Universidade de Brasília

Instituto de Ciências Exatas

Departamento de Estatística

Análise do tempo de permanência do trabalhador formal no mercado de trabalho no Distrito Federal

Rayany de Oliveira Santos

Brasília

2014

Rayany de Oliveira Santos

Bacharel em Estatística

Análise do tempo de permanência do trabalhador formal no mercado de trabalho no Distrito Federal

Relatório apresentado à disciplina Estágio Supervisionado II do curso de graduação em Estatística, Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Orientador: Prof. Dr. **Eduardo Yoshio Nakano**

Brasília

2014

Dedicatória

A Deus, que conhece meu medo, a minha felicidade e os meus sonhos. Conhece minha estrada e sabe exatamente o meu destino.

Aos meus queridos e amados pais, que me ensinaram a ser o que sou.

Agradecimentos

Agradeço ao meu Deus por me permitir transformar sonhos em realidade, por me sustentar, por ser meu refúgio, meu pai e amigo.

Ao Professor **Eduardo Yoshio Nakano** pelo desprendimento, compreensão e paciência. Agradeço por ter se mostrado sempre disponível e ter compartihado seu conhecimento que foi essencial para a realização desse trabalho.

A minha mãe, **Rogéria**, por muitas vezes tornar dela, os meus planos e anseios. Por me amar, me incentivar e nunca me deixar desanimar. Ao meu pai, **Juvenal**, que com sua firmeza nunca permitiu que eu desviasse do caminho certo a trilhar. Agradeço a minha irmã, **Débora**, por torcer pelo meu sucesso e depositar sua confiança em mim.

A minha prima **Eglay**, meus tios e avós por acreditarem na minha capacidade de ser uma boa profissional.

Um agradecimento ao **Emanuel Brasil**, que me estimula constantemente a acreditar que posso ser melhor, por apoiar minhas decisões e demonstrar tanto cuidado, amor e companheirismo.

Aos meus amigos **Mayva Luany**, **Bruno Wencelwski**, **Jessica Delavechia** e, em especial, meus amigos do colégio JK pela fidelidade durante todos esses anos. Aos meus amigos **Ana Luiza**, **Bianca Agapito**, **Lucas Silva**, **Marcos Lima** e todos os colegas que fiz durante o período da graduação.

Resumo

Análise do tempo de permanência do trabalhador formal no mercado de trabalho no Distrito Federal

Neste trabalho, o modelo probabilístico log-normal e o modelo de regressão de Cox foram propostos para analisar dados de sobrevivência relacionados a trabalhadores formais do Distrito Federal a fim de se verificar a influência de covariáveis que pudessem explicar o tempo permanência deles em um emprego. Os parâmetros dos modelos foram estimados através do método de Máxima Verossimilhança. O grande número de observações, que são parte da Relação Anual de Informações Sociais - RAIS, inviabilizou a realização de testes de hipóteses e técnicas gráficas foram as alternativas utilizadas para a tomada de decisões.

Palavras-chave: Análise de Sobrevivência; Modelos de regressão: log-normal e Cox; Máxima Verossimilhança; Relação Anual de Informações Sociais.

Sumário

Capítulo 1

INTRODUÇÃO

Segundo Outhwaite and Bottomore (1996), em seu sentido mais amplo, trabalho é o esforço humano dotado de um propósito e envolve a transformação da natureza através do dispêndio de capacidades mentais e físicas. Com o passar dos anos, a sociedade capitalista passou a inverter o propósito das ocupações que se qualificariam como trabalho e a definição foi limitada a ser sinônimo de emprego remunerado.

O trabalho é um dos principais vínculos entre o desenvolvimento econômico e o social, uma vez que representa um dos principais mecanismos por intermédio dos quais os seus benefícios podem efetivamente chegar às pessoas e, portanto, serem mais bem distribuídos.

Considerando-se um cenário atual, apesar da crise financeira internacional que afetou principalmente os Estados Unidos e a Europa no fim dos anos 2000 e veio a refletir nos países em desenvolvimento, o Brasil vem registrando grandes avanços na área trabalhista, como o crescimento expressivo do emprego formal, sobretudo nas regiões brasileiras mais pobres e com mercados de trabalho menos estruturados (OIT, 2012). Analogamente, o Distrito Federal é uma região do país cujo mercado de trabalho segue no mesmo ritmo aquecido. De acordo com a SETRAB-DF (2013),

a taxa de desemprego continua sendo a menor registrada desde 1992 no Distrito Federal.

Muitos são os desafios, no entanto, a serem enfrentados, relacionados principalmente a desigualdade (de gênero, raça e entre as regiões do país), para que as condições de trabalho no Brasil e no DF possam ser consideradas ideais, para que o quantitativo de pessoas desempregadas diminua cada vez mais e para que as causas dessa situação de inatividade sejam identificadas e políticas públicas sejam criadas com o intuito de até mesmo capacitar melhor o trabalhador. Bases de dados confiáveis são grandes aliadas no processo de encarar tão grande enfrentamento.

O governo brasileiro tem como importante apoio e insumo a Relação Anual de Informações Sociais - RAIS que é uma fonte de dados que tem grande potencial para assistí-lo no que diz respeito ao monitoramento, análise e avaliação do mercado formal de trabalho. É considerada um censo formal de trabalho, já que todos os estabelecimentos legalmente constituídos devem fornecer ao Ministério do Trabalho e Emprego (MTE) as informações referentes a cada um de seus empregados. Entretanto, há falhas no processo da declaração, o que tornam os métodos estatísticos meios eficazes para a análise da RAIS.

A Análise de Sobrevivência é uma área da Estatística que pode ser utilizada em diversas áreas do conhecimento, sendo particularmente importante em pesquisas de saúde. Também é muito utilizada na engenharia em que é conhecida como análise de confiabilidade. Ela avalia o tempo decorrido até a ocorrência de um evento ou situação de interesse e se caracteriza por utilizar a informação de todos os indivíduos

presentes no estudo, inclusive daqueles em que as observações estão incompletas. (Santos, 2013)

1.1 Objetivos

O objetivo geral do trabalho é analisar o tempo de permanência dos trabalhadores formais em seu emprego mais recente, desde a admissão até a demissão, no Distrito Federal a partir do ano 2002 até o ano 2009.

Os objetivos específicos são:

- Identificar quais fatores, tais como: idade do trabalhador, sexo, grau de instrução, raça e cor, podem influenciar o tempo de permanência no emprego;
- Aplicar métodos de Análise de Sobrevivência aos dados da RAIS utilizando o *software* R 3.1.0 (R CORE TEAM, 2013) e SPSS (Statistical Package for Social Sciences).

Capítulo 2

ANÁLISE DE SOBREVIVÊNCIA

2.1 Introdução

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.

2.2 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.

2.2.1 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 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.

2.2.2 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 in-

formações sobre o tempo de vida de indivíduos e a omissão deles pode fazer com que conclusões viciadas sejam feitas.

Alguns mecanismos de censura podem ser considerados, visto que são diversos os motivos para que ela aconteça, e são mostrados a seguir.

Censura à esquerda

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 aprendem a ler. As observações censuradas são caracterizadas pelas crianças que já sabiam ler e não lembravam com que idade isto tinha acontecido.

Censura intervalar

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.

Censura à direita

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 2.1 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.

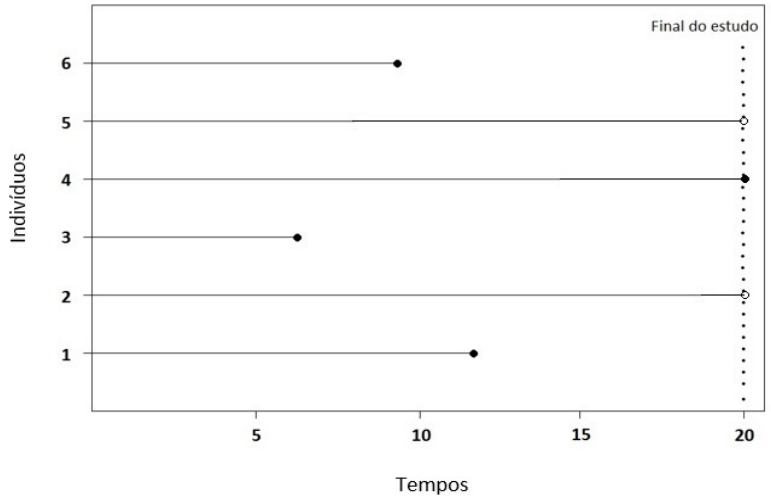

Figura 2.1: Dados com censura tipo I.

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.

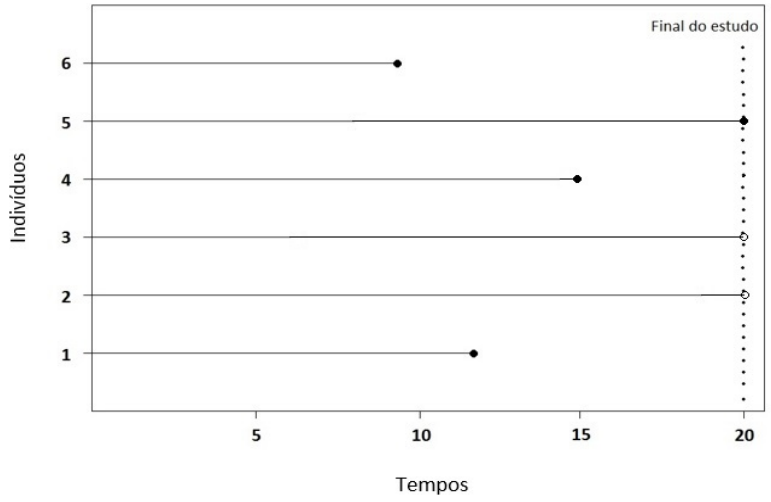

Figura 2.2: Dados com censura tipo II.

A Figura 2.2 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.

   A Figura 2.3 ilustra a censura aleatória. A falha é representada por • e a censura por ∘.

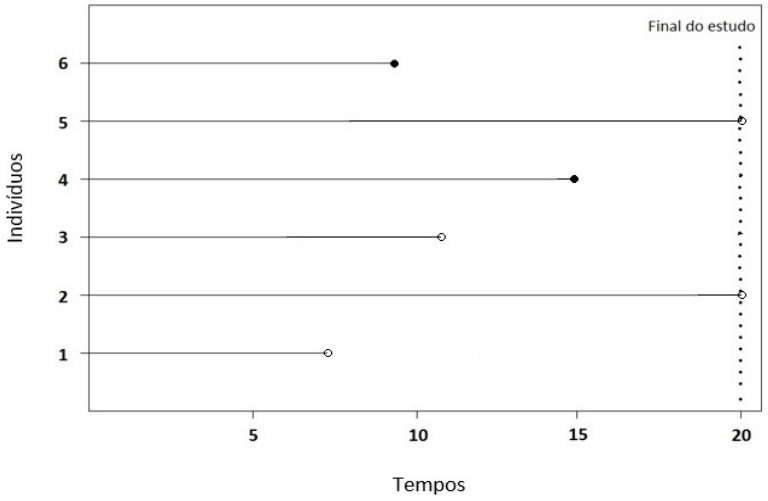

Figura 2.3: Dados com censura aleatória.

2.3 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 so-

brevivência, $S(t)$; 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 Nakano e Carrasco (2006) e Carrasco et al. (2012).

O tempo de sobrevivência, T_S, é dado por:

$$T_S = T_F - T_I,$$

oem 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 $(t_i, \delta_i, \mathbf{x}_i)$.

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

2.3.2 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:

$$F_T(t) = P(T \in (-\infty, t]) = P(T \leq t),$$

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, $F(t) = 1 - S(t)$, onde $S(t)$ representa a função de sobrevivência, descrita abaixo.

2.3.3 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 por (Colosimo e Giolo, 2006):

$$S(t) = P(T \geq t).$$

2.3.4 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:

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

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:

$$H(t) = \int_0^t h(u)du.$$

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

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

2.4 Técnicas Não-Paramétricas

Convencionalmente, a análise estatística descritiva de um estudo consiste na des-

crição 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.

2.4.1 O estimador de Kaplan-Meier

Também chamado de estimador limite-produto, 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)

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

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:

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

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.

2.5 Modelos Probabilísticos em Análise de Sobrevivência

Alguns modelos probabilísticos ou paramétricos, que são distribuições de probabilidade, são bastante adequados para descrever os tempos de vida de estudos em análise de sobrevivência. Entre os que ocupam uma posição de destaque estão o exponencial, o de Weibull e o log-normal.

2.5.1 Distribuição Exponencial

Por possuir apenas um único parâmetro e ter uma função de taxa de falha constante (propriedade chamada de falta de memória), a distribuição exponencial é uma das mais simples usadas para descrever a variável tempo até a falha. A função de densidade de probabilidade para a variável T é dada por:

$$f(t) = \frac{1}{\alpha}\ \exp\left\{-\left(\frac{t}{\alpha}\right)\right\},\ t \geq 0,$$

onde o parâmetro $\alpha > 0$ é o tempo médio de vida e tem a mesma unidade do tempo de falha t.

As funções de sobrevivência $S(t)$ e de taxa de falha $h(t)$ são dadas, respectivamente por:

$$S(t) = \exp\left\{-\left(\frac{t}{\alpha}\right)\right\}$$

e

$$h(t) = \frac{1}{\alpha},\ t \geq 0.$$

2.5.2 Distribuição de Weibull

A distribuição de Weibull é muito popular por possuir aplicabilidade em estudos biomédicos e industriais, além de apresentar uma grande variedade de formas com função de taxa de falha monótona.

