5}$$

O conhecimento de qualquer uma das funções descritas acima implica no conheci-

mento das demais. Isso pode ser mostrado pelas seguintes relações (Colosimo e Giolo, 2006):

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

$$H(t) = \int_0^t h(u)du = -\log S(t)$$

e

$$S(t) = \exp\{-H(t)\} = \exp\left\{-\int_0^t h(u)du\right\}.$$

1.1.3 Estimador de Kaplan-Meier

Convencionalmente, a análise estatística descritiva de um estudo consiste na descrição dos dados, que envolve média, desvio-padrão e técnicas gráficas. No entanto, a presença de censuras é um problema para essas técnicas, pois há um aumento no nível de dificuldade para a interpretação de seus resultados e as censuras dificultam a tentativa de encontrar medidas de tendência central e variabilidade. Assim, o principal componente da análise envolvendo dados de sobrevivência é a própria função de sobrevivência, que pode ser estimada pelo conhecido estimador não-paramétrico de Kaplan-Meier (Kaplan e Meier, 1958) quando há censuras.

O estimador de Kaplan-Meier (Kaplan e Meier, 1958), na sua construção, considera tantos intervalos quantos forem o número de falhas distintas. Assumindo:

- $t_1 < t_2 < ... < t_k$, os k tempos distintos e ordenados de falha,
- d_j o número de falhas em t_j, $j = 1, ..., k$, e
- n_j o número de indivíduos sob risco em t_j, ou seja, os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a t_j.

O estimador é, então definido como: (Colosimo e Giolo, 2006)

$$\hat{S}(t) = \prod_{j:t_j<t}\left(\frac{n_j - d_j}{n_j}\right) = \prod_{j:t_j<t}\left(1 - \frac{d_j}{n_j}\right). \tag{1.6}$$

Ele possui as seguintes propriedades:

1. é não viciado para amostras grandes;

2. é fracamente consistente;

3. converge assintoticamente para um processo gaussiano; e

4. é estimador de máxima verossimilhança de $S(t)$.

Um intervalo aproximado de $100(1-\alpha)\%$ de confiança para $S(t)$ é dado por:

$$\hat{S}(t) \pm z_{\alpha/2}\sqrt{\hat{Var}(\hat{S}(t))}, \tag{1.7}$$

em que

$$\hat{Var}(\hat{S}(t)) = \left[\hat{S}(t)\right]^2 \sum_{j:t_j<t} \frac{d_j}{n_j(n_j-d_j)}.$$

Aqui $z_{\alpha/2}$ é o quantil $\alpha/2$ de uma distribuição normal padrão.

1.2 Modelo de Regressão de Cox

Segundo Colosimo e Giolo (2006), o modelo de Cox permite a análise de dados provenientes de tempo de vida com a presença de covariáveis em um contexto não paramétrico.

Considerando primeiramente um estudo em que existe apenas uma covariável e que tem o objetivo de comparar os tempos de falha de dois grupos em que os indivíduos são selecionados para fazer parte do grupo 0 ou do grupo 1, temos:

$$\frac{h_1(t)}{h_0(t)} = K. \tag{1.8}$$

Aqui $h_0(t)$ é a função de risco do grupo 0, $h_1(t)$ é a função de risco do grupo 1 e K é a razão das taxas de falha, constante para todo tempo t.

Assumindo que x é a variável indicadora de grupo, em que

$$x = \begin{cases} 0, & \text{se grupo 0} \\ 1, & \text{se grupo 1} \end{cases}$$

e $K = \exp\{\beta x\}$, temos o seguinte modelo de Cox para uma única covariável:

$$h(t|x) = h_0(t) \exp\{\beta x\} \tag{1.9}$$

Agora, considerando p covariáveis, de modo que $\mathbf{x} = (x_1, ..., x_p)'$ é um vetor, a expressão geral do modelo de regressão de Cox é dada por (Cox, 1972):

$$h(t|\mathbf{x}) = h_0(t) g(\mathbf{x}'\boldsymbol{\beta}), \tag{1.10}$$

em que $g(\mathbf{x}'\boldsymbol{\beta})$ é uma função não-negativa que deve ser especificada de forma que $g(0) = 1$, geralmente dada por:

$$g(\mathbf{x}'\boldsymbol{\beta}) = \exp\{\mathbf{x}'\boldsymbol{\beta}\} = \exp\{\beta_1 x_1 + ... + \beta_p x_p\} \tag{1.11}$$

Esse modelo é denominado modelo de taxas de falha proporcionais devido a razão das taxas de falha de dois indivíduos diferentes ser constante ao longo do tempo. O modelo de riscos proporcionais de Cox é dito ser um modelo semi-paramétrico pois é composto pelo produto de dois componentes:

- Componente não-paramétrico: função de taxa de falha de base, h_0, que não é especificada;
- Componente paramétrico: $g(\mathbf{x}'\boldsymbol{\beta})$.

Note que o modelo não possui o intercepto β_0 pois o mesmo é absorvido pela constante de proporcionalidade.

1.2.1 Estimação dos Parâmetros

Para a estimação dos parâmetros do modelo, o método de maxima verossimilhança (Colosimo e Giolo, 2006) é inapropriado devido a presença do componente não-paramétrico, $h_0(t)$, na função de verossimilhança. Assim, o método de verossimilhança parcial foi proposto por Cox para condicionar a construção da função de verossimilhança ao conhecimento da história passada de falhas e censuras para eliminar a função de risco base.

Dada uma amostra de n indivíduos com $k \leq n$ falhas distintas nos tempos $t_1 < t_2 ... < t_k$, o conceito de verossimilhança considera o argumento de que a probabilidade condicional da i-ésima observação vir a falhar no tempo t_i conhecendo quais observações estão sob risco em t_i é:

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

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

$$\frac{h_i(t \mid \mathbf{x}_i)}{\sum_{j \in R(t_i)} h_j(t \mid \mathbf{x}_j)} = \frac{h_0(t)\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{\sum_{j \in R(t_i)} h_0(t)\exp\{\mathbf{x}_j'\boldsymbol{\beta}\}} = \frac{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{\sum_{j \in R(t_i)} \exp\{\mathbf{x}_j'\boldsymbol{\beta}\}},$$

em que $R(t_i)$ é o conjunto dos índices das observações sob risco no tempo t_i.

Assim, a função de verossimilhança parcial é dada por:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{\sum_{j \in R(t_i)} \exp\{\mathbf{x}_j'\boldsymbol{\beta}\}} = \prod_{i=1}^{n} \left(\frac{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{\sum_{j \in R(t_i)} \exp\{\mathbf{x}_j'\boldsymbol{\beta}\}} \right)^{\delta_i}, \tag{1.12}$$

em que δ_i é o indicador de falha. Os valores de $\boldsymbol{\beta}$ que maximizam $L(\boldsymbol{\beta})$ são obtidos a partir de $U(\boldsymbol{\beta}) = 0$, que representa o vetor escore de derivadas de primeira ordem da função $l(\boldsymbol{\beta}) = \log(L(\boldsymbol{\beta}))$. Isto é,

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \log \left[x_i - \frac{\sum_{j \in R(t_i)} x_j \exp\{\mathbf{x}_i'\hat{\boldsymbol{\beta}}\}}{\sum_{j \in R(t_i)} \exp\{\mathbf{x}_j'\hat{\boldsymbol{\beta}}\}} \right] = 0. \tag{1.13}$$

A função 1.13 assume que os tempos de sobrevivência são contínuos e não pressupõe a possibilidade de empates nos valores observados. Com isso, a função de verossimilhança parcial foi aproximada por Efron (1977) e é dada por:

$$PL_E(\boldsymbol{\beta}) = \prod_{k=1}^{D} \frac{\sum_{t_i = t_k^*} \exp(\boldsymbol{\beta}' x_i)}{\prod_{j=1}^{d_k} [\sum_{l \in R_k} \exp(\boldsymbol{\beta}' x_l) - \frac{j-1}{d_k} \sum_{t_i = t_k^*} \exp(\boldsymbol{\beta}' x_i)]^{d_k}}, \tag{1.14}$$

em que d_k é o número de falhas no tempo t_k^*, com $k = 1, 2, ..., D$, t_k^* é o tempo de falha do indivíduo k. (Matuda, 2005)

Existem outras propostas de aproximação, como a de Breslow e Peto que é muito utilizada em estudos estatísticos. No entanto, uma desvantagem encontrada é que esta aproximação é adequada somente quando o número de observações empatadas

em qualquer tempo não é grande. A aproximação de Efron, no entanto produz boas estimativas nessas situações e apenas não é tão utilizada como a de Breslow e Peto por requerer mais tempo e esforço computacional.

A expressão 1.15 mostra um intervalo de $(1-\alpha)$ de confiança para β_i, $i = 1, \ldots, k$, utilizado para se fazer inferências sobre os parâmetros do modelo de Cox, após garantida uma boa estimação de máxima verossimilhança que acarrete pouco ou nenhum viés aos mesmos. O intervalo é definido por:

$$\beta_i \pm z_{1-\alpha/2} * (Var(\hat{\beta}_i))^{1/2} \tag{1.15}$$

em que $Var(\hat{\beta}_i)$ é o elemento diagonal correspondente ao parâmetro de β_i da matriz de informação de Fisher observada e $z_{1-\alpha/2}$ é o quantil $(\frac{1-\alpha}{2} * 100)\%$ da distribuição Normal padrão.

1.2.2 Funções relacionadas a $h_0(t)$

No modelo de Cox, as funções relacionadas a função de risco base são importantes. A função de sobrevivência base é dada por (Colosimo e Giolo, 2006):

$$S_0(t) = \exp\{-H_0(t)\}, \tag{1.16}$$

em que $H_0(t)$ é a função de risco acumulada base.

A função de sobrevivência para um conjunto de covariáveis $\mathbf{x}$ é dada por:

$$S(t|x) = [S_0(t)]^{\exp\{\mathbf{x}'\boldsymbol{\beta}\}}. \tag{1.17}$$

Como o método de máxima verossimilhança parcial elimina $h_0(t)$, os estimadores das funções descritas acima são de natureza não-paramétrica. Uma estimativa simples para $H_0(t)$, proposta por Breslow (1972), é expressa por:

$$\hat{H}_0(t) = \sum_{j:t_j<t} \frac{d_j}{\sum_{l \in R_j} \exp\{\mathbf{x}_l'\hat{\boldsymbol{\beta}}\}}, \tag{1.18}$$

em que d_j é o número de falhas em t_j e $\hat{\boldsymbol{\beta}}$ são os estimadores de $\boldsymbol{\beta}$ obtidos pela

verossimilhança parcial.

Assim, a estimativa da função $\hat{S}(t|x)$ é expressa por:

$$\hat{S}(t|x) = [\hat{S}_0(t)]^{\exp\{x'\hat{\beta}\}}, \tag{1.19}$$

em que $\hat{S}_0(t)$ é a função que estima a função de sobrevivência de base que é dada por:

$$\hat{S}_0(t) = \exp\{-\hat{H}_0(t)\}. \tag{1.20}$$

1.2.3 Interpretação dos Parâmetros no Modelo de Cox

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

Para interpretar os coeficientes estimados, é utilizada a propriedade de riscos proporcionais. Para dois indivíduos (i e l) que apresentam os mesmos valores para as covariáveis, exceto para a p-ésima delas, a taxa de falha é dada por:

$$\frac{h(t|\mathbf{x}_i)}{h(t|\mathbf{x}_l)} = \frac{\exp\{\beta_p x_{ip}\}}{\exp\{\beta_p x_{lp}\}} = \exp\{\beta_p(x_{ip} - x_{lp})\} \tag{1.21}$$

A expressão 1.21 é, na realidade, uma razão de riscos e, assim, se, por exemplo, x_p é a covariável dicotômica referente ao sexo do indivíduo, em que $x_{ip} = 1$ (masculino) e $x_{lp} = 0$ (feminino), tem-se que o risco de falha dos indivíduos do sexo masculino é $\exp(\beta_p)$ vezes o risco de falha dos indivíduos do sexo feminino, mantendo-se fixas as demais covariáveis.

1.2.4 Adequação do Modelo de Cox

O modelo de regressão de Cox, apesar de ser bastante flexível devido a presença do componente não-paramétrico, não se ajusta a qualquer situação e, assim como qualquer outro modelo estatístico, requer o uso de técnicas para avaliar sua adequação (Colosimo e Giolo, 2006). Para esse fim, é necessário verificar a suposição de riscos proporcionais, cuja violação pode acarretar sérios vícios na estimação dos coeficientes do modelo, e fazer uma avaliação geral do ajuste do modelo de Cox.

Avaliação da Suposição de Riscos Proporcionais

Encontram-se propostos na literatura algumas técnicas gráficas e testes estatísticos para avaliar a suposição de riscos proporcionais no modelo de Cox. Algumas dessas técnicas serão descritas a seguir.