As funções de densidade de probabilidade, de sobrevivência e de taxa de falha são dadas, respectivamente, por:

$$f(t) = \frac{\gamma}{\alpha^{\gamma}}\ t^{\gamma-1}\ \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\},\ t \geq 0,$$

$$S(t) = \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}$$

e

$$h(t) = \frac{\gamma}{\alpha^{\gamma}}\ t^{\gamma-1},$$

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

A função de risco $h(t)$ é estritamente crescente para $\gamma > 1$, estritamente decrescente quando $\gamma < 1$ e constante para $\gamma = 1$, que é a função de risco da distribuição exponencial, um caso particular da distribuição Weibull.

2.5.3 Distibuição Log-normal

A distribuição log-normal é bastante utilizada para descrever situações clínicas e caracterizar tempos de vida de produtos e indivíduos. A função de densidade de probabilidade é dada por:

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

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

As funções de sobrevivência e de risco de uma variável log-normal não apresentam uma forma analítica explícita e são representadas, respectivamente por:

$$S(t) = 1 - \Phi\left(\frac{\log(t)-\mu}{\sigma}\right)$$

e

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

em que $\Phi(.)$ é a função de distribuição acumulada de uma distribuição normal padrão.

Uma característica especial deste modelo é que a função de risco apresenta formas unimodais, isto é, o risco é inicialmente crescente e para grandes valores de T há um comportamento decrescente da função, como no caso de risco de óbito em recém-nascidos.

2.5.4 Seleção do Modelo Probabílistico

Um método eficaz para identificar uma distribuição apropriada para modelar a variável aleatória T é a curva do Tempo Total em Teste, também conhecida como curva TTT. A curva TTT é o gráfico da função $G(r/n)$ versus r/n, sendo $G(r/n)$ dada por:

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

em que $r = 1, ..., n$ e $T_{i:n}$, $i = 1, ..., r$ são as estatísticas de ordem da amostra.

Deve-se observar que as censuras não são consideradas no momento da construção do gráfico da curva, o que pode induzir a um erro de interpretação e de escolha da distribuição nos casos em que o número de censuras é grande.

A curva TTT pode apresentar várias formas, que devem ser associadas às funções de risco das distribuições que podem modelar a variável T. A Figura 2.4 (Neto et al., 2002) traz as formas que a curva pode assumir.

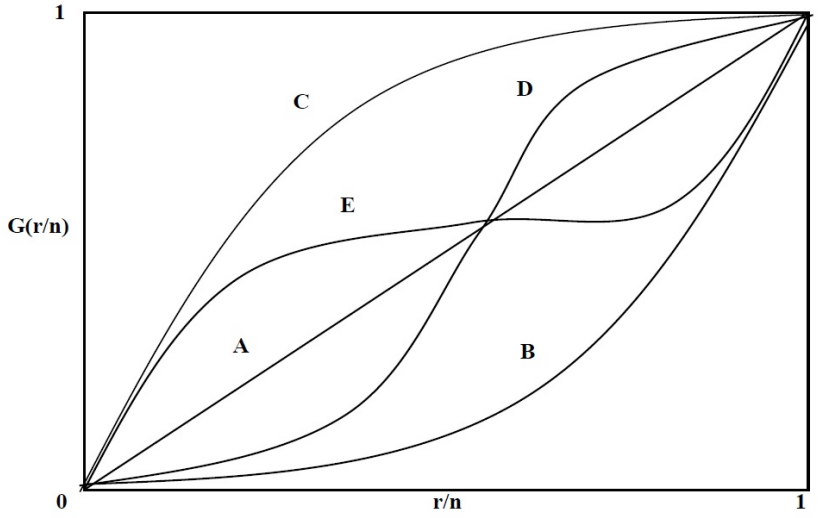

Figura 2.4: Formas da curva do Tempo Total em Teste - TTT.

Quando a curva referente aos dados estudados assume a forma de uma reta diagonal, como no modelo **A**, uma distribuição que possui a função de taxa de falha constante é adequada para modelar os dados. A distribuição exponencial tem função de risco constante para todo tempo de observação.

Quando a curva TTT assume a forma convexa **B** ou côncava **C**, as distribuições que possuem função de risco monotonicamente decrescente ou crescente, respectivamente, são adequadas para modelar os dados. A distribuição Weibull apresenta função de risco decrescente quando seu parâmetro de forma γ é menor que 1 e crescente quando γ é maior que 1.

Já quando a forma da curva é convexa e depois côncava, como no modelo **D**, as distribuições que possuem função taxa de falha com forma de **U**, conhecida como do tipo banheira, são as mais apropriadas. Elas são as modificações da distribuição Weibull: Weibull exponencializada, Weibull modificada, distribuição XTG, Weibull aditiva, entre outras.

Por último, quando a curva TTT tem forma côncava e depois convexa **E**, as distribuições apropriadas são as que possuem função de risco unimodal. Exemplos de distribuições com funções de risco com esse comportamento são a Log-Normal e Log-Logística.

2.6 Estimação dos Parâmetros dos Modelos

Os parâmetros dos modelos probabilísticos devem ser estimados a partir das observações da amostra. Devido pricipalmente a sua incapacidade de incorporar censuras no processo de estimação, o método dos mínimos quadrados, um dos mais

conhecidos na literatura estatística, não é apropriado para estudos de sobrevivência. Já o método de máxima verossimilhança permite incorporar as censuras e possui ótimas propriedades para grandes amostras.

2.6.1 O método de Máxima Verossimilhança

O método de Máxima Verossimilhança (Colosimo e Giolo, 2006) escolhe a distribuição, entre todas aquelas definidas pelos possíveis valores de seus parâmetros, com maior probabilidade de ter gerado a amostra observada, ou seja, a distribuição que melhor explica essa amostra. Em outras palavras, o objetivo do método é encontrar o valor de θ, um parâmetro genérico que pode estar representando um único parâmetro ou um conjunto de parâmetros, que maximiza a função de verossimilhança, $L(\theta)$, dada por:

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

em que $t_1, ..., t_n$ representam as observações de uma certa população.

No caso da análise de sobrevivência, as observações não censuradas da amostra contribuem para $L(\theta)$ com suas funções de densidade $f(t)$ e as observações censuradas contribuem com a função de sobrevivência $S(t)$. Assim, na análise de sobrevivência, as observações podem ser divididas em dois conjuntos: um com r observações não censuradas e outro com $n - r$ observações censuradas. A função de verossimilhança, considerando todos os mecanismos de censuras a direita, a menos de uma constante, é dada por:

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

ou equivalentemente por:

$$L(\theta) \propto \prod_{i=1}^{n} [f(t_i;\theta)]^{\delta_i} [S(t_i;\theta)]^{1-\delta_i} = \prod_{i=1}^{n} [h(t_i;\theta)]^{\delta_i} S(t_i;\theta),$$

em que δ_i é a variável indicadora de falha e $h(t)$ é a função de risco.

Os estimadores são encontrados a partir da resolução do sistema de equações:

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

em que $l(\theta) = logL(\theta)$ é dado por:

$$l(\theta) = \sum_{i=1}^{n} \{\delta_i \log[f(t_i;\theta)] + (1-\delta_i) \log[S(t_i;\theta)]\}.$$

2.7 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:

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

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:

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

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):

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

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:

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

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.

2.7.1 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:

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

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,

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

A função acima 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:

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

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. Uma desvantagem encontrada é que esta aproximação proposta por Breslow e Peto 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 não é tão utilizada como a de Breslow e Peto por requerer mais tempo e esforço computacional.

2.7.2 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):

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

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:

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

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:

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

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:

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

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

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

2.7.3 Adequação do Modelo de Cox

A suposição de taxas de falhas proporcionais no modelo de Cox pode ser avaliada através da análise dos resíduos de Schoenfeld. 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):

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

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:

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

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

Considerando $\beta(t) = \beta$ como uma forma alternativa de representar o modelo de Cox, temos que a suposição de taxas de falha proporcionais é válida se o gráfico de $\beta_q(t)$ versus t for uma linha horizontal. Inclinação zero mostra evidências a favor da proporcionalidade. Uma curva suavizada com bandas de confiança é adicionada ao gráfico para auxiliar o processo de detecção de alguma possível falha na proporcionalidade dos riscos.

Capítulo 3

RELAÇÃO ANUAL DE INFORMAÇÕES SOCIAIS - RAIS

A Relação Anual de Informações Sociais (RAIS) foi instituída pelo Decreto nº 76.900, de 23 de Dezembro de 1975 e 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 (MTE, 2014).

De acordo com o decreto, a RAIS identificará a empresa e o empregado, pelo número de inscrição no Cadastro Geral de Contribuintes - CGC e pelo número de inscrição no Programa de Integração Social ou no Programa de Formação do Patrimônio do Servidor Público - PIS/PASEP, respectivamente.

3.1 Declaração

Todo estabelecimento deve fornecer as informações requeridas de cada um de seus empregados para o Ministério do Trabalho e Emprego (MTE) através da Relação Anual de Informações - RAIS (MTE, 2012).

3.1.1 Quem deve declarar

1. inscritos no CNPJ com ou sem empregados - o estabelecimento que não possuiu empregados ou manteve suas atividades paralisadas durante o ano-base está obrigado a entregar a RAIS Negativa;

2. todos os empregadores, conforme definidos na CLT;

3. todas as pessoas jurídicas de direito privado, inclusive as empresas públicas domiciliadas no País, com registro, ou não, nas Juntas Comerciais, no Ministério da Fazenda, nas Secretarias de Finanças ou da Fazenda dos governos estaduais e nos cartórios de registro de pessoa jurídica;

4. empresas individuais, inclusive as que não possuem empregados;

5. cartórios extrajudiciais e consórcios de empresas;

6. empregadores urbanos pessoas físicas (autônomos e profissionais liberais) que mantiveram empregados no ano-base;

7. órgãos da administração direta e indireta dos governos federal, estadual ou municipal, inclusive as fundações supervisionadas e entidades criadas por lei, com atribuições de fiscalização do exercício das profissões liberais;

8. condomínios e sociedades civis;

9. empregadores rurais pessoas físicas que mantiveram empregados no ano-base;

10. filiais, agências, sucursais, representações ou quaisquer outras formas de entidades vinculadas à pessoa jurídica domiciliada no exterior.

3.1.2 Quem deve ser relacionado

1. empregados contratados por empregadores, pessoa física ou jurídica, sob o regime da CLT, por prazo indeterminado ou determinado, inclusive a título de experiência;

2. servidores da administração pública direta ou indireta, federal, estadual ou municipal, bem como das fundações supervisionadas;

3. trabalhadores avulsos (aqueles que prestam serviços de natureza urbana ou rural a diversas empresas, sem vínculo empregatício, com a intermediação obrigatória do órgão gestor de mão-de-obra, nos termos da Lei nº 8.630, de 25 de fevereiro de 1993, ou do sindicato da categoria);

4. empregados de cartórios extrajudiciais;

5. trabalhadores temporários, regidos pela Lei nº 6.019, de 3 de janeiro de 1974;

6. trabalhadores com Contrato de Trabalho por Prazo Determinado, regido pela Lei nº 9.601, de 21 de janeiro de 1998;

7. diretores sem vínculo empregatício, para os quais o estabelecimento/ entidade tenha optado pelo recolhimento do FGTS (Circular CEF nº 46, de 29 de março de 1995);

8. servidores públicos não-efetivos (demissíveis ad nutum ou admitidos por meio de legislação especial, não regidos pela CLT);

9. trabalhadores regidos pelo Estatuto do Trabalhador Rural (Lei nº 5.889, de 8 de junho de 1973);

10. aprendiz (maior de 14 anos e menor de 24 anos), contratado nos termos do art. 428 da CLT, regulamentado pelo Decreto nº 5.598, de 1º de dezembro de 2005;

11. trabalhadores com Contrato de Trabalho por Tempo Determinado, regido pela Lei nº 8.745, de 9 de dezembro de 1993, com a redação dada pela Lei nº 9.849, de 26 de outubro de 1999;

12. trabalhadores com Contrato de Trabalho por Prazo Determinado, regido por lei estadual;

13. trabalhadores com Contrato de Trabalho por Prazo Determinado, regido por lei municipal;

14. servidores e trabalhadores licenciados;

15. servidores públicos cedidos e requisitados;

16. dirigentes sindicais.

3.1.3 Quem não deve ser relacionado

1. diretores sem vínculo empregatício para os quais não é recolhido FGTS;

2. autônomos;

3. eventuais;

4. ocupantes de cargos eletivos (governadores, deputados, prefeitos, vereadores, etc.), a partir da data da posse, desde que não tenham feito opção pelos vencimentos do órgão de origem;

5. estagiários regidos pela Portaria MTPS nº 1.002, de 29 de setembro de 1967, e pela Lei nº 11.788, de 25 de setembro de 2008;

6. empregados domésticos regidos pela Lei nº 11.324/2006; e

7. cooperados ou cooperativados.

Capítulo 4

BASE DE DADOS

Cada base de dados anual da RAIS é constituída dos trabalhadores que foram admitidos em anos anteriores ou no próprio ano base e que foram demitidos no ano base ou não foram demitidos. Por exemplo, a base de dados do ano 2002 é formada pelos trabalhadores que foram admitidos até ou durante o ano 2002 e que foram demitidos durante esse mesmo ano ou não foram demitidos e continuaram na base.

Durante o passar dos anos, a RAIS vem sendo aprimorada e novas informações, que antes não eram objeto de interesse da relação, se tornaram parte dela. Contudo, mesmo com o avanço significativo da tecnologia, com o aumento do incentivo e também das penalidades para os empregadores declararem corretamente os dados referentes a seus empregados, os registros são entregues incompletos ou incorretos.