(a) Método gráfico descritivo

Essa técnica consiste em dividir os dados em m estratos, usualmente de acordo com alguma covariável. Em seguida, deve-se estimar $H_0(t)$, usando a expressão 1.18, para cada estrato da covariável. Se a suposição de riscos proporcionais for pertinente, as curvas do logaritmo de $H_0(t)$ versus t, ou $\log(t)$, apresentarão diferenças aproximadamente constantes ao longo do tempo. Curvas não paralelas significam desvios da suposição de taxas de falha proporcionais.

(b) Método com coeficiente dependente do tempo

A suposição de taxas de falhas proporcionais no modelo de Cox pode ser avaliada através da análise dos resíduos de Schoenfeld (1982). Assim como no caso do método gráfico descritivo, por ser uma técnica gráfica, conclusões subjetivas estão envolvidas durante a interpretação dos gráficos.

Considerando que o i-ésimo indivíduo com vetor de covariáveis $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})'$ venha a falhar, tem-se para este indivíduo um vetor de resíduos de Schoenfeld $\mathbf{r}_i = (r_{i1}, r_{i2}, ..., r_{ip})$ em que cada componente r_{iq}, para $q = 1, ..., p$, é definido por (Colosimo e Giolo, 2006):

$$r_{iq} = x_{iq} - \frac{\sum_{j \in R(t_i)} x_{jq} \exp\{\mathbf{x}_j' \hat{\beta}\}}{\sum_{j \in R(t_i)} \exp\{\mathbf{x}_j' \hat{\beta}\}}. \tag{1.22}$$

Os resíduos são definidos para cada falha e não são definidos para censuras. Para permitir que a estrutura de correlação dos resíduos seja considerada, uma forma padronizada dos resíduos de Schoenfeld é frequentemente usada e é definida por:

$$\mathbf{s}_i^* = [I(\hat{\beta})]^{-1} \mathbf{r}_i, \tag{1.23}$$

com $I(\hat{\beta})$ a matriz de informação observada.

A suposição de riscos proporcionais avaliada pelos resíduos de Schoenfeld é baseada

em um resultado de Grambsch e Therneau (1994), expresso por:

$$h(t|\mathbf{x}) = h_0(t)\exp\{\mathbf{x}'\boldsymbol{\beta}(t)\} \tag{1.24}$$

Tem-se a restrição de que $\boldsymbol{\beta}(t) = \boldsymbol{\beta}$, como uma forma alternativa de representar o modelo de Cox. A restrição $\boldsymbol{\beta}(t) = \boldsymbol{\beta}$ resulta na proporcionalidade dos riscos. Se $\boldsymbol{\beta}(t)$ não é constante, o impacto de uma ou mais covariáveis no risco pode variar com o tempo. Assim, o gráfico de $\boldsymbol{\beta}_q(t)$ versus t deve ser uma linha horizontal. Grambsch e Therneau (1994), sugerem o gráfico de $s^*_{iq} + \hat{\beta}_q$ versus t, para $q = 1, \ldots, p$, ou alguma função do tempo, g(t). Inclinação zero mostra evidências a favor da proporcionalidade (Colosimo e Giolo, 2006). Para auxiliar na detecção de uma possível falha da suposição de riscos proporcionais, uma curva suavizada, com bandas de confiança, é adicionada a esse gráfico.

(c) Teste de hipótese e coeficiente de correlação

O coeficiente de correlação de Pearson (ρ), entre os resíduos padronizados e g(t), para cada covariável, é uma medida que permite avaliar a suposição de riscos proporcionais. Valores de ρ próximos de zero mostram não haver evidências para rejeitar a suposição de riscos proporcionais.

Um teste para a hipótese global de proporcionalidade dos riscos sobre todas as covariáveis no modelo, levando em consideração $g_q(t) = g_t$, tem como estatística do teste:

$$T = \frac{(g - \bar{g})' S^* I S^* (g - \bar{g})}{d\sum_k (g_k - \bar{g})^2} \tag{1.25}$$

no qual, I é a matriz de informação observada, d é o número de falhas e $S^* = dRI^{-1}$, sendo R a matriz dX_p dos resíduos de Schoenfeld não padronizados.

As hipóteses a serem testadas são:

H_0: Os riscos são proporcionais. H_1: Os riscos não são proporcionais.

A estatística do teste T tem distribuição qui-quadrado com p graus de liberdade. Valores de $T > \chi^2_{p,1-\alpha}$ mostram que há evidências estatística para rejeitar a suposição de riscos proporcionais.

Avaliação Geral do Ajuste do Modelo de Cox

É importante avaliar se o modelo proposto está bem ajustado aos dados. Isso pode ser feito por meio de técnicas gráficas, que avaliam a distribuição dos erros e são utilizadas para rejeitar modelos inapropriados, ou seja, o interesse não é aprovar um modelo particular, até porque, em muitos casos, existe mais de um modelo que pode ser utilizado para o mesmo fim.

Segundo Colosimo e Giolo (2006), os resíduos de Cox-Snell são utilizados com o propósito de avaliar a qualidade geral de ajuste do modelo de Cox. Eles são deifinidos por:

$$\hat{e_i} = \hat{H}(t_i), \tag{1.26}$$

em que $\hat{H}(t_i)$ é a função de risco acumulado obtido do modelo ajustado.

Os resíduos $\hat{e_i}$ vêm de uma população homogênea e, se o modelo for adequado, devem seguir uma distribuição exponencial padrão (Lawless,2003). Para que o modelo exponencial seja adequado, o gráfico de $\hat{e_i}$ versus $\hat{H}(\hat{e_i})$ deve ser aproximadamente uma reta. O gráfico das curvas de sobrevivência dos resíduos, $\hat{S}(\hat{e_i})$, pelo modelo exponencial padrão, $\exp(-\hat{e_i})$, também auxiliam na verificação da qualidade do modelo ajustado: quanto mais próximas, melhor o ajuste do modelo aos dados.

1.3 Regressão Logística

A regressão logística, que é um caso particular do Modelo Linear Generalizado (MLG), surgiu e se desenvolveu na área médica por volta de 1960 com o objetivo de realizar predições e estudar a relação entre uma variável aleatória binária (variável dependente) e um conjunto de variáveis independentes, mas se expandiu rapidamente por muitos campos, devido à facilidade e capacidade em explicar a ocorrência de determinados eventos (Machado, 2015).

Sua finalidade é gerar uma função matemática cuja resposta permita estabelecer a probabilidade de uma observação pertencer a um grupo previamente determinado, em razão do comportamento de um conjunto das variáveis independentes. São poucas as suposições necessárias para a aplicação da técnica, que consistem em:

(a) valor esperado igual a zero para os resíduos;

(b) erros não correlacionados;

(c) variáveis independentes;

(d) ausência de multicolinearidade perfeita entre as variáveis explicativas.

Agresti (1990) define MLG como um modelo linear para transformação da esperança de uma variável aleatória cuja distribuição pertence à família exponencial. Segundo McCullagh e Nelder (1989), um MLG é composto por três elementos fundamentais: um componente aleatório, um componente determinístico ou sistemático e uma função de ligação.

O componente aleatório consiste na variável dependente Y que se deseja modelar, da qual se coletam n observações independentes e cuja distribuição de probabilidade deve pertencer à família exponencial. Mais detalhes sobre a família exponencial podem ser encontrados em Casella e Berguer (2010).

O componente determinístico é definido por um vetor $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$ que consiste em uma combinação linear da forma $\boldsymbol{\eta} = \boldsymbol{X\beta}$ em que $\boldsymbol{X}$ é uma matriz de ordem $n\mathbf{x}p$ de variáveis independentes (preditoras) e $\boldsymbol{\beta}$ é o vetor p-dimensional de parâmetros desconhecidos do modelo.

A função de ligação $g(.)$ é uma função diferenciável e monótona que associa os valores esperados das observações (componente aleatório) com as variáveis independentes (componente sistemático).

Suponha que uma variável aleatória binária Y_i segue uma distribuição de Bernoulli e assume os valores 1 ou 0, caso o indivíduo pertença ao grupo predeterminado ou não, respectivamente. E seja $\boldsymbol{x_i} = (1, x_1, x_2, \ldots, x_p)'$ o vetor de características do indivíduo i e $\pi(\boldsymbol{x_i})$ a proporção de indivíduos que encontram-se no grupo de interesse em função do perfil dos indivíduos. A esperança e variância de Y_i são dadas por:

$$E(Y_i) = \pi_i \tag{1.27}$$

e

$$Var(Y_i) = \pi_i(1 - \pi_i). \tag{1.28}$$

Dado que a distribuição Bernoulli pertence à família exponencial, aplicando MLG

utilizando a função *logit* como função de ligação, temos:

$$g(E(Y_i)) = g(\pi(\boldsymbol{x_i})) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_{p-1} X_{p-1,i} = \boldsymbol{x_i'}\boldsymbol{\beta}, \quad (1.29)$$

podendo também ser escrito da forma:

$$E(Y_i) = \pi_i(\boldsymbol{X}) = \frac{\exp(\beta_0 + \beta_1 X_{1,i} + \ldots + \beta_{p-1} X_{p-1,i})}{1 + \exp(\beta_0 + \beta_1 X_{1,i} + \ldots + \beta_{p-1} X_{p-1,i})}, \quad (1.30)$$

em que $0 \leq \pi_i(\boldsymbol{X}) \leq 1$.

A partir da estimativa dos β's, que será apresentada a seguir, a estimativa de π_i é dada por:

$$\hat{\pi_i} = \frac{\exp(\boldsymbol{x_i'}\hat{\boldsymbol{\beta}})}{1 + \exp(\boldsymbol{x_i'}\hat{\boldsymbol{\beta}})}. \quad (1.31)$$

1.3.1 Estimação dos Coeficientes

Em modelos de regressão logística, a estimativa dos parâmetros é realizada através do método da máxima verossimilhança (Hosmer Lemeshow, 2000). Eles são as únicas quantias desconhecidas que necessitam ser estimadas pois os valores de $\boldsymbol{X}$ são conhecidos.

Sabendo que os dados são oriundos de uma distribuição Bernoulli e uma vez que as observações do conjunto de dados são independentes, a função de verossimilhança é dada por:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\boldsymbol{x_i})^{y_i}(1 - \pi(\boldsymbol{x_i}))^{1-y_i}. \quad (1.32)$$

Pelo princípio do método da máxima verossimilhança, os valores estimados de $\boldsymbol{\beta}$ são aqueles que maximizam $L(\boldsymbol{\beta})$. Para obtenção desses valores, calcula-se a derivada dessa função em relação a cada um dos parâmetros e procura-se pelo ponto crítico onde a derivada é igual a zero.

Aplicando a transformação monotônica logaritmo natural (ln) à função de verossimilhança, em virtude da propriedade de que o logaritmo de um produto é igual à soma dos logaritmos dos fatores, é obtido:

$$\ln[L(\boldsymbol{\beta})] = \sum_{i=1}^{n} y_i \ln[\pi(\boldsymbol{x_i})] + (1 - y_i) \ln[1 - \pi(\boldsymbol{x_i})]. \quad (1.33)$$

Essa transformação é realizada para simplificar matematicamente o cálculo das derivadas, tendo em vista que os resultados da maximização das funções $L(\boldsymbol{\beta})$ e $\ln[L(\boldsymbol{\beta})]$ são exatamente os mesmos (Casella e Berger, 2010).

Dessa forma, diferenciando $\ln[L(\boldsymbol{\beta})]$ e igualando a zero obtém-se as equações de verossimilhança, que são expressões não lineares nos parâmetros e portanto, podem ser solucionadas via métodos numéricos iterativos, como por exemplo o método Newton-Raphson.

Os valores encontrados para $\boldsymbol{\beta}$ são chamados Estimadores de Máxima Verossimilhança (EMV) e indicam a importância de cada variável independente para a ocorrência do evento de interesse (Sicsú, 2010).

Os estimadores possuem diversas características, dentre elas está o conceito de eficiência, que é relacionado com a variância do mesmo, do qual infere-se como estimador mais eficiente o de menor variância. Um estimador é dito consistente quando o mesmo converge, em probabilidade, para o seu valor populacional quando o tamanho da amostra n tende para infinito, e não viesado quando a esperança do estimador é o seu valor populacional, ou seja, $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.

Quando n tende a infinito (comportamento assintótico), visto na prática como n suficientemente grande, o estimador de máxima verossimilhança $\hat{\boldsymbol{\beta}}$ possui distribuição aproximadamente Normal com média $\boldsymbol{\beta}$ e variância tendendo para o limite inferior da desigualdade de Cramer-Rao. Além disso, ele é consistente já que $Var(\hat{\boldsymbol{\beta}}) \to 0$ quando $n \to \infty$ (Ehlers, 2009).

A significância dos estimadores pode ser testada através do Teste da Razão de Verossimilhança, que tem o intuito de comparar os valores observados da variável resposta com os valores preditos obtidos dos modelos com e sem a variável que se deseja testar. Outro teste que pode ser realizado é o Teste de Wald.

Após garantirmos uma boa estimação de máxima verossimilhança, que acarrete pouco ou nenhum viés aos parâmetros, um intervalo de $(1-\alpha)$ de confiança para β_i, $i = 1, \ldots, k$, utilizado para se fazer inferências, é dado por:

$$\beta_i \pm z_{1-\alpha/2} * (Var(\hat{\beta}_i))^{1/2} \tag{1.34}$$

em que $Var(\hat{\beta}_i)$ é o elemento diagonal correspondente ao parâmetro de β_i da matriz

de informação de Fisher observada e $z_{1-\alpha/2}$ é o quantil $(\frac{1-\alpha}{2} * 100)\%$ da distribuição Normal padrão.