No presente estudo serão utilizadas as bases de dados dos anos 2002 a 2009 da RAIS, com informações apenas dos trabalhadores alocados na região do Distrito Federal. Neste capítulo serão apresentadas as variáveis que compõem as bases de dados e as modificações feitas para validar e corrigir os dados declarados.

4.1 Variáveis

A Relação Anual de Informações Sociais apresenta bases de dados grandes, tanto no que diz respeito a quantidade de variáveis quanto de indivíduos.

Uma observação importante a se fazer é que, apesar de novas variáveis terem sido criadas e incluídas nas bases com o passar dos anos, apenas as variáveis que estão presentes simultaneamente nas bases de 2002 a 2009 foram consideradas. Além disso, devido ao presente estudo ser objeto da análise de sobrevivência, somente as variáveis que continham informações do indivíduo anteriores a sua contratação foram mantidas. Por exemplo, a variável **CAUSA DE DESLIGAMENTO** foi excluída da base pois a informação relativa a ela só foi adquirida após a admissão do trabalhador.

A seguir, são apresentadas as variáveis que, após essa seleção, continuaram no estudo e uma breve descrição das mesmas. E logo após são apresentadas as variáveis que não permaneceram na base.

4.1.1 Variáveis que permaneceram na base

1. **CLASSE CNAE**: Classe da atividade econômica do estabelecimento informante, segundo a Classificação CNAE/95 (CNAE 1.0, revisada pela CONCLA em 2002).

2. **DATA DE ADMISSAO**: Data de admissão do trabalhador.

3. **DATA DE NASCIMENTO**: Data de nascimento do trabalhador.

4. **DIA DESLIGAMENTO**: Dia em que o trabalhador foi desligado do atual trabalho.

5. **GR INSTRUÇÃO**: Grau de instrução do trabalhador, categorizado em:

   (a) Analfabeto.

   (b) Até o 5º ano incompleto do Ensino Fundamental ou que se tenha alfabetizado sem ter frequentado escola regular.

   (c) 5º ano completo do Ensino Fundamental.

   (d) do 6º ao 9º ano do Ensino Fundamental incompleto.

   (e) Ensino Fundamental completo.

   (f) Ensino Médio incompleto.

   (g) Ensino Médio completo.

   (h) Educação Superior incompleta.

   (i) Educação Superior completa.

   (j) Mestrado completo.

   (k) Doutorado completo.

6. **IND CEI VINCULADO**: Indica se o empregado/servidor está ligado ao CEI (Cadastro Específico do INSS) vinculado. O estabalecimento declara ter CEI se possuir obra de construção civil.

7. **IND PAT**: Indicador de estabelecimento participante do PAT (Programa de Alimentação do Trabalhador).

8. **IND SIMPLES**: Indicador de estabelecimento optante pelo SIMPLES. O Sistema Integrado de Pagamento de Impostos e Contribuições das Microempresas e Empresas de Pequeno Porte (Simples) é um regime tributário diferenciado, simplificado e favorecido, aplicável às pessoas jurídicas consideradas como microempresas e empresas de pequeno porte.(SEF-SP)

9. **MES DESLIGAMENTO**: Mês em que o trabalhador foi desligado do atual trabalho.

10. **MUNICIPIO**: Município de localização do estabelecimento.

11. **NACIONALIDADE**: Nacionalidade do trabalhador.

12. **NAT JURIDICA**: Natureza jurídica da empresa. Categorizado em:

    (a) Administração Pública que inclui:

        i. Órgão Público do Poder Executivo Federal.

        ii. Órgão Público do Poder Executivo Estadual ou do Distrito Federal.

        iii. Órgão Público do Poder Executivo Municipal.

        iv. Órgão Público do Poder Legislativo Federal.

        v. Órgão Público do Poder Legislativo Estadual ou do Distrito Federal.

        vi. Órgão Público do Poder Legislativo Municipal.

        vii. Órgão Público do Poder Judiciário Federal.

        viii. Órgão Público do Poder Judiciário Estadual.

        ix. Autarquia Federal.

x. Autarquia Estadual ou do Distrito Federal.

xi. Autarquia Municipal.

xii. Fundação Federal.

xiii. Fundação Estadual ou do Distrito Federal.

xiv. Fundação Municipal.

xv. Órgão Público Autônomo Federal.

xvi. Órgão Público Autônomo Estadual ou do Distrito Federal.

xvii. Órgão Público Autônomo Municipal.

xviii. Comissão Polinacional.

xix. Fundo Público.

xx. Associação Pública.

(b) Entidades Empresariais

i. Empresa Pública

ii. Sociedade de Economia Mista

iii. Sociedade Anônima Aberta

iv. Sociedade Anônima Fechada

v. Sociedade Empresária Limitada Sociedade Empresária em Nome Coletivo

vi. Sociedade Empresária em Comandita Simples

vii. Sociedade Empresária em Comandita por Ações

viii. Sociedade em Conta de Participação

ix. Empresário (Individual)

x. Cooperativa

xi. Consórcio de Sociedades

xii. Grupo de Sociedades

xiii. Estabelecimento, no Brasil, de Sociedade Estrangeira

xiv. Estabelecimento, no Brasil, de Empresa Binacional Argentino-Brasileira

xv. Empresa Domiciliada no Exterior

xvi. Clube/Fundo de Investimento

xvii. Sociedade Simples Pura

xviii. Sociedade Simples Limitada

xix. Sociedade Simples em Nome Coletivo

xx. Sociedade Simples em Comandita Simples

xxi. Empresa Binacional

xxii. Consórcio de Empregadores

xxiii. Consórcio Simples

(c) Entidades sem Fins Lucrativos

i. Serviço Notarial e Registral (Cartório).

ii. Fundação Privada.

iii. Serviço Social Autônomo.

iv. Condomínio Edilício.

v. Comissão de Conciliação Prévia.

vi. Entidade de Mediação e Arbitragem.

vii. Partido Político.

viii. Entidade Sindical.

ix. Estabelecimento, no Brasil, de Fundação ou Associação Estrangeiras.

x. Fundação ou Associação Domiciliada no Exterior.

xi. Organização Religiosa.

xii. Comunidade Indígena.

xiii. Fundo Privado.

xiv. Associação Privada.

(d) Pessoas Físicas

i. Empresa Individual Imobiliária.

ii. Segurado Especial.

iii. Contribuinte individual.

iv. Candidato a Cargo Político Eletivo.

v. Leiloeiro.

(e) Instituições Extraterritoriais

i. Organização Internacional

ii. Representação Diplomática Estrangeira

iii. Outras Instituições Extraterritoriais

13. **PIS**: O PIS/PASEP, Programa de Integração Social e o Programa de Formação

do Patrimônio do Servidor Público, são contribuições sociais de natureza tributária utilizadas para constituir um fundo de ajuda ao trabalhador. O PIS é destinado aos funcionários de empresas privadas regidos pela Consolidação das Leis do Trabalho (CLT), enquanto o PASEP é destinado aos servidores públicos regidos pelo Regime jurídico estatutário federal. (INFOMONEY, 2005)

14. **PORT DEFICIENCIA**: Indica se o trabalhador possui deficiência.

15. **SEXO**: Sexo do trabalhador: Masculino ou feminino.

16. **TAMESTAB**: Tamanho do estabelecimento baseado no número de trabalhadores: Zero; Até 4; De 5 a 9; De 10 a 19; De 20 a 49; De 50 a 99; De 100 a 249; De 250 a 499; De 500 a 999; 1000 ou mais.

17. **TIPO SALARIO**: Tipo de salário do empregado/servidor, de acordo com o contrato de trabalho: Mensal, quinzenal, semanal, diário, horário, por tarefa, outros tipos.

4.1.2 Variáveis que não permaneceram na base

As seguintes variáveis foram excluídas da análise por:

- possuírem informações sobre os indivíduos que não foram fornecidas anteriormente ou no ato da contratação:

    1. **CAUSA DESLIGAMENTO**: Causa do desligamento do trabalhador.

- tratarem sobre o rendimento do trabalhador e a quantidade de horas trabalhadas. Houve grande dificuldade para padronizar os valores declarados a fim

de que se tornem comparáveis. Por exemplo, alguns indivíduos declararam o quanto recebem por tarefa enquanto outros declararam o quanto recebem por mês e outros, ainda, o quanto recebem por dia. Como não há maneira de relacionar esses valores, decidiu-se pela exclusão das seguintes variáveis:

1. **HORAS CONTRATUAIS**: Quantidade de horas contratuais por semana.
2. **REM MEDIA (R$)**: Remuneração média do trabalhador (valor nominal).
3. **REM MEDIA SM**: Remuneração média do ano em salários mínimos (quando acumulada representa massa salarial).
4. **REM DEZEMBRO (R$)**: Remuneração do trabalhador em dezembro (valor nominal).
5. **REM DEZEMBRO**: Remuneração de dezembro em salários mínimos (quando acumulada representa massa salarial)
6. **SALARIO CONTRATUAL (R$)**: Salário Contratual do trabalhador (valor nominal).
7. **ULTIMA REM (R$)**: Última Remuneração do trabalhador (valor nominal).

- conterem informações cadastrais e pessoais dos empregados e empresas:

1. **CEI VINCULADO**: número do CEI vinculado do estabelecimento.
2. **CPF**: CPF do trabalhador.

3. **IDENTIFICADOR (CNPJ OU CEI)**: Identificador do estabelecimento.

- trazerem informações já identificadas em outras variáveis:

  1. **OCUPAÇÃO**: Classificação Brasileira de Ocupações criada em 1994 e abrange categorias como químico, fisíco, médico, etc. A variável **CLAS CNAE** já traz informações sobre qual área pertence o emprego do indivíduo.

  2. **TIPO ESTAB**: Tipo de estabelecimento: CNPJ ou CEI. As empresas/entidades que possuírem CNPJ e CEI, simultaneamente, devem informar na declaração somente o CNPJ. Essa variável traz praticamente a mesma informação da variável **IND CEI VINCULADO**.

- não existir a possibilidade de se encontrar a informação referente a variável. Nesse caso, a variável **TIPO ADMISSAO** só mostra dados referentes aos indivíduos admitidos no ano base. Os indivíduos que estão na base de um certo ano, porém não foram admitidos naquele ano, não têm informação válida para fins do estudo.

4.2 Validação e correção dos dados

Como já mencionado, a RAIS enfrenta o problema dos empregadores que declaram informações erradas ou incompletas dos seus empregados. Para que as técnicas e métodos de Análise de Sobrevivência pudessem ser aplicados, algumas modificações

nas bases de dados foram realizadas, como por exemplo a criação de novas variáveis, descritas a seguir.

4.2.1 PIS/PASEP inválidos

O PIS/PASEP é um número cadastrado de onze dígitos e possui o formato:

$$X_1X_2X_3X_4X_5X_6X_7X_8X_9X_{10} - Y,$$

em que $i = 1, 2, \ldots, 10$ mostra qual a posição do dígito e Y é o dígito verificador que é calculado através dos seguintes passos:

Primeiramente, soma-se o produto dos dígitos com os números mostrados a seguir.

$$S = (X_1.3)+(X_2.2)+(X_3.9)+(X_4.8)+(X_5.7)+(X_6.6)+(X_7.5)+(X_8.4)+(X_9.3)+(X_{10}.2)$$

Posteriormente, encontra-se a diferença entre 11 e o resto da divisão de S pelo número 11, denotada abaixo.

$$D = 11 - mod\left(\frac{S}{11}\right)$$

Se $D = 11$ ou $D = 10$, o dígito verificador Y é igual a 0. Quando $0 \leq D < 10$, Y assume o valor de D.

Nas bases de dados analisadas, foram encontrados vários PIS/PASEP que não eram válidos, ou seja, o valor do dígito verificador informado não era o mesmo encontrado ao se realizar o cálculo acima. Logo, conclui-se que, por algum motivo, as empresas informaram o número erroneamente.

A solução encontrada para esse problema foi a de selecionar apenas os indivíduos que possuiam o PIS informado válido. Os outros deixaram de fazer parte do estudo.

4.2.2 Seleção dos trabalhadores do DF a partir do ano 2002

Nas bases de dados, a variável **MUNICIPIO** representa o município de localização do estabelecimento. Como o estudo está interessado no mercado de trabalho do Distrito Federal, selecionou-se apenas os municípios cujos códigos se iniciam pelo número 53, que abrangem Brasília e outros.

Foi realizado também um truncamento nas bases, que resultou na exclusão dos trabalhadores admitidos antes do ano 2002.

4.2.3 Criação de chaves identificadoras

Para identificar individualmente cada um dos trabalhadores, foi criada uma chave que é formada pela concatenação das variáveis **PIS** e **DIASNASC**, sendo que essa última foi criada e é calculada pela quantidade de dias existente entre a data 14 de Outubro de 1582 (primeiro dia do calendário Gregoriano (IBM)), que é uma data base assumida pelo *software* SPSS, e a data de nascimento do empregado.

4.2.4 Seleção do emprego mais recente do trabalhador

Após as modificações citadas acima, uniu-se as bases dos 8 anos e identificou-se as chaves repetidas. Foram observados casos de trabalhadores que só foram admitidos uma única vez após o ano 2002 e trabalhadores que foram admitidos mais de uma vez. Para fins do estudo, selecionou-se apenas o último emprego do trabalhador, sendo considerados tanto os casos quando aconteceu a demissão quanto quando não aconteceu. Assim, cada chave que identifica os indivíduos está relacionada a apenas um emprego.