1.3.2 Interpretação dos Coeficientes

Na Regressão Logística, os coeficientes das variáveis independentes podem ter diversas interpretações, uma vez que eles influenciam o *logit* (logaritmo natural da razão de chance), a razão de chance e as probabilidades.

No logit, a estimativa do parâmetro indica a alteração na variável dependente por unidade de variação da variável independente. Ou seja, caso uma variável x_1 tenha coeficiente 10 e todas as outras permaneçam constantes, então o acréscimo de uma unidade em x_1 implica no acréscimo de 10 no *logit*. Apesar de simples, em termos práticos, essa interpretação não apresenta nenhum significado intuitivo e nem melhora a qualidade da informação disponível.

Para a interpretação do coeficiente sob a razão de chance, basta fazer e^{β_i} para identificar o impacto dessa variável. Assim, o efeito dos coeficientes sobre a razão de chance é de natureza multiplicativa. Desse modo, um coeficiente igual a 0 significa que o efeito da variável na resposta é nulo, uma vez que a razão de chance fica multiplicada por 1. Isso posto, conclui-se que, coeficientes positivos contribuem para elevar a razão de chance e a probabilidade, e coeficientes negativos reduzem esses valores.

É importante destacar que a relação estabelecida entre as variáveis explicativas e a variável dependente no modelo logístico não é linear.

1.3.3 Avaliação do ajuste do Modelo

O teste de Hosmer e Lemeshow é a forma mais usual para a avaliação do ajuste do modelo de regressão logística de resposta binária. Esse teste avalia o modelo ajustado comparando as frequências observadas e as esperadas, propondo dois tipos de agrupamento que se baseam nas probabilidades estimadas.

Primeiramente, as observações são classificadas e os eventos de probabilidade são estimados. As observações são, então, divididas em cerca de 10 grupos. Seja N o número total de indivíduos e M o número alvo de indivíduos para cada grupo é dada

por:

$$M \cong [0,1 * N + 0,5]. \tag{1.35}$$

O número de grupos pode ser menor do que 10 se houver menos do que 10 padrões de variáveis explicativas. Devendo haver pelo menos três grupos mínimos para que a estatística de Hosmer-Lemeshow possa ser determinada.

A estatística proposta, por meio de simulação, segue uma distribuição Qui-quadrado quando não há replicação em qualquer uma das subpopulações.

A estatística do teste é dada por:

$$H = \sum_{g=1}^{G} \frac{(O_g - N_g \pi_g)^2}{N_g \pi_g (1 - \pi_g)}, \tag{1.36}$$

em que N_g é a frequência total de individuos no g-ésimo grupo $g = 1, 2, \ldots, G$; O_g é a frequência total de resultados de evento no g-ésimo grupo e π_g é a probabilidade média estimada previsto de um resultado de eventos para o g-ésimo grupo.

A estatística de Hosmer-Lemeshow é comparada com uma distribuição qui-quadrado com $g - 2$ graus de liberdade. Maiores valores da estatística do teste em relação ao p-valor indicam uma falta de ajuste do modelo.

1.3.4 Avaliação do Modelo

Um dos principais mecanismos para avaliar um modelo de Regressão Logística é o *Log Likelihood Value*. Esse indicador tem o objetivo de verificar a capacidade de estimação geral do modelo. Quanto mais próximo de zero, melhor o grau de adequação do modelo.

Mas o *Log Likelihood Value* sozinho oferece pouca informação sobre a qualidade do modelo. Assim, outras medidas são importantes.

O teste de Hosmer e Lemeshow é outra estratégia que pode contribuir para a avaliação do modelo Logístico. Nesse caso, busca-se comparar os valores preditos com os observados. Dessa forma, caso haja diferenças significativas entre eles, conclui-se que o modelo não é capaz de produzir estimativas confiáveis.

1.4 RH para Ciência, Tecnologia e Inovação

O estudo dos recursos humanos que dão suporte à produção e difusão de conhecimentos científicos e tecnológicos é o foco da atividade de investigação do Centro de Gestão e Estudos Estratégicos (CGEE), que é uma organização Social supervisionada pelo Ministério da Ciência, Tecnologia, Inovações e Comunicações - MCTIC e tem como missão "subsidiar processos de tomada de decisão em temas relacionados à ciência, tecnologia e inovação, por meio de estudos em prospecção e avaliação estratégica baseados em ampla articulação com especialistas e instituições do SNCTI". (CGEE, 2013)

A atividade, que teve como marco principal a elaboração e publicação do livro Doutores 2010: estudo da demografia da base técnico-científica brasileira, tem como ponto de partida a discussão sobre o perfil e a evolução dos programas de pós-graduação e dos correspondentes titulados, além do perfil da inserção dos egressos no mercado profissional. O objetivo maior é a geração de informações e estudos sobre as dimensões e características dos Recursos Humanos de Ciência, Tecnologia e Inovação (RHCTI) brasileiros.

Além da linha dos Estudos da demografia da base técnico-científica brasileira, que abrange três livros - Doutores 2010 [...], Mestres 2012 [...] e Mestres e Doutores 2015 [...] -, uma avaliação do Programa Institucional de Bolsas de Iniciação Científica (Pibic) foi tomada como base e motivação para o desenvolvimento deste trabalho.

1.4.1 Doutores 2010: estudos da demografia da base técnico-científica brasileira

O livro Doutores 2010: estudos da demografia da base técnico-científica brasileira (CGEE, 2010), trouxe pela primeira vez um retrato estatístico aprofundado da população de doutores titulados no país, abordando não apenas a quantidade e a variedade de programas e titulados em programas de pós-graduação, mas também a situação do emprego dos doutores.

Foram apresentadas informações detalhadas sobre a formação de doutores titulados no Brasil no período 1996-2008 e sobre o emprego destes no ano de 2008. Os dados

sobre formação foram basicamente fornecidos pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) do Ministério da Educação através das bases do Coleta Capes, que é alimentado de forma regular pelos programas de pós-graduação e contém informações detalhadas sobre os programas, inclusive sobre os indivíduos que neles obtiveram seus títulos.

Os dados sobre emprego foram provenientes da Relação Anual de Informações Sociais (RAIS) do ano de 2008, fornecida pelo Ministério do Trabalho e Emprego (MTE). A RAIS é um questionário que empregadores brasileiros, públicos ou privados, enviam anualmente ao Ministério com informações individualizadas sobre todos seus empregados.

1.4.2 Mestres 2012: estudos da demografia da base técnico-científica brasileira

Em 2012, após a repercussão positiva do primeiro livro e a observância de evidências[1] do fato de informações ou análises ali divulgadas terem tido impacto na política brasileira de pós-graduação, o CGEE veio a publicar Mestres 2012: estudos da demografia da base técnico-científica brasileira (CGEE, 2012), com abordagem e metodologia muito semelhantes as da publicação anterior. Nesse caso, a população alvo da análise era constituída dos mestres titulados no Brasil entre 1996 e 2012 e o ano de emprego era 2012. Aqui também as fontes dos dados foram o Coleta Capes e a Relação Anual de Informações Sociais (RAIS).

Na segunda parte da publicação, foram analisadas separadamente as populações de mestres e doutores do Brasil e comparadas com os brasileiros de todos os níveis educacionais através dos microdados que continham os resultados da amostra do Censo 2010, disponibilizados pelo IBGE.

[1] A elaboração do novo Plano Nacional de Pós-Graduação - PNPG 2010-2020 (Capes 2010) utilizou extensivamente informações, análises e dados divulgados pelo livro Doutores 2010 (CGEE, 2012).

1.4.3 Mestres e Doutores 2015: estudos da demografia da base técnico-científica brasileira

A partir da experiência adquirida com a elaboração dos livros anteriores, que consiste da expansão e consolidação da capacidade de o CGEE adquirir, atualizar e tratar bases de dados de interesse para o estudo dos recursos humanos em CT&I, foi possível a construção e publicação de Mestres e Doutores 2015: estudos da demografia da base técnico-científica brasileira (CGEE, 2015).

Essa publicação avança em novos campos e a primeira diferença significativa em relação às publicações antecessoras é uma abordagem conjunta do contigente de mestres e doutores titulados no Brasil entre 1996 e 2014. Já a dimensão tradicional de análise da inserção dos egressos no mercado de trabalho se tornou mais abrangente e passou a constar de 6 anos (2009-2014). E ainda, nos três últimos capítulo do livro, é apresentada uma análise do conjunto de mestres e doutores engajados nas entidades empresariais, públicas ou privadas.

Mais uma vez, os dados sobre formação foram disponibilizados pela Capes. O sistema informatizado Coleta Capes subsidiou, até 2013, o processo de avaliação realizado pela Capes, bem como os programas de fomento e delineamento de políticas institucionais. A partir daí, a Plataforma Sucupira se tornou a ferramenta utilizada para a disponibilização de informações, processos e procedimentos.

1.4.4 A Formação de Novos Quadros para CT&I: Avaliação do Programa Institucional de Bolsas de Iniciação Científica - Pibic

Em 2017, o CGEE publicou um resumo executivo com os resultados de uma avaliação do Programa Institucional de Bolsas de Iniciação Científica - Pibic (CGEE, 2017). O Pibic foi inaugurado no final dos anos 80 pelo CNPq como um novo canal de distribuição de bolsas de iniciação científica e, desde os anos 90, se consolidou como um programa permanente do CNPq, envolvendo todas as unidades da federação, dezenas de instituições de ensino e pesquisa e milhares de alunos e orientadores.

Outras avaliações do Programa já haviam sido realizadas em anos anteriores,

trazendo como um destaque a consolidação da demanda de egressos do programa pelos cursos de pós-graduação e, com isso, o aumento do fluxo de estudantes para o mestrado (Marcuschi, 1996). Assim, uma nova avaliação se fez necessária devido aos resultados obtidos pelo programa e às mudanças que ocorreram ao longo dos anos no cenário do ensino superior brasileiro.

Os resultados do trabalho foram apresentados em três partes: a primeira analisa a experiência dos bolsistas e orientadores a partir de um questionário; a segunda mostra a trajetória dos egressos avaliada através da análise da conclusão de cursos de mestrado e doutorado e tempo despendido com a formação; a terceita parte traz uma avaliação de impacto do Pibic para mensurar seus efeitos na formação pós-graduada e inserção no mercado de trabalho, realizada por meio da técnica Propensity Score Matching.

Nessa última parte do trabalho, foram utilizados dados de bolsistas egressos da Instituição de Ensino Superior Unesp e, a partir dos resultados de uma regressão logística aplicada aos dados, foi possível concluir que, apesar de a participação no Pibic não contribuir diretamente com o emprego em instituições que atuam em atividades típicas de professores e pesquisadores - "Educação"e "Atividades Profissionais, Científicas e Técnicas-, aumenta as chances de conclusão do mestrado e do doutorado e isso sim influencia o percurso profissional para atividades de caráter acadêmico.

Capítulo 2

Análise de dados

As bases que serviram como fonte de dados para este trabalho foram escolhidas por já terem se mostrado boas alternativas em estudos da influência que informações sobre a pós-graduação exercem em políticas públicas brasileiras, como pode-se notar na Seção 1.4.

Apesar de serem confiáveis, em parte devido a obrigatoriedade de declaração, as bases de dados precisaram ser submetidas a tratamentos para que as variáveis se tornassem mais adequadas na análise do tempo de ingresso dos doutores em atividades acadêmicas, que é o principal objetivo do trabalho. Na Seção 2.2, para conhecimento, os tratamentos nas bases de dados serão apresentados.

2.1 Bases de dados

As principais fontes de dados utilizadas neste estudo foram o Coleta Capes e a Plataforma Sucupira (que veio a substituir o primeiro em 2013), sistemas de informação sobre a pós-graduação brasileira criados e mantidos pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) do Ministério da Educação (MEC), e a Relação Anual de Informações Sociais (RAIS) do Ministério do Trabalho e Emprego (MTE), provedora das informações sobre emprego formal no Brasil.

Como mencionado na Seção 1.4.1, o Coleta Capes e a Plataforma Sucupira são alimentados de forma regular pelos programas de pós-graduação. Esses sistemas contêm informações detalhadas sobre os programas e indivíduos que neles obtiveram

seus títulos e são utilizadas pelo processo de avaliação e credenciamento de programas de pós-graduação e pelo sistema de concessão a estes de bolsas e auxílios à pesquisa.

Obteve-se acesso às bases do Coleta Capes entre 1996 e 2012 e da Plataforma Sucupira entre 2013 e 2014, que continham informações sobre os indivíduos que obtiveram títulos de doutorado no período 1996-2014 - Cadastro de Pessoas Físicas (CPF), nome, sexo e idade - e sobre questões acadêmicas - data da titulação, informações sobre a instituição de ensino superior (nome, natureza jurídica e UF) e sobre o programa de pós-graduação (nome, nota recebida na avaliação da Capes, área do conhecimento a que o programa pertence e nome do orientador). É importante destacar que, para fins da análise, os dados sobre a obtenção do título de doutorado no Brasil foram obtidos exclusivamente a partir das bases fornecidas pela Capes, que são o Coleta Capes e a Plataforma Sucupira.