4.2.5 Criação da data de demissão

Para a criação da variável que denomina a data de demissão uniu-se as variáveis **DIA DESLIGAMENTO**, **MES DESLIGAMENTO** e **ANO BASE** que representam o dia que o trabalhador foi demitido, o mês que ele foi demitido e o ano em que o trabalhador foi declarado, respectivamente, sendo que a variável **ANO BASE** foi criada.

A variável **MES DESLIGAMENTO** estava presente em todas as bases de dados e, para que o tempo de sobrevivência dos trabalhadores fosse calculado, ela teve que ser recodificada. Nos casos em que o indivíduo foi censurado, a variável apresentava valor igual a 0 e passou a ser 12, o que representa a situação do trabalhador não ter sido desligado até o último mês do ano base.

A variável **DIA DESLIGAMENTO** também apresentava valor igual a 0 nos casos em que o indivíduo foi censurado e passou a ter valor igual a 31. Logo, um trabalhador que não falhou possui a data de demissão igual a 31/12/AAAA, em que AAAA simboliza aqui o ano base.

A base do ano 2002 não apresentava originalmente a variável **DIA DESLIGAMENTO** e teve que ser recodificada tanto nos casos de falha quanto de censura. Quando o indivíduo era censurado, possuia o mês de desligamento igual a 0 e passou a ter o dia de desligamento igual a zero, que em seguida passou a ser igual a 31, da mesma forma que aconteceu nas bases dos outros anos. Quando o indivíduo era desligado, passava a ter o dia de desligamento igual ao último dia do mês de desligamento. Por exemplo, se o mês de desligamento era fevereiro, o dia de desligamento

assumido foi 28 pois 2002 não é ano bissexto.

4.2.6 Cálculo do tempo de sobrevivência

Para a aplicação das técnicas de Análise de Sobrevivência, foi criada a variável **TEMPO** que denota o tempo de sobrevivência, ou seja, o tempo compreendido entre a data de admissão e a data de demissão do trabalhador. O resultado é calculado pela diferença entre as datas de demissão e admissão e é dado em dias.

4.2.7 Criação da variável indicadora de falha ou censura

Foi criada também a variável **STATUS** que indica se o indivíduo experimentou o evento de interesse, a demissão, ou se foi censurado, podendo não ter sido mais acompanhado durante os anos por algum motivo, como a falta de declaração, ou não ter sido demitido até dia 31 de dezembro de 2009. Ela é denotada por δ_i, expressa por:

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

4.2.8 Recodificação da variável CLASSE CNAE

Como já mencionado anteriormente, a variável **CLASSE CNAE** representa a classe da atividade econômica do estabelecimento. Com isso, precisou-se recodificar essa variável devido a quantidade de áreas econômicas que podem ser declaradas pelas empresas. A recodificação foi feita com base em um documento feito pelo IBGE chamado Classificação Nacional de Atividades Econômicas - Fiscal e, no nível mais agregado, as categorias individuais da CNAE estão organizadas em 17 seções, discriminadas na Tabela 4.1.

Tabela 4.1: Recodificação Parcial de **CLASSE CNAE**.

SEÇÃO	DIVISÕES	DESCRIÇÃO CNAE
A	01,02	Agricultura, pecuária, silvicultura e exploração florestal
B	05	Pesca
C	10,11,13,14	Indústrias extrativas
D	15 a 33	Indústrias de transformação
E	40,41	Produção e distribuição de eletricidade, gás e água
F	45	Construção
G	50,51,52	Comércio; reparação de veículos automotores, objetos pessoais e domésticos
H	55	Alojamento e alimentação
I	60,61,62,63,64	Transporte, armazenagem e comunicações
J	65,66,67	Intermediação financeira, seguros, previdência complementar e serviços relacionados
K	70,71,72,73,74	Atividades imobiliárias, aluguéis e serviços prestados às empresas
L	75	Administração pública, defesa e seguridade social
M	80	Educação
N	85	Saúde e serviços sociais
O	90,91,92,93	Outros serviços coletivos, sociais e pessoais
P	95	Serviços domésticos
Q	99	Organismos internacionais e outras instituições extraterritoriais

Apesar do número de categorias ter diminuído bastante, essa ainda é uma grande quantidade no que se diz respeito a análise de sobrevivência. Assim, realizou-se a regressão de riscos proporcionais de Cox apenas entre as 17 categorias da variável **CLASSE CNAE** para agregar os dados em menos níveis, através do valor de β, mas que ainda assim sejam semelhantes entre si. O nível de referência utilizado foi a seção Q, referente a *Organismos internacionais e outras instituições extraterritoriais*, que assume β igual a zero.

O modelo de Cox foi escolhido aqui, ao invés do Log-normal, por possuir um conjunto menor de suposições.

A Tabela 4.2 mostra as seções e seus respectivos β's ordenados, assim como a

qual categoria passarão a fazer parte. Observa-se que as novas categorias foram criadas com base nos β's que possuem valores próximos.

Tabela 4.2: Recodificação Final de **CLASSE CNAE**.

SEÇÃO	BETA	NOVA CATEGORIA
A	1,114	1
F	1,113	1
H	0,858	2
G	0,739	2
D	0,607	2
C	0,596	2
K	0,578	2
O	0,507	2
P	0,409	3
N	0,321	3
I	0,298	3
B	0,277	3
M	0,249	3
Q	0	3
J	-0,266	4
E	-0,454	4
L	-0,843	5

Assim, foram criadas 5 novas categorias para a variável **CLASSE CNAE** que serão chamadas a partir daqui de Categoria 1, Categoria 2, Categoria 3, Categoria 4 e Categoria 5.

4.2.9 Identificação da idade do trabalhador

A idade do trabalhador foi calculada pela diferença entre a data de admissão e a data de nascimento, em anos. Outro ajuste, relacionado a exclusão de alguns trabalhadores devido a erro de declaração, foi realizado. Como já mencionado na seção *Quem deve ser relacionado* do Capítulo 3, os aprendizes (maiores de 14 anos e menores de 24 anos) entram nas bases de dados, logo só foram considerados os trabalhadores maiores de 14 anos.

Tomando como base as faixas de idade utilizadas nas pirâmides etárias pelo IBGE, exceto pelas modificações que foram incluir a idade 14 anos na faixa de 15 a 19 anos e agregar as idades maiores ou iguais a 60 anos, devido as pequenas frequências observadas nessas categorias, a variável que representa a idade do trabalhador no momento da contratação foi categorizada nas seguintes faixas: 14 a 19 anos, 20 a 24 anos, 25 a 29 anos, 30 a 34 anos, 35 a 39 anos, 40 a 44 anos, 45 a 49 anos, 50 a 54 anos, 55 a 59 anos, 60 a 64 anos, 65 anos ou mais.

Assim como no caso da variável **CLASSE CNAE**, foi utilizada a regressão de riscos proporcionais de Cox entre as categorias da variável **IDADE** para se alcançar um número menor de categorias. A Tabela 4.3 mostra a primeira recodificação ordenada, em razão da variável ser ordinal, seus respectivos β's e a recodificação final.

Tabela 4.3: Recodificação Final de **IDADE**.

CATEGORIA ANTIGA	BETA	NOVA CATEGORIA
14 a 19 anos	-0,545	1
20 a 24 anos	-,046	2
25 a 29 anos	-,093	2
30 a 34 anos	-,127	3
35 a 39 anos	-,142	3
40 a 44 anos	-,148	3
45 a 49 anos	-,140	3
50 a 54 anos	-,138	3
55 a 59 anos	-,074	4
60 a 64 anos	-,018	4
65 anos ou mais	0	4

Assim a primeira categoria continua sendo chamada de *14 a 19 anos*, a segunda passa a ser *20 a 29 anos*, a terceira passa a ser *30 a 54 anos* e a quarta fica sendo *55 anos ou mais*.

4.2.10 Recodificação da variável GR INSTRUÇÃO

A variável **GR INSTRUÇÃO**, que mostra qual o Grau de instrução do trabalhador, foi recodificada através da tentativa em diminuir o número de categorias, sendo utilizado apenas o critério em unir categorias próximas e com níveis parecidos. A Tabela 4.4 mostra a recodificação final.

Tabela 4.4: Recodificação Final de **GR INSTRUÇÃO**.

CATEGORIA ANTIGA	NOVA CATEGORIA
Analfabeto	1
Até o 5º ano incompleto do Ensino Fundamental	2
5º ano incompleto do Ens. Fundamental	2
Do 6º ao 9º ano do Ens. Fundamental incompleto	2
Ensino Fundamental Completo	2
Ensino Médio Incompleto	3
Ensino Médio Completo	3
Ensino Superior Incompleto	4
Ensino Superior Completo	4
Mestrado	5
Doutorado	5

A primeira categoria continua a ser chamada de Analfabeto, a segunda passou a ser Ensino Fundamental - Completo e Incompleto, a terceira passou a ser Ensino Médio - Completo e Incompleto, a quarta passou a ser Ensino Superior - Completo e Incompleto e a quinta se tornou Mestrado ou Doutorado.

4.2.11 Recodificação da variável NACIONALIDADE

Durante o processo de declaração, a empresa tem a opção de escolher a nacionalidade do trabalhador entre 23 categorias disponíveis, como *brasileira*, *argentina*, *coreana*, inclusive *entre outras*. Devido a baixas frequências encontradas, decidiu-se diferenciar apenas a nacionalidade brasileira das outras.

4.2.12 Recodificação da variável TAMANHO ESTAB

Como já descrito na Seção *Variáveis*, a variável **TAMANHO ESTAB** retrata o tamanho do estabelecimento a partir da quantidade de funcionários. A Tabela 4.5 mostra a recodificação dessa variável que foi feita baseada nos valores de β obtidos pela regressão de Cox.

Tabela 4.5: Recodificação Final de **TAMANHO ESTAB**.

CATEGORIA ANTIGA	BETA	NOVA CATEGORIA
Zero	1,736	Zero
Até 4	1,074	Até 249
De 5 a 9	1,061	Até 249
De 10 a 19	1,065	Até 249
De 20 a 49	1,026	Até 249
De 50 a 99	0,996	Até 249
De 100 a 249	0,885	Até 249
De 250 a 499	0,678	250 ou mais
De 500 a 999	0,572	250 ou mais
De 1000 ou mais	0	250 ou mais

Assim, 3 novas categorias foram criadas: Zero, Até 249 e 250 ou mais.

4.2.13 Recodificação da variável TIPO SALARIO

Devido a pequena frequência encontrada nos tipos de salário *quinzenal*, *semanal*, *diário*, *horário*, *por tarefa* e *outros tipos*, decidiu-se por uni-los. Assim, a variável passou a ser categorizada como Mensal e Outros tipos.

Capítulo 5

RESULTADOS

A análise dos dados será feita primeiramente através de uma análise descritiva. Logo após serão feitas uma análise não-paramétrica e uma análise paramétrica na tentativa de encontrar modelos que representem bem o comportamento dos dados.

Uma observação importante a se fazer é que, devido a grande quantidade de obsevações na base de dados (1.645.284 indivíduos), não é viável a realização de testes de hipóteses pois em todos eles, a hipótese nula seria rejeitada. Com isso, as decisões para chegar em resultados foram tomadas, em grande parte, baseadas em técnicas gráficas.

5.1 Análise descritiva dos dados

A base de dados utilizada para a análise, após as recodificações das variáveis, possui 1.645.959 indivíduos, visto que foram considerados apenas os trabalhadores do Distrito Federal que começaram a trabalhar entre os anos de 2002 e 2009 e foram demitidos até 31 de dezembro de 2009 ou não foram demitidos, ou seja, o acompanhamento deles só foi feito até essa última data. Entretanto, percebeu-se que 675 deles tinham tempo, considerando falhas ou censuras, igual a 0 dias e

adotou-se o critério de não se admitir essa situação. Logo, a base passou a ter 1.645.284 trabalhadores. O menor tempo observado passou a ser 1 dia e o maior tempo encontrado foi 2.921 dias, que representam aproximadamente 8 anos.

Como já exposto no capítulo que trata da metodogia do presente estudo, a análise descritiva tradicional não pode ser aqui utilizada em razão da presença de censuras. Com isso, foi construído o gráfico da curva de Kaplan-Meier sem considerar nenhuma covariável, que é apresentado a seguir na Figura 5.1.

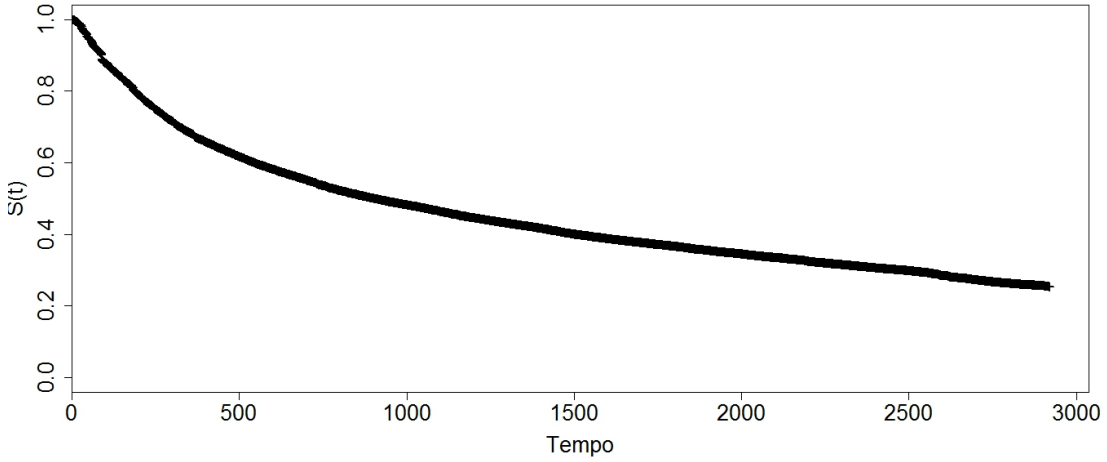

Figura 5.1: Curva estimada pelo método não-paramétrico de Kaplan-Meier para os tempos de sobrevivência dos trabalhadores do DF.

Destaca-se a quantidade de censuras que pode ser observada na Figura 5.1: 53,9% dos tempos são censurados enquanto 46,1% deles são tempos de falha.

A seguir, é mostrada na Tabela 5.1 as frequências relativas às covariáveis que permaneceram na base de dados, já sendo consideradas as recodificações finais citadas no Capítulo 4.

Tabela 5.1: Frequências absolutas e relativas das covariáveis.

VARIÁVEL	FREQ. ABSOLUTA	FREQ. RELATIVA
CLASSE CNAE		
Categoria 1	172.621	10,49%
Categoria 2	889.139	54,04%
Categoria 3	180.666	10,98%
Categoria 4	45.027	2,73%
Categoria 5	357.828	21,74%
GR INSTRUÇÃO		
Analfabeto	6.672	0,4%
Ens. Fundamental	511.225	31,07%
Ens. Médio 3	782.589	47,57%
Ens. Superior 4	339.279	20,62%
Mestrado/Doutorado 5	5.219	0,32%
IDADE		
14 a 19 anos	242.736	14,75%
20 a 29 anos	745.759	45,33%
30 a 54 anos	617.860	37,55%
mais de 55 anos	38.929	2,37%
IND CEI VINCULADO		
Não	1.582.262	96,17%
Sim	63.022	3,83%
IND PAT		
Não	1.059.462	64,39%
Sim	585.822	35,61%
IND SIMPLES		
Não	1.303.198	79,21%
Sim	342.086	20,79%
NACIONALIDADE		
Brasileira	1.643.322	99,88%
Outra	1.962	0,12%
NAT JURIDICA		
Administração Pública	361.911	21,99%
Entidades Empresariais	1.171.355	71,19%
Entidades sem fins lucrativos	96.465	5,86%
Pessoas Físicas	14.637	0,89%
Instituições Extraterritoriais	914	0,05%
PORT DEFICIENCIA		
Sim	12.659	0,77%
Não	1.632.625	99,23%
SEXO		
Masculino	1.059.086	64,37%
Feminino	586.198	35,63%
TAMANHO ESTAB		
0 funcionários	52.073	3,16%
1 a 249 funcionários	882.403	53,63%
250 ou mais funcionários	710.808	43,20%
TIPO SALÁRIO		
Mensal	1.586.962	96,45%
Outro	58.322	3,54%

[1]As categorias da variável **CLASSE CNAE** estão especificadas no Capítulo 4 na seção *Validação e correção dos dados*, na subseção *Recodificação da variável CLASSE CNAE*.

Através da Tabela 5.1, percebe-se que mais da metade dos trabalhadores estão na Categoria 2 da CNAE, ou seja, tem o trabalho relacionado a Alojamento e alimentação ou a Comércio, reparação de veículos automotores, objetos pessoais e domésticos ou a Indústria de Transformação ou a Indústrias Extrativas ou a Atividades imobiliárias ou outros serviços coletivos, sociais e pessoais. A Categoria 4, que abrange as áreas de Intermediação financeira, seguros, previdência complementar e Produção e distribuição de eletricidade, gás e água, é a que menos possui trabalhadores.

No que se refere ao grau de instrução dos indivíduos, poucos são os analfabetos, que representam 0,4% do total. 47,57% possuem o Ensino médio completo ou pelo menos chegaram a ingressar nesse nível de ensino. Apenas 0,32% das pessoas tem mestrado ou doutorado.

Nota-se que 64,37% dos trabalhadores são homens, 45,33% têm entre 20 e 29 anos e quase 100% deles são brasileiros. O número de deficientes é menor que 1%. A grande maioria recebe salário mensal.

Em relação as empresas, o que mais chama a atenção é que 71,19% delas são entidades empresariais e 35,61% participam do PAT.

A seguir são apresentados os gráficos das curvas estimadas por Kaplan-Meier das covariáveis para que as comparações entre suas categorias sejam feitas.

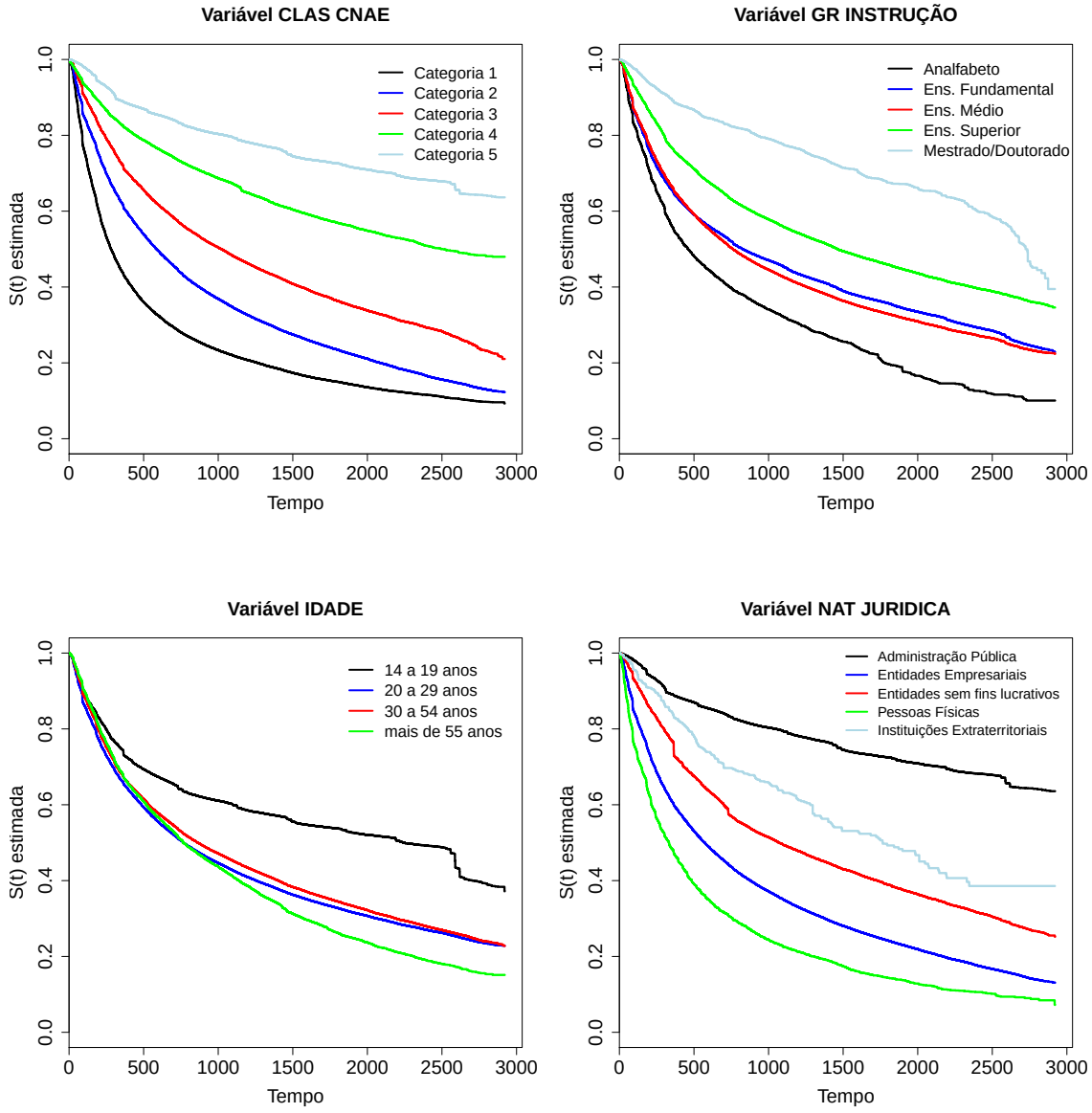

Figura 5.3: Curvas de sobrevivência das covariáveis CLASSE CNAE, GR INSTRUÇÃO, IDADE e NAT JURÍDICA estimadas por Kaplan-Meier.

A partir da Figura 5.3, pode-se observar que os trabalhadores pertencentes a Categoria 1 da variável **CLASSE CNAE**, que realizam atividades na área da agricultura, pecuária, silvicultura, exploração florestal e construção correm maior risco de saírem do emprego, enquanto os que pertencem a Categoria 5 têm menor risco, visto que realizam atividades relacionadas a administração pública, defesa e seguri-

dade social. Isso pode ocorrer devido a estabilidade alcançada pelos servidores no setor público e pode ser visto também através do gráfico da variável **NAT JURIDICA**: Os servidores da administração pública tendem a permanecer mais tempo na função do que quando comparados a trabalhadores de empresas que possuem outra natureza jurídica.

Como já esperado, os trabalhadores analfabetos são os que têm menor tempo de sobrevivência e os que possuem mestrado ou doutorado são os que têm maior tempo. Os que possuem Ensino Fundamental ou Ensino Médio completo ou incompleto têm aproximadamente o mesmo risco de sairem do emprego, o que pode indicar uma preferência por parte das empresas por pessoas com nível de ensino igual ou mais avançado que o Ensino Superior.

O gráfico da variável **IDADE** mostra que a única faixa etária que se diferencia das outras e tem maior tempo de sobrevivência é a de 14 a 19 anos, enquanto os indivíduos que se encontram nas outras faixas têm maior chance de não permanecerem no emprego.

A continuação da apresentação dos gráficos das outras variáveis se encontra nas Figuras 5.5, 5.6 e 5.7 a seguir.

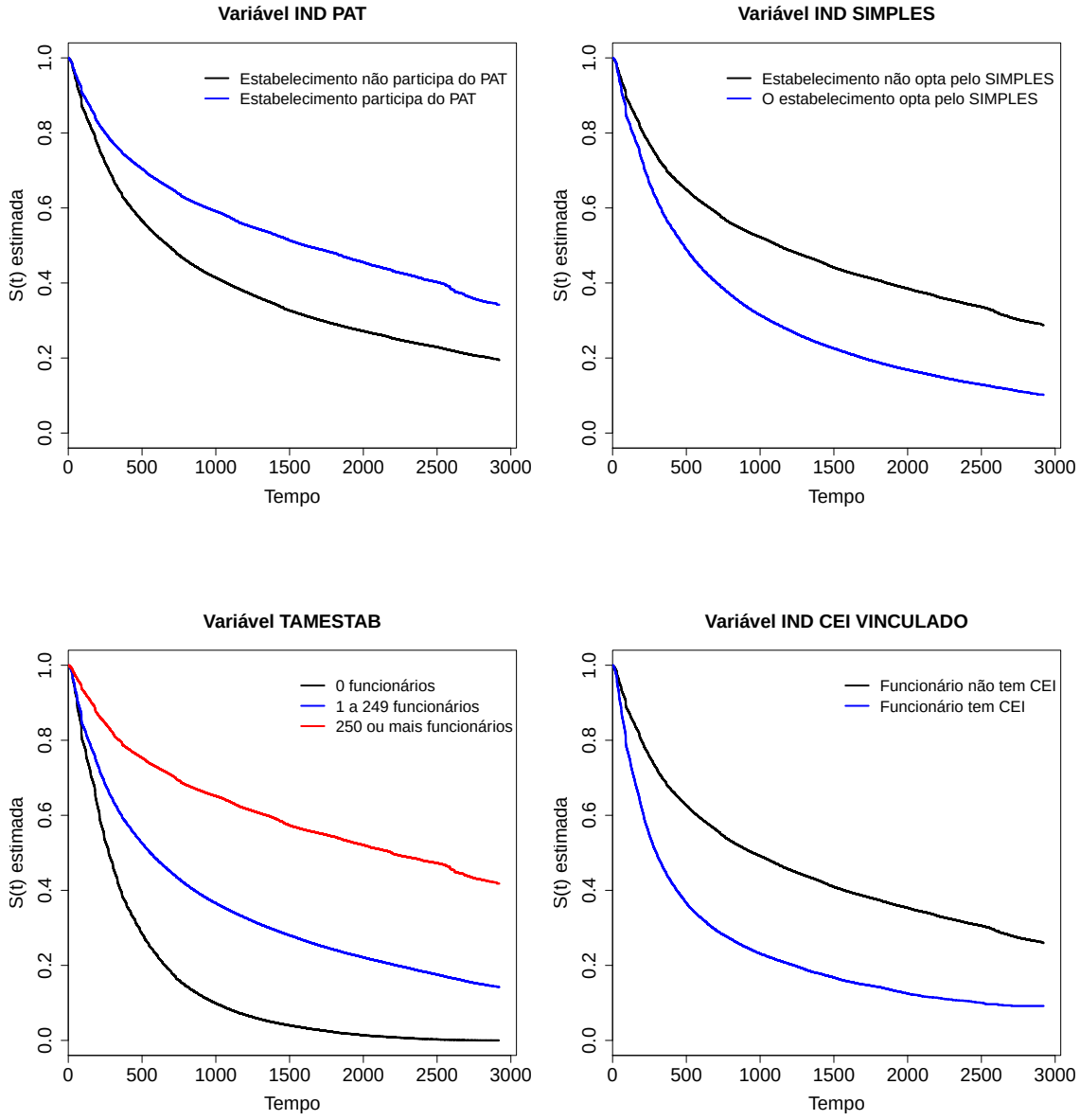

Figura 5.5: Curvas de sobrevivência das covariáveis IND PAT, IND SIMPLES, TAMANHO ESTAB e IND CEI VINCULADO estimadas por Kaplan-Meier.