A Relação Anual de Informações Sociais (RAIS) tem como objetivos suprir às necessidades de controle da atividade trabalhista no Brasil, prover dados para a elaboração de estatísticas do trabalho e disponibilizar informações do mercado de trabalho às entidades governamentais. Todo estabelecimento deve fornecer as informações requeridas de cada um de seus empregados para o MTE através da RAIS (MTb, SPPE, DES, CGET, 2016).

As bases da RAIS são anuais e cada uma delas contém dados dos empregados formais ativos em algum período durante aquele ano. Existem variáveis que indicam a data de admissão do indivíduo (que pode ter acontecido naquele mesmo ano ou em anos anteriores), se ele foi demitido durante o ano e, em caso afirmativo, a data de demissão. Um indivíduo pode possuir mais de um registro em cada base caso possua mais de um vínculo formal ou tenha mudado de emprego durante o ano vigente. Para esses caso, a escolha do vínculo a ser utilizado na análise será apresentada na Seção 2.2.

Através das bases RAIS 2010-2014, foi possível a obtenção de informações relativas aos doutores empregados a partir dos CPFs obtidos nas bases do Coleta Capes e Plataforma Sucupira - data de admissão, quantidade de horas contratuais, remuneração média do ano, tempo no emprego e ocupação - e aos estabelecimentos empregadores - UF, tamanho, atividade econômica e natureza jurídica. A variável escolaridade, também encontrada na base de dados, traz informações sobre o grau de instrução

do empregado mas não foi utilizada aqui devido o número informado de doutores ser muito inferior ao encontrado nas bases da Capes. Acredita-se que isso ocorra devido ao mal preenchimento dessa variável por parte do responsável no estabelecimento empregador.

2.2 Tratamento da base de dados

Antes da apresentação dos resultados do estudo, é importante deixar claro quais os tratamentos adotados, além do processo de preparação em que foram padronizados os nomes das variáveis e alguns caracteres, para que as bases de dados fossem unificadas.

Os tratamentos consistem tanto em validação de dados quanto em seleção de grupos de indivíduos de interesse.

2.2.1 Validação de CPF

A base de dados do Coleta Capes e da Plataforma Sucupira utilizada para a geração dos resultados do livro Mestres e Doutores 2015: estudos da demografia da base técnico-científica brasileira (CGEE, 2015), que trazem informações sobre os egressos da pós-graduação no Brasil, é a mesma utilizada nesse estudo com exceção dos dados disponíveis sobre mestres. Aqui, optou-se por trabalhar apenas com os doutores devido a especifidade do objetivo geral, que é a análise do tempo até o ingresso em uma atividade acadêmica, não sendo incluídos os doutorandos que entram na carreira acadêmica por possuírem o título de mestrado.

Como mencionado no livro, o número de CPF inválidos de mestres e doutores diminuiu significativamente após o ano 2000 e os casos em que isso ocorre com maior frequência são aqueles em que são utilizados o número do passaporte para o registro de alunos estrangeiros. Eventuais erros de digitação também foram identificados. Para correção do maior número possível de registros, foram utilizados dados da Plataforma Lattes, onde o critério para substituição de um CPF inválido foi a identificação de indivíduos com mesmo nome, área do conhecimento, nível de titulação e instituição de ensino superior.

A validação de CPF também foi feita nos dados da RAIS para que fosse possível

a união dessa base com a base dos egressos do doutorado. O procedimento adotado para a validação é apresentado a seguir.

O Cadastro de Pessoas Físicas - CPF é formado por 11 dígitos númericos que seguem a forma $X_1X_2X_3.X_4X_5X_6.X_7X_8X_9 - YZ$. A validação ocorre a partir dos 9 primeiros dígitos e verificando-se os resultados de dois cálculos simples, mostrados a seguir, correspondem aos dígitos verificadores Y e Z.

O primeiro passo para a validação do dígito Y consiste na soma do produto dos 9 primeiros dígitos pela sequência decrescente de números de 10 a 2, representado pela expressão 2.1.

$$S_1 = X_1 \times 10 + X_2 \times 9 + X_3 \times 8 + X_4 \times 7 + X_5 \times 6 + X_6 \times 5 + X_7 \times 4 + X_8 \times 3 + X_9 \times 2 \quad (2.1)$$

Em seguida, multiplica-se a soma S_1 por 10 e divide por 11. Caso o resto da divisão seja igual a Y, a primeira parte da validação está correta. É importante observar que se o resto da divisão for igual a 10, Y é válido se for igual a 0.

O processo da validação do dígito Z é bastante semelhante. Como mostra a expressão 2.2, soma-se o produto dos 9 dígitos e Y pela sequência decrescente de 11 a 2. Dessa forma, tem-se que:

$$\begin{aligned} S_2 = & X_1 \times 11 + X_2 \times 10 + X_3 \times 9 + X_4 \times 8 + X_5 \times 7 + X_6 \times 6 + \\ & X_7 \times 5 + X_8 \times 4 + X_9 \times 3 + Y \times 2 \end{aligned} \quad (2.2)$$

Assim como na verificação de Y, multiplica-se a soma S_2 por 10 e divide por 11. O dígito Z é válido se o resto da divisão for igual a ele. Novamente, se o resto da divisão for igual a 10, Z é válido se for igual a 0.

2.2.2 Seleção de egressos titulados entre 2010 e 2013

Como mencionado na Seção 2.1, os dados sobre egressos da pós-graduação a que se teve acesso eram referentes aos anos 1996 a 2014, enquanto que a base com informações sobre emprego formal trazia dados a partir de 2010 (RAIS 2010). Tendo como objetivo a análise do tempo entre a data de titulação do egresso no doutorado e a data de admissão em um vínculo cujo estabelecimento empregador tivesse ativi-

dade econômica principal relacionada à alguma atividade acadêmica - assunto tratado na subseção 2.2.4 -, decidiu-se considerar apenas os titulados a partir de 2010. Isso se deve à necessidade de acompanhamento da trajetória do egresso, que não seria possível caso o ano de titulação fosse inferior ao primeiro ano de emprego disponível.

A título de exemplo, caso os egressos que titularam no doutorado em 2009 fossem considerados na análise, não haveria certeza de que eles ingressaram no mercado de trabalho naquele mesmo ano e não chegaram a ser demitidos. Logo, não estariam presentes nos registros da base RAIS 2010. Apenas seria factível o acompanhamento dos titulados em 2009 que ingressaram no mercado de trabalho também em 2009 e permaneceram empregados até, no mínimo, 2010. Essa perda de informação seria maior e imensurável caso fossem considerados todos os egressos desde 1996.

Os egressos titulados em 2014 não foram incluídos na análise devido ao prazo relativamente curto para obter emprego, pois a base da RAIS mais recente disponível para esse estudo era a do ano 2014. Segundo CGEE (2010), uma das razões que contribuem para os doutores demorarem a obter emprego é associada ao fato de uma grande proporção dos doutores brasileiros trabalharem em instituições públicas e o acesso aos quadros funcionais dessas instituições depende de concursos públicos, que geralmente são processos complexos e demorados, que seguem periodicidade irregular. Outras razões que contribuem para essa demora são o fato dos doutores possuírem qualificação altamente especializada e esperarem por oportunidades em outros setores de atividade adequadas à sua formação específica, além do fato de recém-doutores frequentemente prolongarem as atividades de P&D ou ensino ou continuarem com trabalhos que não constam como emprego formal, tais como bolsista de pós-doutorado, professor colaborador, e outros.

Apesar dos doutores titulados em 2013 também poderem ser considerados recém-doutores, eles fizeram parte do grupo dos egressos analisados no presente trabalho por terem participação relativa maior e mais expressiva do que os titulados em 2014 em relação ao total de titulados a partir de 2010.

2.2.3 Seleção do título mais recente

É possível encontrar casos de indivíduos que obtiveram mais de um título de doutorado nas bases do Coleta Capes e Plataforma Sucupira no período 1996-2014. Ao considerar o período 2010-2013, essa mesma situação ocorre mas com frequência menor. No entanto, nesse último caso, algumas pessoas que apareceram como tendo apenas um título haviam obtido outro título no período 1996-2009 ou até mesmo antes de 1996, de quando não há dados disponíveis. Assim, devido aos motivos apresentados na Subseção 2.2.2, como o período 2010-2013 foi o estabelecido, naturalmente seleciona-se o vínculo mais recente desses indivíduos.

Tomando como base essa circunstância, decidiu-se por adotar o vínculo mais recente dos indivíduos que apresentaram mais de um título entre 2010 e 2013. Uma preocupação que surgiu até a tomada dessa decisão foi o fato de ocorrer uma possível perda de informação sobre o emprego do egresso quando a data de admissão fosse posterior a data de titulação mais antiga e anterior a data de titulação mais recente, devido a metodologia adotada na seleção dos vínculos empregatícios que será tratada na Subseção 2.2.4. No entanto, após o cruzamento das bases e uma análise inicial tanto a partir do vínculo mais antigo quanto do mais recente no período considerado, foi possível perceber que, na maioria dos casos, a data de admissão era posterior a de titulação nos dois casos. E, então, optou-se pelo título mais recente para que o tempo decorrido até a admissão fosse, em geral, reduzido.

2.2.4 Seleção de vínculos relacionados a atividades acadêmicas com data de admissão mais próxima a da titulação

Os estabelecimentos empregadores que possuem "Educação"ou "Atividades profissionais, científicas e técnicas"como atividade econômica principal concentram grande parte dos mestres e doutores e incluem os professores e grande parte dos pesquisadores (CGEE, 2015). Essas atividades correspondem às seções M e P da Classificação Nacional de Atividades Econômicas (CNAE 2.0), ver Tabela 2.1.

Para estudar o tempo que doutores levam para ingressar em atividades dessa natureza, e assim analisar a formação desse contingente que se dedica a docência e a pesquisa, foram selecionados apenas os vínculos de indivíduos em estabelecimentos

que possuiam seção da CNAE correspondente a "Educação"ou "Atividades profissionais, científicas e técnicas". A Tabela 2.1 mostra quais divisões da CNAE estão agregadas pelas seções M e P.

Ao selecionar os vínculos nas bases RAIS 2010 a 2013 a partir do CPF dos doutores encontrados nas bases da Capes, foi possível perceber que muitos indivíduos estavam presentes em mais de uma base, o que indicava, na maioria dos casos, a continuação em um mesmo emprego por mais de um ano, caso em que a data de admissão era única. Em outros casos, como quando o egresso acumulava vínculos empregatícios no mesmo período, ou seja, trabalhava em dois locais diferentes ao mesmo tempo, ou quando acontecia mudança de vínculo, seja por demissão ou opção própria, existia uma data de admissão para quantos vínculos o empregado estivesse ligado.

Tabela 2.1: Divisões pertencentes as seções Atividades profissionais, científicas e técnicas e Educação da CNAE

Seção/Divisão da CNAE	**Descrição**
M	**Atividades profissionais, científicas e técnicas**
69	Atividades jurídicas, de contabilidade e de auditoria
70	Atividades de sedes de empresas de consultoria em gestão empresarial
71	Serviços de arquitetura e engenharia, testes e análises técnicas
72	Pesquisa e desenvolvimento científico
73	Publicidade e pesquisa de mercado
74	Outras atividades profissionais, científicas e técnicas
75	Atividades veterinárias
P	**Educação**
85	Educação

A partir daí, para aplicação da metodogia estatística, surgiu a necessidade em se escolher apenas um vínculo por egresso e optou-se então pelo vínculo cuja data de admissão fosse a mais proxima da data de titulação no curso de doutorado e fosse também posterior a esta. Assim, foi definido que o tempo entre as duas datas fosse de no mínimo um dia. É importante notar que esse procedimento foi realizado após a seleção de vínculos em estabelecimentos de CNAE "Educação"ou "Atividades profissionais, científicas e técnicas"e, consequentemente, foram desconsiderados os

casos em que indivíduos conseguiram emprego em atividade de outra natureza até que surgisse uma oportunidade na área acadêmica.

Dessa forma, a variável resposta será definida como o tempo entre a obtenção do título de doutorado no Brasil e o ingresso em um vínculo de emprego formal cuja atividade principal possua natureza acadêmica.

Capítulo 3

Resultados

Os resultados da análise dos dados utilizados na análise do tempo entre a titulação no doutorado e admissão no mercado de trabalho formal estão apresentados nesse capítulo, que está organizado em três seções.

A primeira, denominada Análise descritiva dos dados, apresenta a tabela de frequências das covariáveis que serão utilizadas nos modelos de regressão de Cox e logística, apresentados nas Seções 3.2 e 3.3, e o gráfico da curva de Kaplan-Meier para os tempos de sobrevivência dos doutores titulados no Brasil. A Seção 3.3 traz ainda um comparativo dos modelos de Cox e logístico.