Pela análise de cada um dos gráficos apresentados na Figura 5.5, observa-se que os estabelecimentos que participam do programa de alimentação ao trabalhador são os contratantes dos indivíduos que passam mais tempo no emprego, assim como os estabelecimentos que não adotam o SIMPLES. Essa última situação pode indicar que as maiores empresas são as que os trabalhadores tem menor risco de saírem do

emprego e isso pode ser evidenciado pela análise do gráfico da variável **TAMANHO ESTAB**: o tempo de vida dos trabalhadores que exercem atividades em empresas que possuem 250 ou mais funcionários é maior que o tempo de trabalhadores de empresas menores.

Aparentemente, são os maiores estabelecimentos, que têm mais de 250 funcionários, os contratantes dos indivíduos que passam mais tempo no emprego e os trabalhadores de empresas que declaram ter empregados com Cadastro Específico do INSS, pelo fato de possuírem obra de construção civil, têm maior risco de sair do emprego.

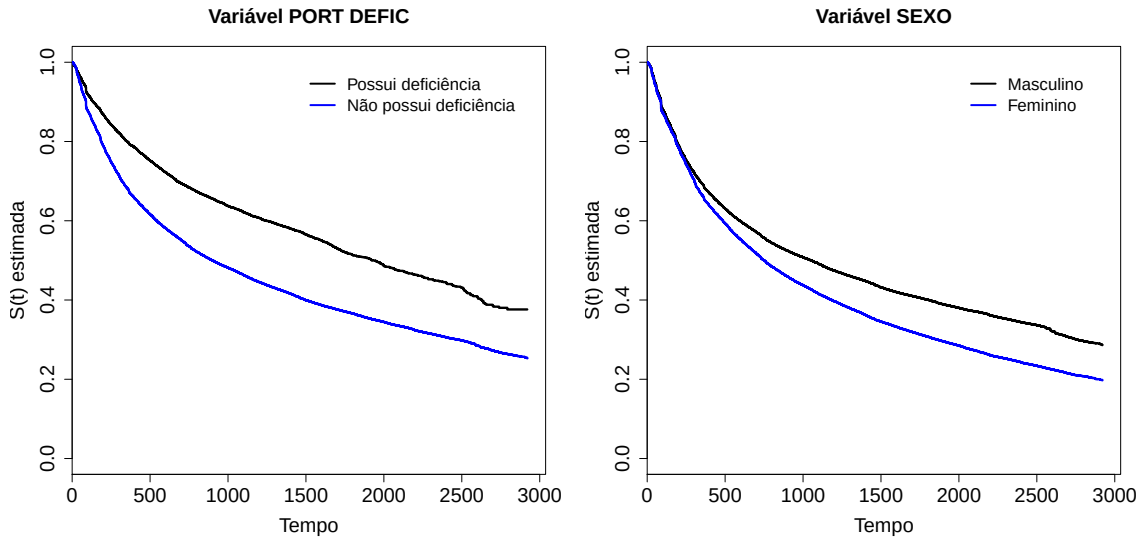

Figura 5.6: Curvas de sobrevivência das covariáveis PORT DEFICIENCIA e SEXO estimadas por Kaplan-Meier.

Através dos gráficos da Figura 5.6, percebe-se que os trabalhadores com deficiência, assim como os do sexo masculino, tem menor risco de saírem de sua atual função. O mesmo acontece com quem recebe salário mensalmente, como é possível observar na Figura 5.7.

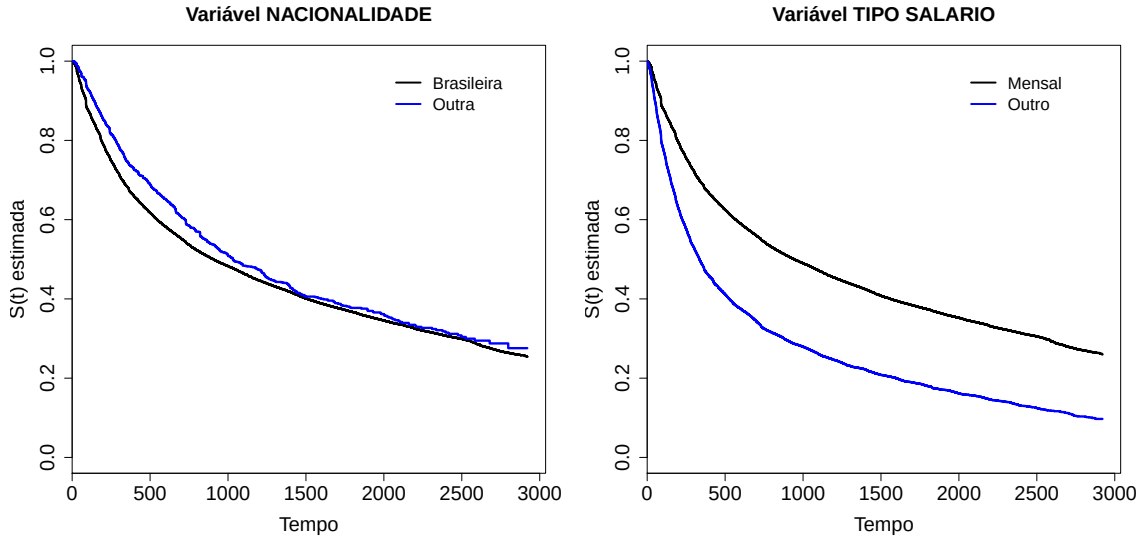

Figura 5.7: Curvas de sobrevivência das covariáveis NACIONALIDADE e TIPO SALARIO estimadas por Kaplan-Meier.

A Figura 5.7 também mostra que as categorias da variável **NACIONALIDADE** não possuem diferenças significativas aparentemente. Decidiu-se, então, retirá-la da análise final e ela não fará parte do modelo. Como já visto na Tabela 5.1, quase 100% dos indivíduos presentes na base de dados são brasileiros.

5.2 Modelo Probabilístico

Na tentativa de se encontrar um modelo paramétrico para descrever os tempos de vida, foi utilizado o método da curva do Tempo Total em Teste (Curva TTT), cujo gráfico é apresentado na Figura 5.8.

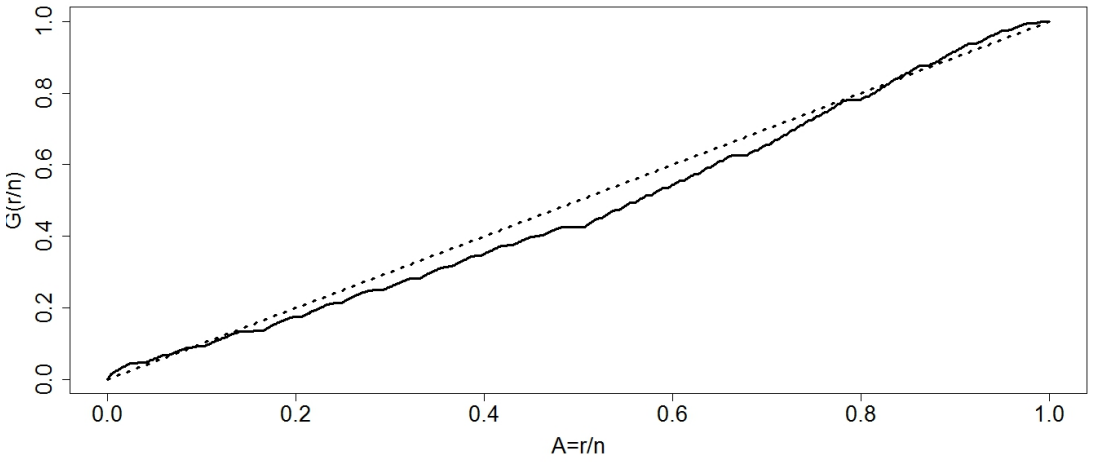

Figura 5.8: Curva do Tempo Total em Teste para os dados dos trabalhadores do DF.

Nota-se que a forma da curva não está bem definida. Pode-se dizer que ela assume a forma de uma reta diagonal, o que leva a tentativa de modelar os dados por distribuição que possui a função de risco constante, a exponencial. Pode-se dizer também que a curva assume a forma convexa que é relacionada a distribuições que possuem função de risco monotonicamente descrescente, como no caso da distribuição Weibull. Logo, uma tentativa de modelar os dados através dela será feita.

Como mencionado na metodologia, no momento da construção do gráfico da curva TTT, as censuras não são consideradas. Assim, por tentativa e devido a curva não ter apresentado um comportamento bem diferenciado, os dados foram modelados através da distribuição Log-Normal que possui função de risco unimodal.

A fim de permitir a comparação e a melhor escolha da distribuição, os gráficos a seguir mostram as curvas de sobrevivência estimadas por Kaplan-Meier e pelas distribuições sem a presença de covariáveis. Para uma melhor visualização, os símbolos que marcavam as censuras foram desconsiderados. As estimativas paramétricas da

função de sobrevivência foram obtidas pelo comando SURVREG da biblioteca Survival do *software* R. O primeiro gráfico, apresentado na Figura 5.9, mostra os dados modelados pela distribuição exponencial.

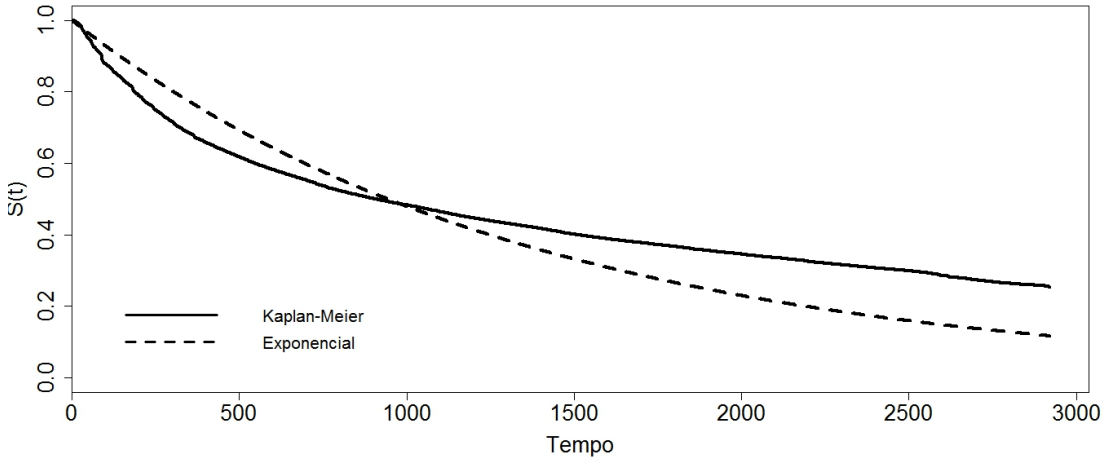

Figura 5.9: Curvas de sobrevivência estimadas pelo modelo exponencial e por Kaplan-Meier.

Percebe-se que não houve um bom ajustamento, pois as curvas deveriam coincidir ou ficarem bem próximas. Assim, a Figura 5.10 mostra a tentativa em modelar os dados pela distribuição Weibull.

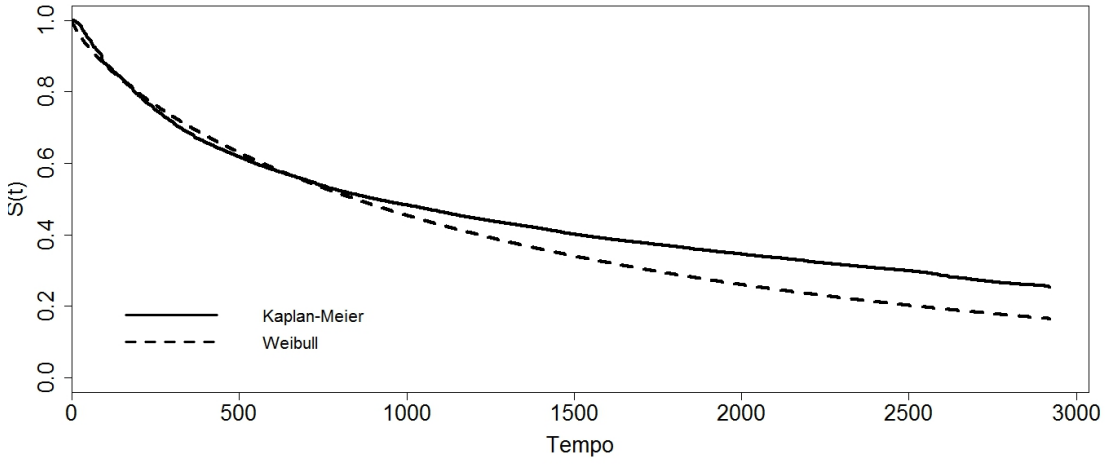

Figura 5.10: Curvas de sobrevivência estimadas pelo modelo Weibull e por Kaplan-Meier.

Observa-se que o ajustamento pela distribuição Weibull foi melhor que pela exponencial mas, ainda assim, não parece ser o mais adequado. Houve a seguir uma tentativa de usar a distribuição log-normal para ser usado como modelo. A Figura 5.11 mostra a comparação entre as curvas estimadas por Kaplan-Meier e pelo modelo log-normal.

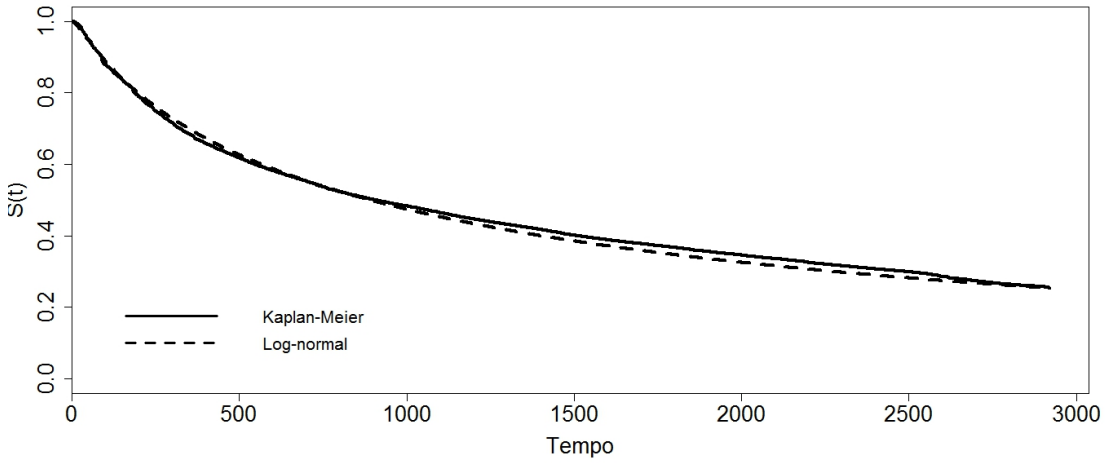

Figura 5.11: Curvas de sobrevivência estimadas pelo modelo lognormal e por Kaplan-Meier.

Pode-se perceber que o ajustamento ficou bom pois não existe grande diferença entre as curvas. Assim, decidiu-se utilizar a distribuição log-normal como modelo. A expressão da estimativa da função de sobrevivência, considerando um modelo sem covariáveis, é dada por:

$$\hat{S}(t) = 1 - \Phi\left(\frac{\log(t) - 6,79}{1,79}\right)$$

5.2.1 Seleção de covariáveis

Após a validação e correção, apresentada no capítulo 4, a base de dados passou a ter 12 covariáveis. Destas, apenas 11 foram consideradas na ánalise porque a

variável **NACIONALIDADE** foi retirada em função de suas categorias não apresentarem diferenças significativas, como já apresentado na seção *Análise descritiva dos dados*. Devido a grande quantidade de observações, não é conveniente utilizar os métodos de seleção de variáveis como *stepwise*, *backward* e *forward*. Assim, foram utilizadas técnicas gráficas para decidir quais delas farão parte do modelo final. Foram construídos gráficos para comparar as curvas de sobrevivência das categorias das variáveis estimadas por Kaplan-Meier e pelo modelo log-normal, que foi definido como o melhor para ajustar as observações. Novamente, os marcadores dos tempos de censura foram desconsiderados para uma melhor comparação das curvas. Os gráficos se encontram nas Figuras 5.12, 5,13 e 5,14.

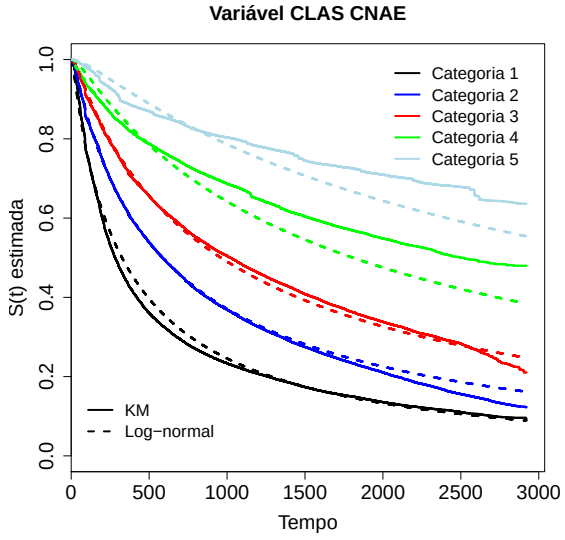

Figura 5.12: Curvas de sobrevivência das categorias da covariável CLASSE CNAE estimadas pelo modelo log-normal e por Kaplan-Meier.

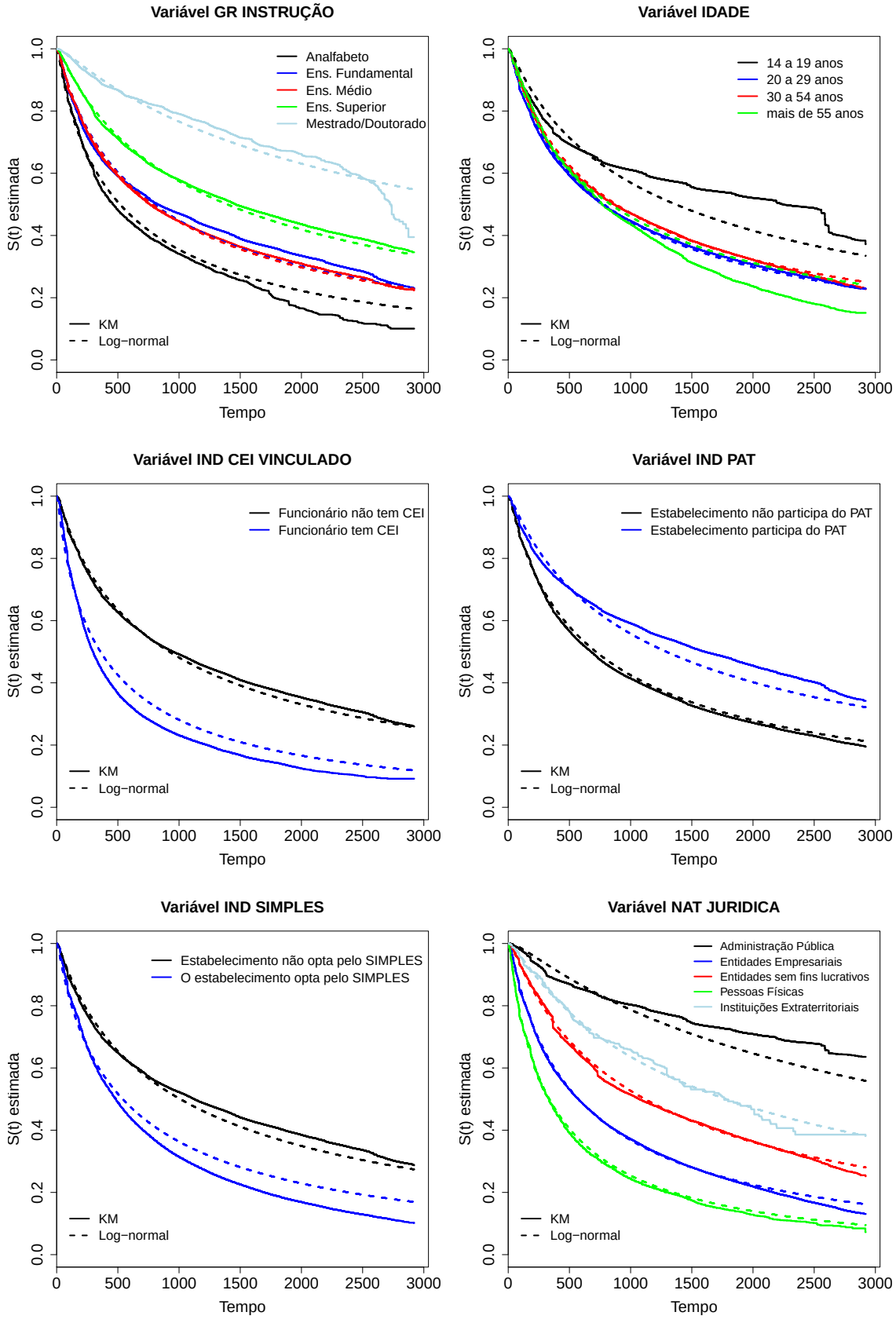

Figura 5.13: Curvas de sobrevivência das categorias das covariáveis estimadas pelo modelo log-normal e por Kaplan-Meier.

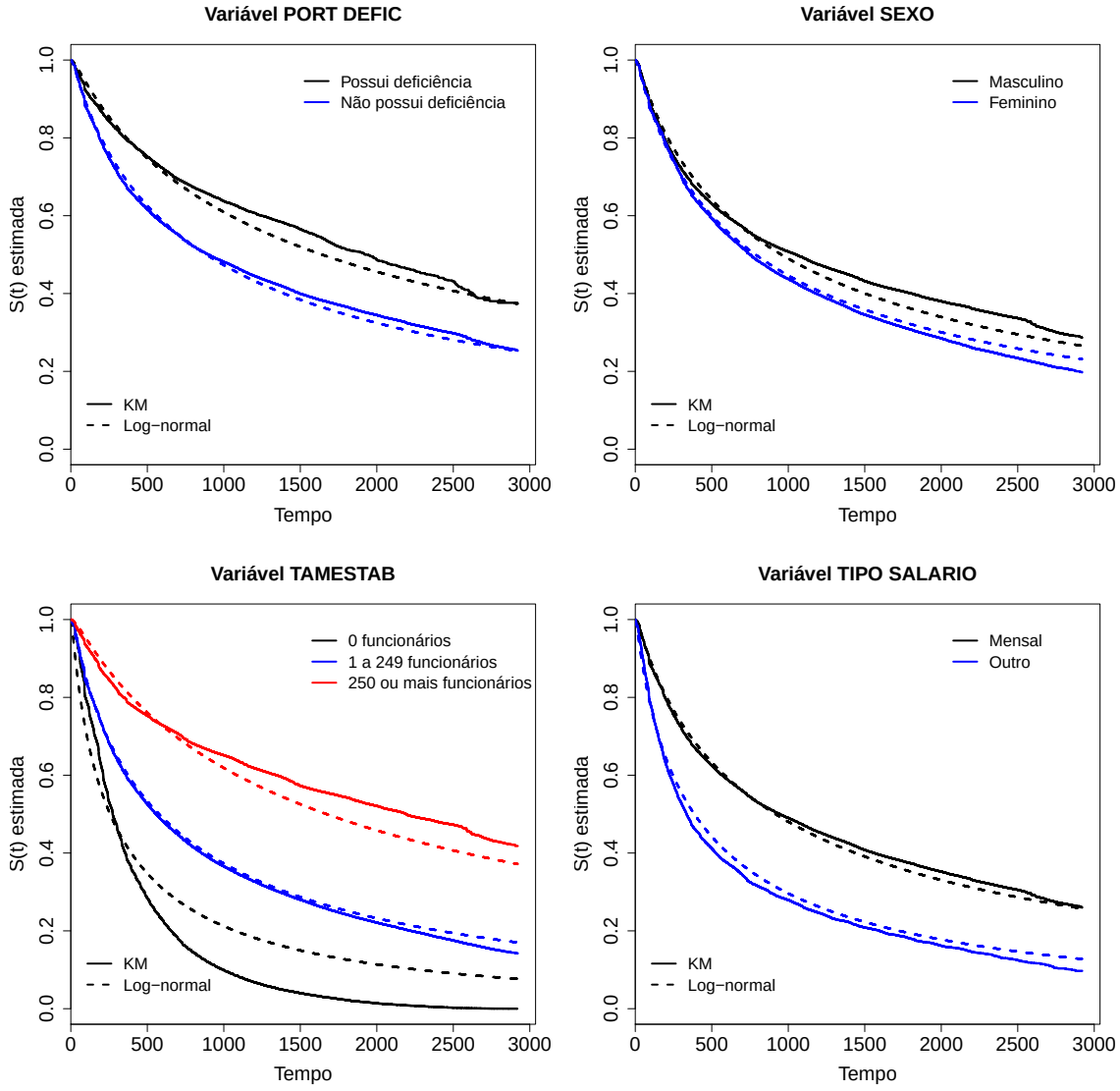

Figura 5.14: Curvas de sobrevivência das categorias das covariáveis estimadas pelo modelo log-normal e por Kaplan-Meier.

Por estarmos tratando de um modelo paramétrico, era esperado que as curvas estimadas pelo modelo log-normal não se ajustassem perfeitamente as estimadas pelo método de Kaplan- Meier. No entanto, nenhuma covariável apresentou uma grande diferença entre as estimativas do modelo log-normal com as estimativas de Kaplan-Meier. Assim, decidiu-se manter todas as 11 covariáveis no modelo.

5.2.2 Modelo Log-normal com covariáveis

O modelo log-normal foi ajustado com as 11 covariáveis selecionadas. A inclusão dessas covariáveis foi feita considerando a função ligação identidade (Agresti, 2007) para representar o parâmetro μ através do vetor de covariáveis $\mathbf{x}$. Essa relação pode ser expressa por:

$$\mu(\mathbf{x}) = \mathbf{x}_0'\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{11} x_{11}.$$

As estimativas de máxima verossimilhança de $\boldsymbol{\beta}$ com seus erros padrões (EP), assim como do intervalo de 95% de confiança para $\boldsymbol{\beta}$ do modelo log-normal são apresentadas na Tabela 5.2.