É importante destacar o grande número de observações presentes na base de dados, logo os resultados dos testes de hipóteses aplicados no processo de ajuste e avaliação de modelos devem ser interpretados com cautela.

3.1 Análise descritiva dos dados

Após o tratamento feito na base de dados, detalhado na Seção 2.2, o número de observações presentes passou a 52.859, referente aos doutores titulados no Brasil entre 2010 e 2013.

A Tabela 3.1 mostra a frenquência absoluta e relativa das covariáveis utilizadas no estudo. Para o ajuste dos modelos de regressão de Cox e logística, foram desconsiderados os 139 doutores que não tinham sua idade no momento da titulação informada na base de dados.

A tabela mostra que pouco mais da metade dos doutores são do sexo feminino e mais de 80% possuem até 45 anos, sendo que aproximadamente um terço se encontra na faixa etária entre 31 e 35 anos.

Tabela 3.1: Frequências absolutas e relativas das covariáveis

Variável	**Frequência absoluta**	**Frequência relativa %**
Sexo		
Feminino	27.813	52,62
Masculino	25.046	47,38
Idade		
Até 30 anos	10.778	20,39
31 a 35 anos	17.356	32,83
36 a 45 anos	14.274	27,00
Mais de 45 anos	10.312	19,51
Não informado	139	0,26
Região		
Norte	916	1,73
Nordeste	6.682	12,64
Sudeste	33.944	64,22
Sul	8.920	16,88
Centro-Oeste	2.397	4,53
Grande área		
Ciências agrárias	6.903	13,06
Ciências biológicas	5.454	10,32
Ciências da saúde	10.003	18,92
Ciências exatas e da terra	5.203	9,84
Ciências humanas	9.096	17,21
Ciências sociais aplicadas	4.497	8,51
Engenharias	5.617	10,63
Linguistíca, letras e artes	3.040	5,75
Multidisplinar	3.046	5,76
Nota da Capes		
Nota 3	972	1,84
Nota 4	12.822	24,26
Nota 5	19.845	37,54
Nota 6	11.304	21,39
Nota 7	7.916	14,98

No período considerado, a região Sudeste foi a que mais titulou doutores, cerca de 64%, seguida pela região Sul (16,9%) e pela região Nordeste (12,6%). As regiões Centro-Oeste e Norte foram responsáveis pela titulação de apenas 6,26% dos doutores. Uma discussão mais profunda sobre a desconcentração da pós-graduação nos estados

e regiões do país pode ser encontrada no livro Mestres e doutores 2015: estudos da demografia da base técnico-científica brasileira (CGEE, 2015).

As grandes áreas do conhecimento em que mais se titularam doutores foram as ciências da saúde e ciências humanas, agregando juntas mais de 19 mil indivíduos. As ciências agrárias, engenharias, ciências biológicas, exatas e da terra e sociais aplicadas titularam, cada uma, entre 4.497 e 6.903 doutores.

Sobre a nota obtida pelos programas de pós-graduação na avaliação trienal da Capes, 37,5% dos titulados obtiveram nota 5, 26,1% nota 3 ou 4 e 36,4% nota 6 ou 7. A Capes avalia os programas de pós-graduação a cada três anos e as notas atribuídas na avaliação vigoram pelo intervalo de três anos. Novos programas podem ser credenciados pela CAPES no intervalo entre as avaliações periódicas mas somente são credenciados programas que receberam nota igual ou superior a 3 no momento do credenciamento. Os novos programas permanecem com a nota recebida no credenciamento até a segunda avaliação trienal que vier a ocorrer após o momento em que se deu o credenciamento do programa. Os programas mais bem avaliados recebem a nota 7. A nota 5 é a mais elevada que pode ser atribuída a programas de mestrado, que não estão vinculados a programas de doutorado.

Esse trabalho tem como objetivo a análise do tempo que doutores levam entre a titulação e o ingresso no mercado em atividades de caráter acadêmico, definidas como vínculos empregatícios em estabelecimentos empregadores de CNAE "Atividades profissionais, científicas e técnicas"ou "Educação". Para esse fim, foram utilizadas técnicas e métodos da análise de sobrevivência.

Como todo estudo da área, esse também é caracterizado pela presença de censuras, que aqui são 35.103, cerca de 66,6% dos dados. Esse dado por ser explicado por falta de tempo de acompanhamento de alguns indivíduos, sobretudo os que se titularam em anos mais recentes, como 2013 em que houve apenas 1 ano para verificar se o evento de interesse ocorreu. Há também o caso de indivíduos que ingressaram no mercado em atividade de caráter acadêmico antes mesmo de concluir o doutorado.

Aqui também não foi possível a utilização de técnicas de análise estatística tradicionais. Assim, segue a Figura 3.1 com a curva de Kaplan-Meier que apresenta a função de sobrevivência estimada para cada tempo t, tempo em dias entre a titulação no doutorado ao ingresso em atividade acadêmica. Nota-se que nos maiores tempos a

estimativa da função é próxima a 0,6, um valor relativamente alto que pode ter como justificativa os mesmos motivos utilizados na tentativa de explicar a quantidade de censuras presente na base.

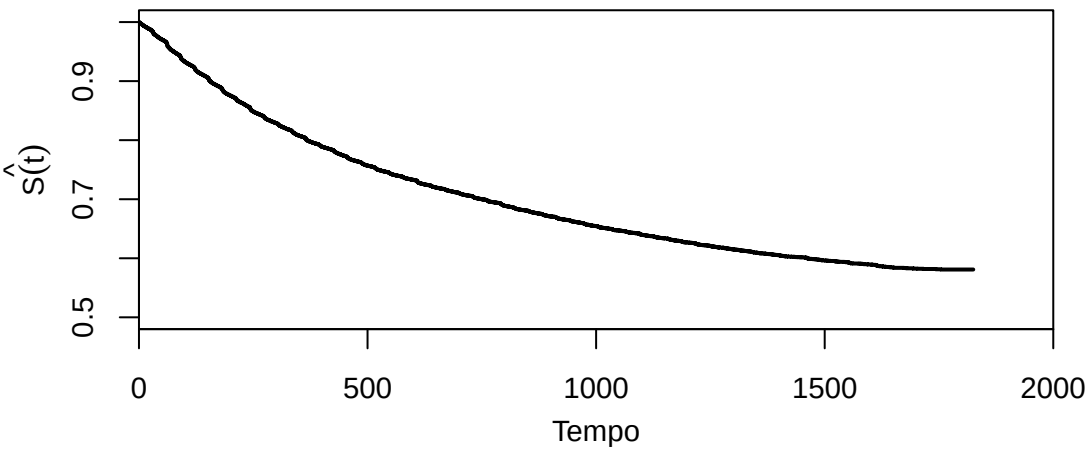

Figura 3.1: Curva estimada pelo método não-paramétrico de Kaplan-Meier para os tempos de sobrevivência dos doutores titulados no Brasil

3.2 Modelo de regressão de Cox

O modelo de regressão de Cox foi ajustado aos dados na tentativa de encontrar um modelo que representasse bem o comportamento deles em um contexto semi-paramétrico. É necessário a avaliação da suposição básica para seu uso que consiste na proporcionalidade das taxas de falha e pode ser realizada através da análise dos gráficos das curvas de sobrevivência para cada covariável.

A Figura 3.2 mostra os gráficos de Kaplan-Meier para cada uma das 5 covariáveis incluídas no modelo. É importante destacar que a variável natureza jurídica da instituição de ensino superior, que possuia as categorias federal, estadual, municipal e particular, estava presente originalmente na base mas não foi incluída no modelo pois as curvas de sobrevivência de suas categorias não apresentaram diferenças entre si.

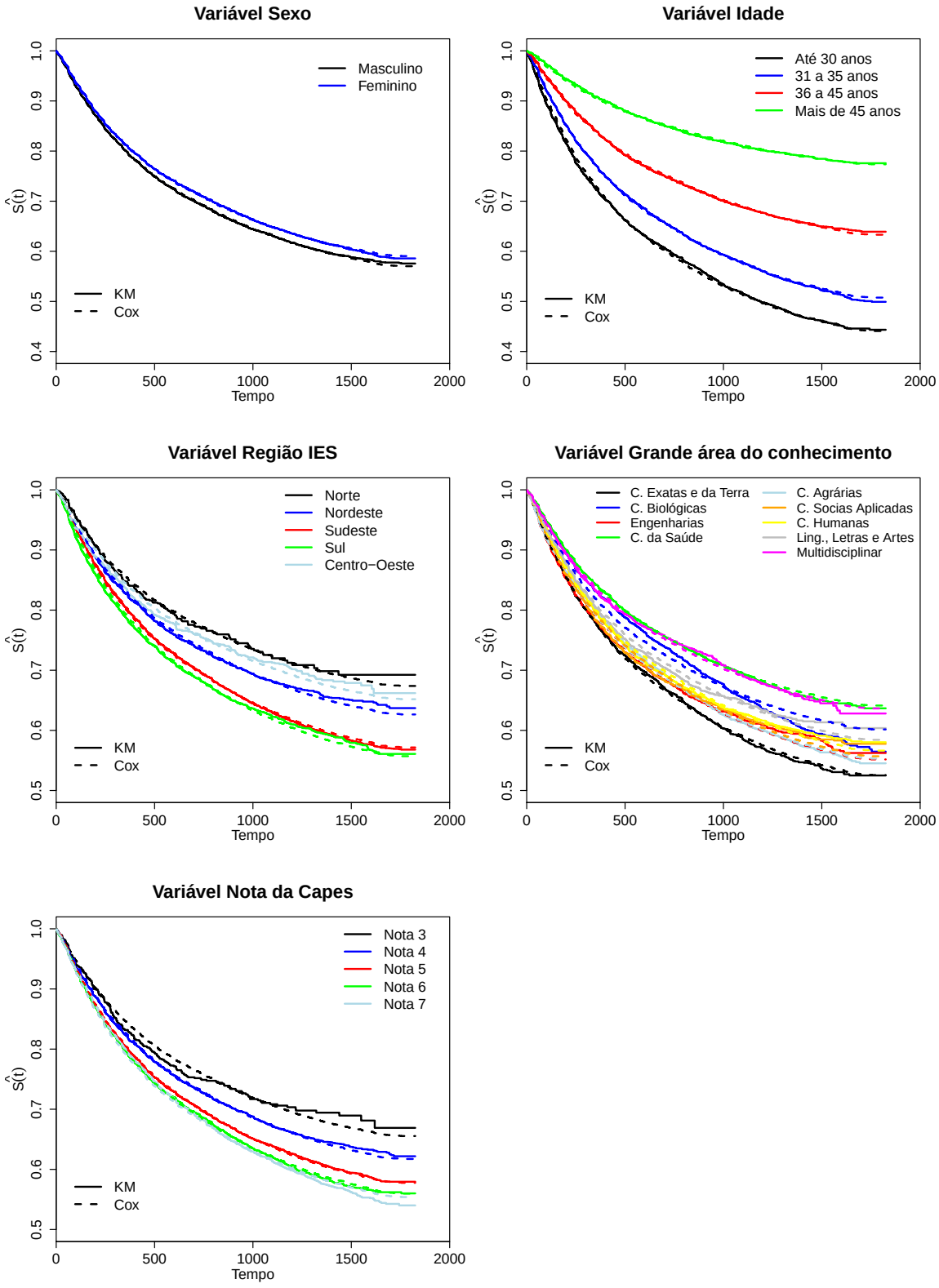

Figura 3.2: Curvas de sobrevivência das categorias das covariáveis estimadas pelo modelo de regressão de Cox e por Kaplan-Meier

Na análise de cada uma das covariáveis, a princípio, as categorias da variável

sexo não aparentam ser diferentes. No entanto, como será visto mais adiante, essa se mostrou uma variável significativa tanto no modelo de Cox ajustado individualmente para as covariáveis quanto no modelo incluindo todas elas. E ainda, o resultado da avaliação geral do ajuste se mostrou satisfatório, como é possível notar na Figura 3.4. As curvas estimadas por Kaplan-Meier e por Cox possuem boa aproximação.

O mesmo ocorre com a aproximação das curvas das variáveis idade, região IES e nota da Capes, com destaque para a primeira delas. A suposição de riscos proporcionais não parece ser violada, apesar do pequeno desencontro nos tempos finais para as outras duas.

Para a variável grande área do conhecimento, algumas das categorias parecem se confundir e outras se cruzam com o passar do tempo t. Ainda assim, dada a importância da informação carregada pela variável, o resultado da avaliação do ajuste mostrada na Figura 3.4 para a covariável e o resultado do teste de tendências apresentado na Tabela 3.2, optou-se por não retirar grande área do conhecimento do modelo.

Em um primeiro momento, o modelo de regressão de Cox foi ajustado para cada uma das covariáveis separadamente. A Tabela 3.2 mostra as estimativas dos coeficientes β's e do risco relativo em cada caso considerado, assim como o resultado do teste de tendências que permite avaliar a suposição de riscos proporcionais a partir do coeficiente de correlação de Pearson.

Considerando o teste de tendências e admitindo um nível de significância de 5%, apenas as categorias 36 a 45 anos e mais de 45 anos da variável idade estariam violando a suposição de riscos proporcionais, o que não é perceptível na Figura 3.2. Como mencionado anteriormente, devido o grande número de observações, deve-se interpretar com cuidado o resultado dos testes de hipótese presentes nesse estudo e deve-se combinar com os resultados obtidos através de técnicas gráficas. Note que o coeficiente de correlação é baixo ($|\rho| \leq 0,021$) e o teste somente foi significativo devido ao grande poder do teste que conta com uma amostra de mais de 50 mil observações.

Agora, dando enfase no risco relativo entre as categorias das covariáveis, o risco de ingressar em atividade de carater acadêmico não é muito diferente entre o sexo feminino e masculino, já que o risco relativo das mulheres é 94% do risco dos homens.

Tabela 3.2: Estimativas dos coeficientes $\boldsymbol{\beta}$ no modelo de regressão de Cox ajustado individualmente para cada covariável e teste de tendências

Variável	$\boldsymbol{\beta}$	$\boldsymbol{e^{\beta}}$ (I.C. 95%)	p	ρ^{*}	p**
Sexo					
Masculino	0	1	-	-	-
Feminino	-0,06	0,94 (0,91;0,97)	<0,001	0,013	0,072
Idade					
Até 30 anos	0	1	-	-	-
31 a 35 anos	-0,19	0,83 (0,79;0,86)	<0,001	-0,002	0,772
36 a 45 anos	-0,58	0,56 (0,54;0,58)	<0,001	-0,020	0,007
Mais de 45 anos	-1,16	0,31 (0,29;0,33)	<0,001	-0,021	0,006
Região					
Norte	0	1	-	-	-
Nordeste	0,17	1,18 (1,03;1,36)	0,01	0,014	0,070
Sudeste	0,35	1,42 (1,24;1,61)	<0,001	0,013	0,082
Sul	0,39	1,48 (1,30;1,70)	<0,001	0,010	0,165
Centro-Oeste	0,08	1,08 (0,93;1,26)	0,29	0,013	0,089
Grande área					
Ciências exatas e da terra	0	1	-	-	-
Ciências biológicas	-0,23	0,79 (0,74;0,84)	<0,001	-0.011	0,153
Engenharias	-0,08	0,92 (0,87;0,98)	0,01	-0,006	0,447
Ciências da saúde	-0,37	0,69 (0,65;0,73)	<0,001	-0,005	0,511
Ciências agrárias	-0,08	0,92 (0,87;0,98)	0,008	0,002	0,808
Ciências sociais aplicadas	-0,09	0,91 (0,85;0,97)	0,005	0,002	0,788
Ciências humanas	-0,12	0,88 (0,84;0,94)	<0,001	-0,007	0,343
Linguistíca, letras e artes	-0,18	0,84 (0,78;0,90)	<0,001	-0,002	0,802
Multidisplinar	-0,35	0,70 (0,65;0,76)	<0,001	-0,002	0,833
Nota da Capes					
Nota 3	0	1	-	-	-
Nota 4	0,13	1,14 (1,00;1,29)	0,04	0,009	0,250
Nota 5	0,26	1,29 (1,14;1,46)	<0,001	0,017	0,123
Nota 6	0,31	1,37 (1,21;1,55)	<0,001	0,010	0,163
Nota 7	0,33	1,39 (1,23;1,58)	<0,001	0,014	0,057

Notas: A categoria com $\boldsymbol{\beta} = 0$ é o nível de referência.
(*) ρ é o coeficiente de correlação entre o tempo de sobrevivência e o resíduo de Schoenfeld.
(**) valor-p do teste de tendências (1.25).

Em relação a idade, o grupo com maior risco relativo é o de titulados com até 30 anos. O risco de ingressar em atividade de carater acadêmico para titulados com mais de 45 anos é de apenas 31% do risco do primeiro grupo.

Os indivíduos titulados na região Norte são os que possuem menor risco relativo. Os titulados nas regiões Sul e Sudeste possuem riscos 48% e 42% maiores que os

da região Norte, respectivamente. A estimativa do β da região Centro-Oeste não se mostrou significante no modelo ajustado.

Em relação a grande área do conhecimento, o risco de nenhuma grande área ultrapassou o risco da grande área ciências exatas e da terra, nem mesmo ao serem considerados os intervalos de confiança, apesar das áreas engenharias, ciências agrárias e ciências sociais aplicadas apresentarem riscos bem próximos ao daquela área (acima de 90%). A área que possui menor risco relativo (69% do risco de exatas e da terra) é ciências da saúde.

Por fim, assim como esperado, o risco relativo de doutores que titularam em programas com notas mais altas é maior em relação aos programas com nota imediatamente inferior. O risco de um indivíduo titulado em um programa que recebeu nota 7 é 39% maior do que o risco de um indivíduo titulado em um programa nota 3.

A verificação da qualidade do ajuste do modelo de riscos proporcionais de Cox foi realizada por meio dos resíduos de Cox-Snell. Segundo Lawless (2003), os resíduos de Cox-Snell vêm de uma população homogênea e devem seguir uma distribuição exponencial com média 1. Quanto mais a função de sobrevivência dos resíduos do modelo de Cox se aproxima da função de sobrevivência da distribuição exponencial, melhor é o ajuste do modelo. Os gráficos são apresentados nas Figuras 3.3 e 3.4.

Pode ser observado que, para todas as covariáveis, a função de sobrevivência dos resíduos se ajustou bem à função de sobrevivência da exponencial padrão. Isso indica que os modelos estão bem ajustados aos dados. E, como aqui o interesse não é a escolha de um melhor modelo, decidiu-se por manter todas as covariáveis no modelo final.

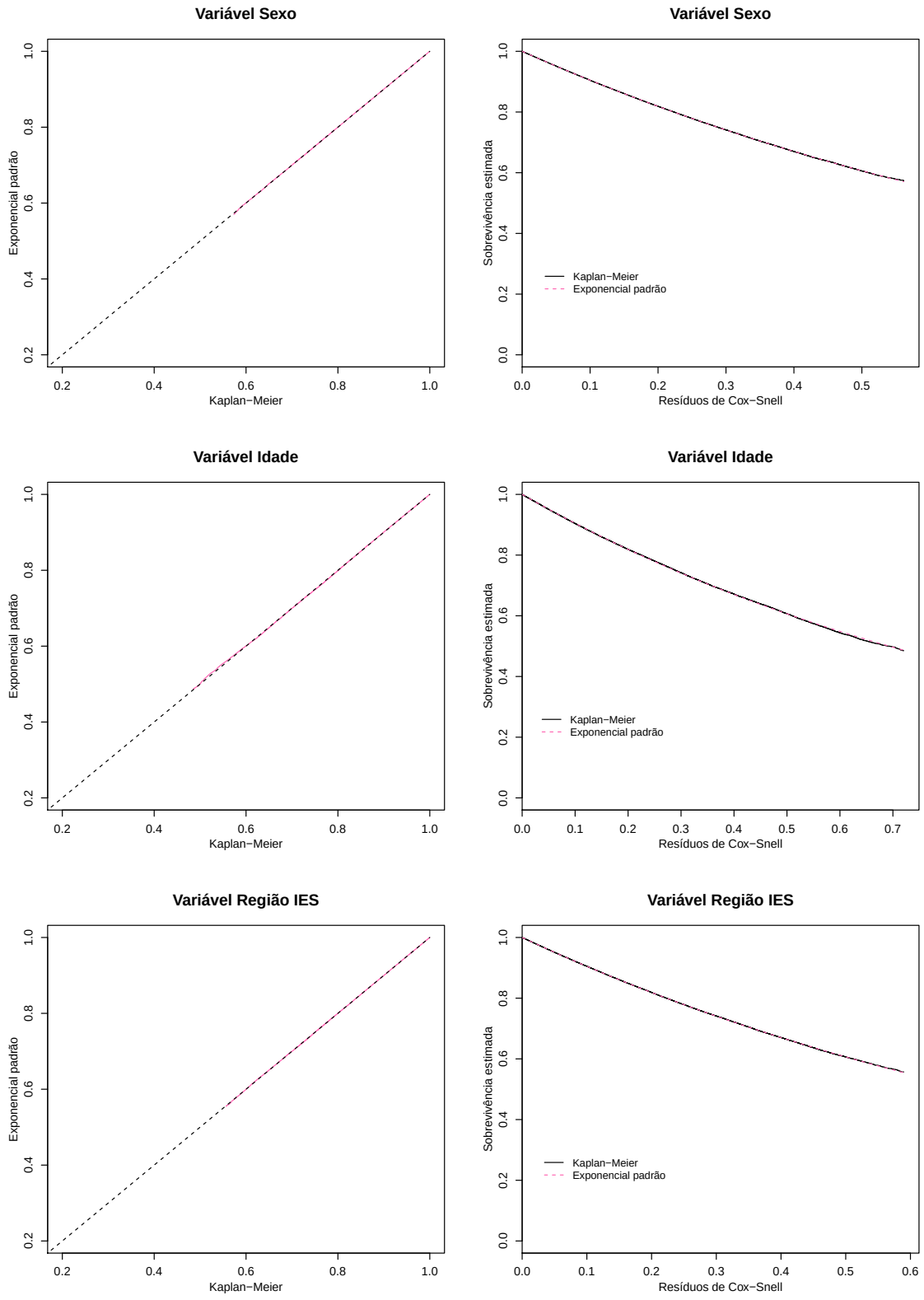

Figura 3.3: Comparação das funções de sobrevivência dos resíduos de Cox-Snell e da distribuição Exponencial para as variáveis Sexo, Idade e Região IES

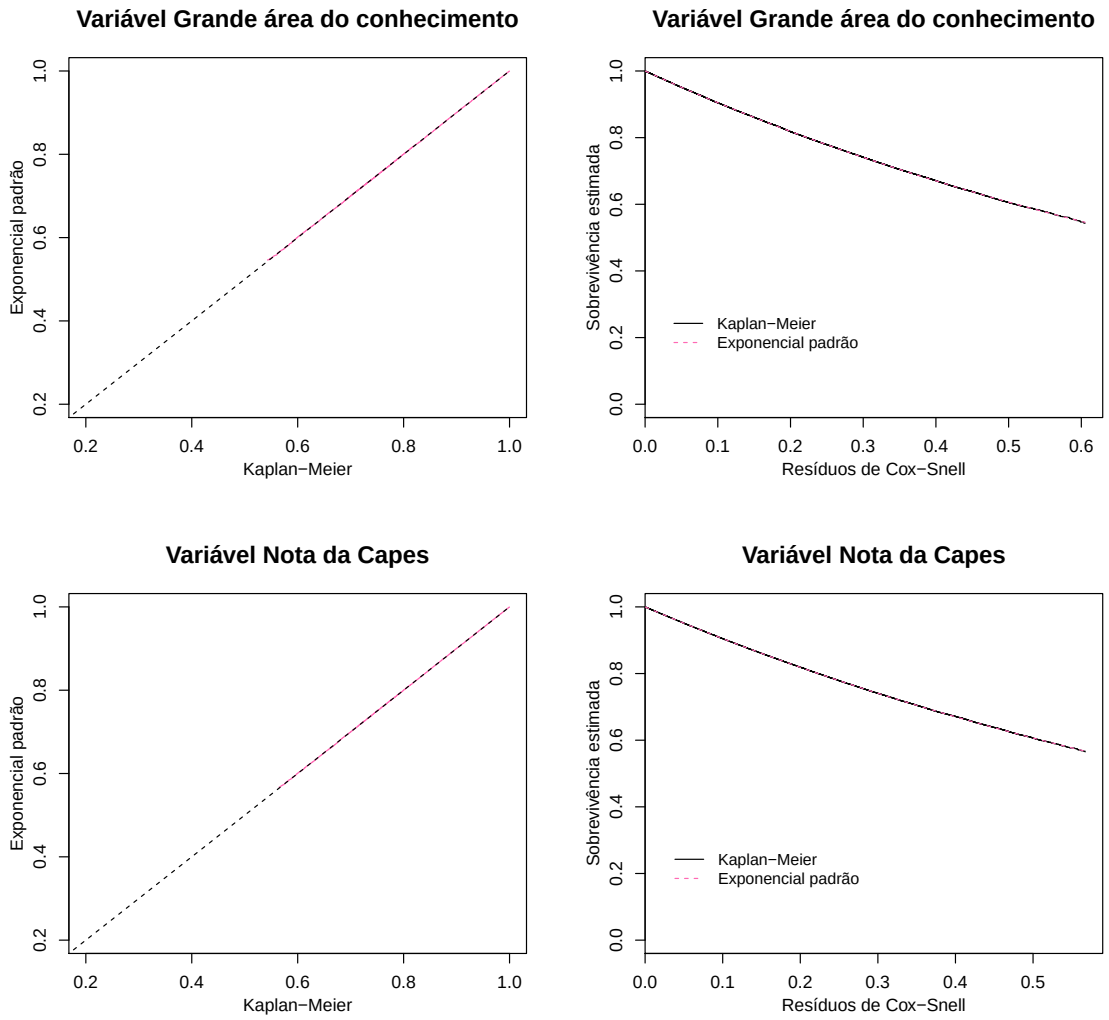

Figura 3.4: Comparação das funções de sobrevivência dos resíduos de Cox-Snell e da distribuição Exponencial para as variáveis Grande área do conhecimento e Nota da Capes

A partir daí, ajustou-se o modelo de regressão de Cox para 5 covariáveis: sexo, idade, região da IES, grande área do conhecimento e nota da Capes. Para o cálculo da sobrevivência, é necessária a estimativa da função de sobrevivência base $\hat{S}_0(t)$ para cada tempo t. A Tabela 3.3 mostra essas estimativas para os tempos a cada 30 dias devido o grande número de tempos distintos observados.

Tabela 3.3: Estimativas da função de sobrevivência base $\hat{S}_0(t)$ do modelo de Cox múltiplo

t	$\hat{S}_0(t)$	t	$\hat{S}_0(t)$	t	$\hat{S}_0(t)$
30	0,977	510	0,674	990	0,551
60	0,952	540	0,663	1020	0,545
90	0,920	570	0,655	1050	0,541
120	0,895	600	0,646	1080	0,536
150	0,871	630	0,636	1110	0,530
180	0,848	660	0,628	1140	0,524
210	0,825	690	0,621	1170	0,518
240	0,806	720	0,613	1200	0,513
270	0,786	750	0,606	1230	0,509
300	0,769	780	0,599	1260	0,504
330	0,754	810	0,590	1290	0,500
360	0,738	840	0,583	1320	0,495
390	0,723	870	0,577	1350	0,490
420	0,712	900	0,570	1380	0,484
450	0,698	930	0,563	1410	0,477
480	0,685	960	0,558	1440	0,470

A Tabela 3.4 traz as estimativas dos parâmetros do modelo de Cox ajustado para as covariáveis em conjunto, assim como a estimativa do erro relativo com intervalo de 95% de confiança.

Logo nota-se que todas as estimativas dos β's das categorias das variáveis região e nota da Capes não são significativas a um nível de significância de 5%. O mesmo ocorre para a estimativa de 3 categorias da variável grande área do conhecimento. No entanto, preferiu-se não ajustar outro modelo apenas para as covariáveis cujas estimativas das categorias fossem significativas e sim utilizar esse modelo apenas para realizar previsão e não para explicar algo por meio das covariáveis.

Assim, a função de sobrevivência estimada para um indivíduo com vetor de covariáveis $\mathbf{x} = (x_1, x_2, \ldots, x_5)'$ é dada por:

$$\hat{S}(t|\mathbf{x}) = [\hat{S}_0(t)]^{\exp\{\mathbf{x}'\hat{\boldsymbol{\beta}}\}}, \tag{3.1}$$

em que $\mathbf{x}'\hat{\boldsymbol{\beta}} = \hat{\beta}_{sexo} + \hat{\beta}_{idade} + \hat{\beta}_{regiao} + \hat{\beta}_{area} + \hat{\beta}_{nota}$.

Tabela 3.4: Estimativas dos coeficientes $\boldsymbol{\beta}$ do modelo de regressão de Cox ajustado para o conjunto das covariáveis

Variável	$\boldsymbol{\beta}$	$\boldsymbol{e^{\beta}}$ (I.C. 95%)	p
Sexo			
Masculino	0	1	-
Feminino	-0,06	0,94 (0,91;0,96)	<0,001
Idade			
Até 30 anos	0	1	-
31 a 35 anos	-0,19	0,82 (0,79;0,86)	<0,001
36 a 45 anos	-0,50	0,60 (0,57;0,63)	<0,001
Mais de 45 anos	-0,84	0,43 (0,41;0,46)	<0,001
Região			
Norte	0	1	-
Nordeste	-0,05	0,95 (0,84;1,08)	0,45
Sudeste	-0,00	1,00 (0,89;1,14)	0,95
Sul	0,09	1,09 (0,96;1,24)	0,16
Centro-Oeste	0,02	1,02 (0,88;1,17)	0,81
Grande área			
Ciências exatas e da terra	0	1	-
Ciências biológicas	-0,25	0,78 (0,73;0,83)	<0,001
Engenharias	0,15	1,16 (1,09;1,23)	<0,001
Ciências da saúde	-0,04	0,96 (0,90;1,02)	0,19
Ciências agrárias	-0,04	0,95 (0,90;1,02)	0,16
Ciências sociais aplicadas	0,17	1,19 (1,11;1,28)	<0,001
Ciências humanas	0,18	1,20 (1,13;1,28)	<0,001
Linguistíca, letras e artes	0,16	1,17 (1,08;1,27)	<0,001
Multidisplinar	0,04	1,05 (0,97;1,14)	0,24
Nota da Capes			
Nota 3	0	1	-
Nota 4	-0,01	0,99 (0,88;1,11)	0,83
Nota 5	-0,01	0,99 (0,88;1,11)	0,86
Nota 6	-0,01	0,99 (0,88;1,11)	0,90
Nota 7	-0,07	0,93 (0,82;1,05)	0,25

Nota: A categoria com $\boldsymbol{\beta} = 0$ é o nível de referência.

A partir das estimativas apresentadas nas Tabelas 3.3 e 3.4, pode ser calculada, por exemplo, a probabilidade de um indivíduo do sexo feminino, com até 30 anos, da região Sul, com título em ciências biológicas e em um programa de pós-graduação nota 7 permanecer mais de 1 ano sem ingressar em vículo formal de emprego para exercer atividade de caráter acadêmico. Será considerado um período de 360 dias pois é o valor mais próximo de 365 dias apresentado na Tabela 3.3.

Para isso, tem-se que $\mathbf{x}'\hat{\beta} = -0,06 + 0 + 0,09 - 0,25 - 0,07 = -0,29$ e $\hat{S}_0(360) =$

$0,738$. E assim, a função de sobrevivência estimada para um indivíduo com as características citadas é dada por:

$$\hat{S}(360|\mathbf{x}) = [\hat{S}_0(360)]^{\exp\{\mathbf{x}'\hat{\beta}\}} = 0,738^{\exp\{-0,29\}} = 0,8$$

A figura 3.5, apresenta a função de sobrevivência do resíduo de Cox-Snell para o modelo de Cox multivariado com a função de sobrevivência da distribuição exponencial com média 1. Pode-se observar que a função de sobrevivência dos resíduos se ajustou bem à função de sobrevivência exponencial padrão, indicando um bom ajuste do modelo de Cox múltiplo.

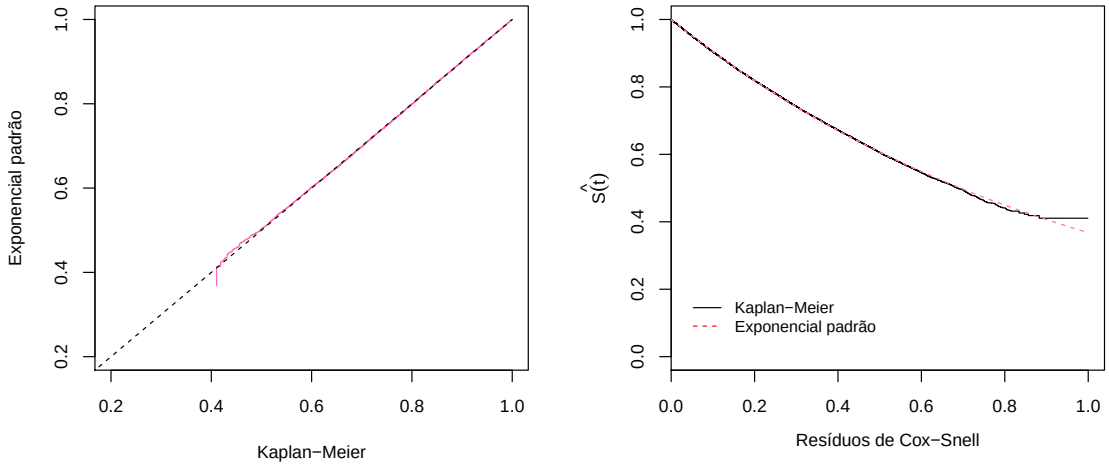

Figura 3.5: Comparação das funções de sobrevivência dos resíduos de Cox-Snell do modelo de Cox múltiplo e da distribuição Exponencial padrão.

A próxima seção apresenta um modelo de regressão logística para a chance de ingresso de doutores em vínculo empregatício para exercer atividades de caráter acadêmico em até um ano. O motivo da escolha do tempo até o ingresso será detalhado a seguir.

3.3 Modelo de regressão logística

Ainda para alcançar o objetivo da análise do tempo de ingresso em atividades de caráter acadêmico, foi proposto o ajuste do modelo de regressão logística aos dados, sendo consideradas as mesmas covariáveis do modelo de Cox (Seção 4.2) para efeito de comparação.

A partir daí, foram ajustados 5 modelos: para a chance do indivíduo estar empregado em até 6 meses, em até 1 ano, em até 2 anos, em até 3 anos e em até 4 anos a partir da data da titulação no doutorado. O intuito era o de comparar o complementar da probabilidade do evento de interesse ocorrer em até cada um desses anos, através do modelo logístico, e a sobrevivência estimada estimada pelo modelo de Cox.

A Figura 3.6 mostra o comportamento dos pontos referentes as estimativas da função de sobrevivência para 6 meses e para 1, 2, 3 e 4 anos versus a probabilidade de ingresso após cada um desses anos, estimada pelo modelo logístico.

Fica bem claro que o modelo logístico mostrado no gráfico após 1 ano (Figura 3.6) é o que mais coincide com o modelo de sobrevivência. E é curioso o fato de os modelos se distanciarem quanto maior o tempo. Ou seja, para esse estudo, caso o interesse seja a estimativa da probabilidade de ingresso em atividade de caráter acadêmico após 1 ano da titulação, qualquer um dos dois modelos apresentariam resultados similares.

O teste de Hosmer-Lemeshow foi utilizado para avaliar o ajuste dos modelos. Como mostra a Tabela 3.5, com exceção do modelo 1, nota-se que o valor-p cresce a medida que se aumenta o tempo considerado. Esse comportamento sugere que quanto maior o tempo considerado, maior é o ajuste do modelo logístico.

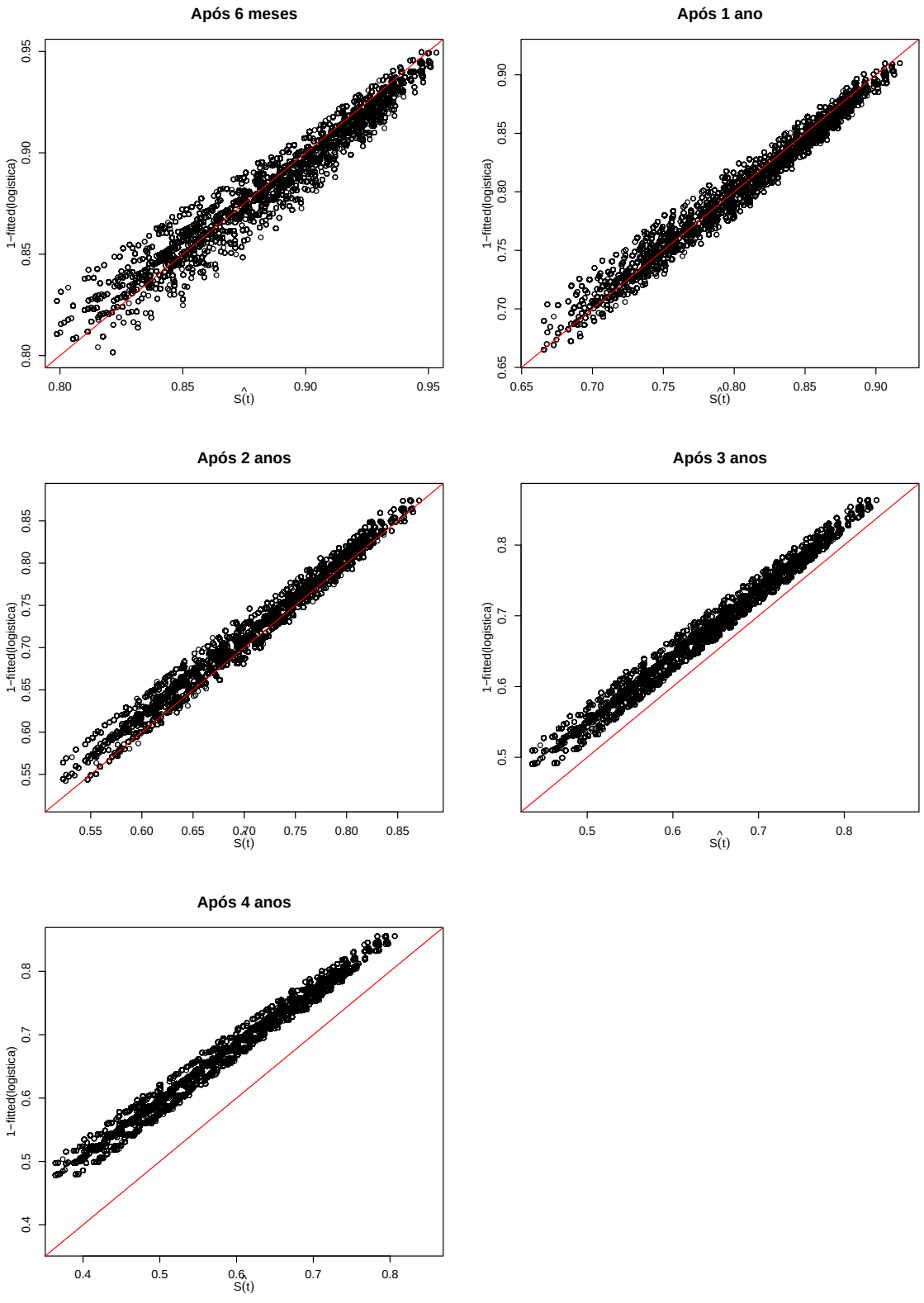

Figura 3.6: Gráficos das estimativas da função de sobrevivência do modelo de Cox versus valores ajustados do modelo logístico para probabilidade de ingresso em atividades de caráter acadêmico após 6 meses, 1 ano, 2 anos, 3 anos e 4 anos

Tabela 3.5: Teste de Hosmer-Lemeshow para ajuste do modelo

Modelo	Tempo	χ^2	p
1	Após 6 meses	9,26	0,32
2	Após 1 ano	12,99	0,11
3	Após 2 anos	9,39	0,31
4	Após 3 anos	7,04	0,53
5	Após 4 anos	4,66	0,79

Devido a aproximação entre os modelos apresentada na Figura 3.6, a Tabela 3.6 apresenta as estimativas dos coeficientes do modelo para a chance de ingresso no mercado de trabalho em até 1 ano e a estimativa da *odds ratio*.

É possível observar que, mantendo as outras covariáveis constantes, a chance de um indivíduo do sexo feminino ingressar em uma atividade de caráter acadêmico em até 1 ano é 92% da chance de um indivíduo do sexo masculino. Os doutores até 30 anos são os que possuem maior chance de ingresso e, com o aumento da idade, a chance vai diminuindo: Em relação aos doutores com 30 anos, os de 31 a 35 possuem 81% da chance e os de 36 a 45 possuem apenas 59%. Para os maiores de 45, a chance cai apenas para 42% do valor de referência.

Os titulados na região Sul são os que possuem maior chance de ingresso no tempo considerado, 8% maior do que a chance dos titulados na região Norte. Os titulados nas outras regiões possuem chance menor de ingresso do que esta última, que foi tomada como nível de referência, ressaltando que apenas a estimativa da região Centro-Oeste é significativa.

Tomando a grande área ciências exatas e da terra como referência, as áreas que possuem maior chance que ela são multidisciplinar, engenharias, ciências sociais aplicadas, linguística, letras e artes e ciências humanas, essa última possuindo chance 22% maior. As outras áreas, apesar de possuírem chance menor, não se distanciam muito da referência.

Surpreendentemente, no caso da variável nota da Capes, os indivíduos que titularam em programas nota 7 possuem 85% da chance de ingresso em até 1 do que os indivíduos que titularam em programas nota 3. Uma possível explicação é que talvez esses alunos continuem os estudos, seja como bolsista de curso de pós-doutorado ou outro.

Tabela 3.6: Estimativas dos coeficientes $\boldsymbol{\beta}$ do modelo de regressão logística para chance de ingresso em atividades de caráter acadêmico em até 1 ano

Variável	$\boldsymbol{\beta}$	$\boldsymbol{e^{\beta}}$	p
β_0	-0,91	0,40	<0,001
Sexo			
Masculino	0	1	-
Feminino	-0,08	0,92	<0,001
Idade			
Até 30 anos	0	1	-
31 a 35 anos	-0,21	0,81	<0,001
36 a 45 anos	-0,52	0,59	<0,001
Mais de 45 anos	-0,87	0,42	<0,001
Região			
Norte	0	1	-
Nordeste	-0,11	0,89	0,23
Sudeste	-0,02	0,98	0,74
Sul	0,08	1,08	0,40
Centro-Oeste	-0,06	0,94	<0,001
Grande área			
Ciências exatas e da terra	0	1	-
Ciências biológicas	-0,26	0,77	<0,001
Engenharias	0,13	1,14	<0,001
Ciências da saúde	-0,07	0,93	0,12
Ciências agrárias	-0,10	0,90	0,02
Ciências sociais aplicadas	0,13	1,14	0,01
Ciências humanas	0,20	1,22	<0,001
Linguistíca, letras e artes	0,18	1,19	<0,001
Multidisplinar	0,01	1,01	0,88
Nota da Capes			
Nota 3	0	1	-
Nota 4	-0,07	0,93	0,44
Nota 5	-0,08	0,92	0,36
Nota 6	-0,04	0,96	0,61
Nota 7	-0,16	0,85	0,08

Notas:A categoria com $\boldsymbol{\beta} = 0$ é o nível de referência.
β_0 é o intercepto do modelo.

A partir das estimativas apresentadas na Tabela 3.6, pode ser calculada, por exemplo, a probabilidade de um indivíduo do sexo feminino, com até 30 anos, da região Sul, com título em Ciências biológicas e em um programa de pós-graduação nota 7 ingressar em atividade de caráter acadêmico em até um ano. Para este indivíduo, o preditor linear é dado por $\mathbf{x}'\hat{\boldsymbol{\beta}} = -0,91 - 0,08 + 0 + 0,08 - 0,26 - 0,16 = -1,33$.

Assim, a probabilidade estimada segundo o modelo logístico é dada por:

$$p = \frac{e^{\mathbf{x}'\hat{\boldsymbol{\beta}}}}{1 + e^{\mathbf{x}'\hat{\boldsymbol{\beta}}}} = \frac{e^{-1,33}}{1 + e^{-1,33}} = 0,2092 \tag{3.2}$$

Assim, a probabilidade desse indivíduo permanecer mais de um ano sem ingressar em vínculo formal de emprego para exercer atividade de caráter acadêmico é estimada, através modelo logístico, como $1 - p = 0,7908$, que é próxima a estimativa dessa mesma probabilidade apresentada pelo modelo de Cox (0,8).

Capítulo 4

Conclusão

A partir dos resultados obtidos, pode-se dizer que o modelo de Cox é adequado para ajustar a distribuição do tempo entre a obtenção do título de doutorado no Brasil e o ingresso em um vínculo empregatício formal cuja atividade principal esteja ligada a área acadêmica através das 5 covariáveis selecionadas, principalmente com o objetivo de estimar a probabilidade de um doutor com um determinado perfil experimentar o evento de interesse.

O mesmo ocorreu para os modelos de regressão logística ajustados aos dados a partir das mesmas 5 covariáveis, tendo como um dos objetivo a comparação com o modelo de Cox. Foram ajustados cinco modelos que estimavam a chance de ingresso em um vínculo empregatício da mesma categoria descrita acima em até 1 ano, 2 anos, 3 anos e 4 anos. Foi possível, através do teste de Hosmer-Lemeshow, a conclusão de que os modelos se encontravam bem ajustados, posto que quanto mais tempo até o ingresso, menores as evidências de rejeição da suposição de bom ajuste.

O modelo de regressão logística que mais se aproxima do modelo de regressão de Cox, dentre os 5 ajustados, é o que estima a chance de ingresso no mercado de trabalho em até 1 ano. No entanto, a escolha por um deles depende do interessado no estudo tanto quanto considerar esse espaço de tempo ou qualquer outro, observado o tempo máximo observado. Isso porque, como mencionado, o modelo de Cox deve ser prioritariamente utilizado com o objetivo de se fazer estimativas de probabilidades e não como modelo explicativo. Pode-se notar, por meio da mudança dos valores dos coeficientes dos modelos de Cox simples e múltiplo (Tabelas 3.2 e 3.4) um possí-

vel problema de multicolinearidade, impossibilitando a interpretação das covariáveis como fatores de risco (de empregabilidade) no modelo de Cox múltiplo. Contudo, por ser um modelo não-paramétrico, o modelo de Cox possui um ajuste muito bom com as categorias das covariáveis. Já para o modelo logístico, a avaliação da qualidade do ajuste varia em cada um dos tempos considerados.

Evidencia-se aqui a boa qualidade dos dados observada nas bases de pós-graduação. Quase todos os campos encontravam-se preenchidos e não precisaram de tratamento extensivo. Como proposta futura, pode-se tentar o ajuste de um modelo paramétrico para que tempos superiores aos observados possam ser previstos. Ainda, um modelo com fração de cura pode ser ajustado, visando estimar a proporção de doutores que não ingressarão à carreira acadêmica.

Referências Bibliográficas

Agresti, A. (1990). *Categorical data analysis.* New York: John Wiley.

Brunello, G. H. V. & Nakano, E. Y. (2015). *Inferência bayesiana no modelo weibull discreto em dados com presença de censura.* TEMA - Tend. Mat. Apl. Comput., v.16, n.2, p.97-110.

Carrasco, C. G., Tutia, M. H., & Nakano, E. Y. (2012). Intervalos de confiança para os parâmetros do modelo geométrico com inflação de zeros. TEMA: Tendências em Matemática Aplicada e Computacional, v.13, n.3, p.247-255.

Casella, G.; Berger, R. (2010). *Inferência Estatística*, (2 ed.). Cengage Learning.

CGEE (2010). *Doutores 2010: estudos da demografia da base técnico-científica brasileira.* Brasília, DF.

CGEE (2012). *Mestres 2012: Estudos da demografia da base técnico-científica brasileira.* Brasília, DF.

CGEE (2013). Estatuto social centro de gestão e estudos estratégicos. http://www.cgee.org.br/arquivos/cgee_estatuto.pdf.

CGEE (2015). *Mestres e doutores 2015 - Estudos da demografia da base técnico-científica brasileira.* Brasília, DF.

CGEE (2017). *A formação de novos quadros para CT&I: avaliação do programa institucional de bolsas de iniciação científica (PIBIC).* Brasília, DF.

Colosimo, E. & Giolo, S. (2006). *Análise de sobrevivência aplicada.* ABE - Projeto Fisher. Edgard Blucher.

Cox, D. R. (1972). Regression model and life tables (with discussion). Journal Royal Statistical Society, B, 34, p.187-202.

Efron, B. (1977). The efficiency of cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72(359):557–565.

Ehlers, R. S. Inferência estatística - notas de aula. Departamento de Matemática Aplicada e Estatística da USP. Disponível em <http://www.icmc.usp.br/ ehlers/inf/inf.pdf>. Acesso em 27/04/2015.

Grambsch, P. M. & Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. Biometrika, 81, 3, 515-526.

Hosmer, D. W.; Lemeshow, S. (1999). *Applied Survival Analysis - Regression Modeling of Time to Event Data*, (1 ed.). Estados Unidos da América: John Wiley Sons, Inc.

Machado, A. R. (2015). Collection scoring via regressão logística e modelo de riscos proporcionais de cox. Dissertação (Mestrado em estatística). Universidade de Brasília.

Magalhães, M. N. (2006). *Probabilidade e Variáveis Aleatórias*, (2 ed.). EDUSP.

Maia, M. A. & Nakano, E. Y. (2016). *Análise do tempo até a re-hospitalização de pacientes com esquizofrenia via modelo de riscos proporcionais de Cox.* Semina: Ciências Exatas e Tecnológicas, v.37, n.2, p.51-58.

Marcuschi, L. A. (1996). Avaliação do programa institucional de bolsas de iniciação científica (pibic) do cnpq e propostas de ação. *Recife: UFPE.*

Matuda, N. S. (2005). *Fragilidade gama e variância robusta: extensões do modelo semiparamétrico de Cox.* PhD thesis.

McCULLAGH, P.; NELDER, J. (1989). *Generalized Linear Models*, (2 ed.). Londres: Chapman Hall. 532p.

MTb, SPPE, DES, CGET (2016). *Manual de Orientação da Relação Anual de Informações Sociais (RAIS): ano-base 2016.* Brasília, DF.

Nakano, E. Y. & Carrasco, C. G. (2006). Uma avaliação do uso de um modelo contínuo na análise de dados discretos de sobrevivência. Tema: Tendências em Matemática Aplicada e Computacional, v.7, n.1, p.91-100.

Nakano, E. Y. & Cunha, J. F. (2012). *Análise do efeito da camuflagem no tempo de segregação em regiões texturizadas utilizando o modelo de riscos proporcionais de Cox.* Semina: Ciências Exatas e Tecnológicas, v.33, n.2, p.141-148.

Santos, R. O. & Nakano, E. Y. (2015). *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.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. Biometrika, 69, 239-241.

Sicsú, A. L. (2010). *Credit Scoring: desenvolvimento, implantação, acompanhamento*. São Paulo: Blucher.

Silva, J. F., Liebano, R. E., Corrêa, J. B., Matsushita, R. Y., & Nakano, E. Y. (2017). *Análise do tempo para o alívio da intensidade da dor em pacientes com dor lombar crônica não específica via modelo de riscos proporcionais de Cox.* Ciência e Natura, v. 39, n.2, p.233-243.