Assim, a função de sobrevivência para um indivíduo que possui vetor de covariáveis $\mathbf{x}$ é estimada por:

$$\hat{S}(t|\mathbf{x}) = 1 - \Phi\left(\frac{\log(t) - \hat{\mu}(\mathbf{x})}{\hat{\sigma}}\right),$$

com $\hat{\mu}(\mathbf{x}) = \hat{\beta}_0 + \hat{\beta}_{CLASCNAE} + \hat{\beta}_{GRINSTR} + \hat{\beta}_{IDADE} + \hat{\beta}_{CEI} + \hat{\beta}_{PAT} + \hat{\beta}_{SIMPLES} + \hat{\beta}_{NATJURIDICA} + \hat{\beta}_{PORTDEFIC} + \hat{\beta}_{SEXO} + \hat{\beta}_{TAMESTAB} + \hat{\beta}_{TPSALARIO}$.

Tabela 5.2: Estimativas dos parâmetros do modelo Log-normal.

	$\beta \pm$EP	β(IC 95%)
β_0	6,57178 ± 0,039	(6,495 ; 6,649)
σ	1,6	-
CLASSE CNAE		
Categoria 1	0	
Categoria 2	0,579 ± 0,006	(0,568 ; 0,590)
Categoria 3	1,021 ± 0,007	(1,008 ; 1,035)
Categoria 4	1,447 ± 0,011	(1,425 ; 1,469)
Categoria 5	1,108 ± 0,026	(1,057 ; 1,159)
GR INSTRUÇÃO		
Analfabeto	0	
Ens. Fundamental	0,157 ± 0,022	(0,113 ; 0,200)
Ens. Médio	0,121 ± 0,022	(0,077 ; 0,165)
Ens. Superior	0,071 ± 0,022	(0,027 ; 0,115)
Mestrado/Doutorado	1,123 ± 0,038	(1,049 ; 1,197)
IDADE		
14 a 19 anos	0	
20 a 29 anos	-0,089 ± 0,005	(-0,098 ; -0,080)
30 a 54 anos	0,013 ± 0,005	(0,004 ; 0,023)
mais de 55 anos	-0,205 ± 0,01	(-0,225 ; -0,185)
IND CEI VINCULADO		
Não	0	
Sim	0,002 ± 0,008	(-0,014 ; 0,018)
IND PAT		
Não	0	
Sim	0,387 ± 0,004	(0,380 ; 0,394)
IND SIMPLES		
Não	0	
Sim	0,067 ± 0,004	(0,059 ; 0,075)
NAT JURIDICA		
Administração Pública	0	
Entidades Empresariais	-1,121 ± 0,025	(-1,170 ; -1,071)
Entidades sem fins lucrativos	-0,551 ± 0,026	(-0,602 ; -0,500)
Pessoas Físicas	-1,026 ± 0,029	(-1,084 ; -0,968)
Instituições Extraterritoriais	-0,336 ± 0,07	(-0,473 ; -0,199)
PORT DEFICIENCIA		
Sim	0	
Não	-0,514 ± 0,018	(-0,550 ; -0,479)
SEXO		
Masculino	0	
Feminino	-0,194 ± 0,003	(-0,200 ; -0,188)
TAMANHO ESTAB		
0 funcionários	0	
1 a 249 funcionários	0,661 ± 0,007	(0,647 ; 0,676)
250 ou mais funcionários	0,838 ± 0,008	(0,822 ; 0,854)
TIPO SALARIO		
Mensal	0	
Outro	-0,538 ± 0,008	(-0,553 ; -0,523)

[1]Nota: as categorias da variável **CLASSE CNAE** estão especificadas no Capítulo 4, na seção *Validação e correção dos dados*, na subseção *Recodificação da variável CLASSE CNAE*.

[2]Nota: as classes com $\beta = 0$ são os níveis de referência das variáveis.

Como exemplo, temos que a probabilidade de um indivíduo com **CLASSE CNAE**=Categoria 2, **GR INSTRUÇÃO**=Ensino Médio, **IDADE**=14 a 19 anos, **IND CEI VINCULADO**=Não, **IND PAT**=Sim, **IND SIMPLES**=Não, **NAT JURIDICA**=Entidades empresariais, **PORT DEFICIENCIA**=Não, **SEXO**=Feminino, **TAMANHO ESTAB**=1 a 249 funcionários, **TIPO SALARIO**=Mensal permanecer no trabalho por mais que 990 dias é:

$$\hat{S}(990|\mathbf{x}) = 1 - \Phi\left(\frac{\log(990) - \hat{\mu}(\mathbf{x})}{1,6}\right) = 1 - \Phi(0,2543) = 0,3996.$$

em que $\hat{\mu}(\mathbf{x}) = 6,57178 + 0,579 + 0,121 + 0 + 0 + 0,387 + 0 - 1,121 - 0,514 - 0,194 + 0,661 + 0 = 6,49078$.

Em outras palavras, essa é a probabilidade de um indivíduo cujo trabalho está relacionado a Categoria 2 da CNAE, que tenha Ensino Médio (completo ou incompleto), tenha entre 14 e 19 anos, não tenha CEI, não possua deficiência, seja mulher, receba salário mensal, que trabalhe em um estabelecimento que participa do PAT, não opte pelo SIMPLES, seja uma entidade empresarial e tenha de 1 a 249 funcionários permanecer no trabalho por mais que 990 dias.

5.3 Modelo de regressão de Cox

O modelo de regressão de Cox foi utilizado nesse estudo com o interesse em se avaliar o poder da explicação das covariáveis. A suposição básica para seu uso é que as taxas de falha sejam proporcionais e a avaliação dessa proporcionalidade pode ser observada a partir dos gráficos das curvas de sobrevivência das covariáveis já mostrados na análise descritiva dos dados. Nessa seção, os gráficos de Kaplan-Meier

são apresentados novamente nas Figuras 5.15, 5.16 e 5.17 mas já com a presença da curva estimada pelo modelo de Cox.

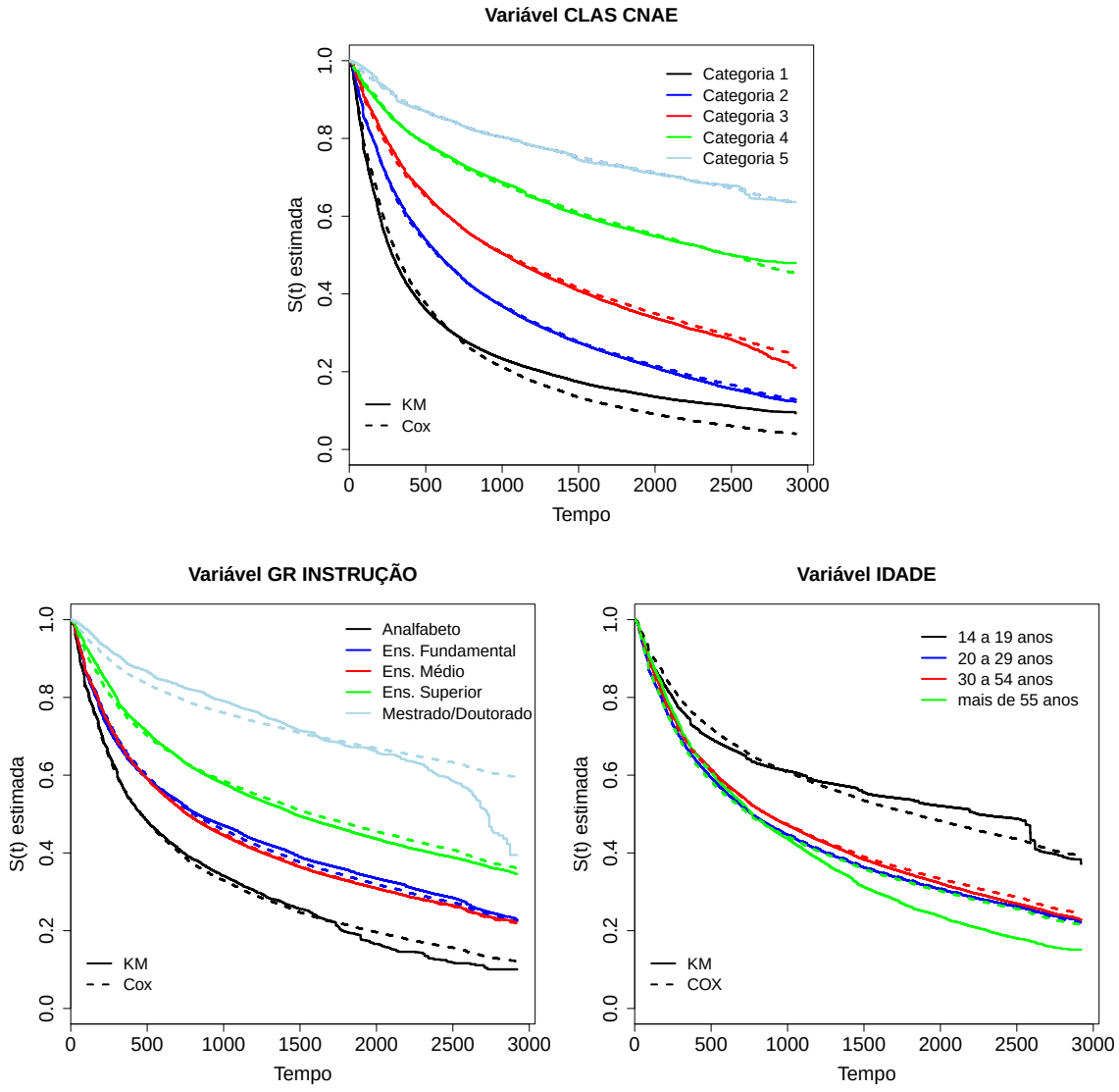

Figura 5.15: Curvas de sobrevivência das categorias das covariáveis estimadas pelo modelo de regressão de Cox e por Kaplan-Meier.

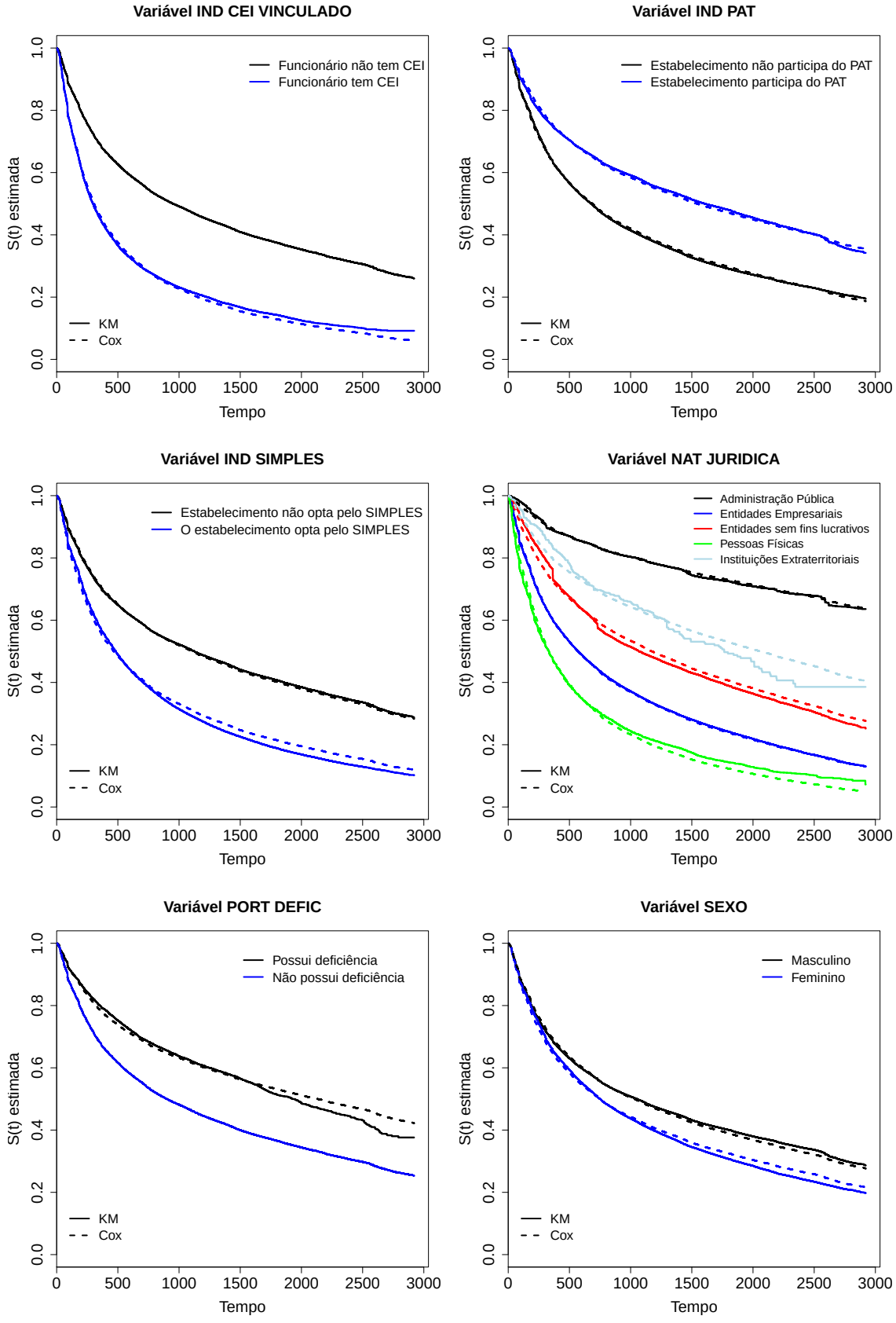

Figura 5.16: Curvas de sobrevivência das categorias das covariáveis estimadas pelo modelo de regressão de Cox e por Kaplan-Meier.

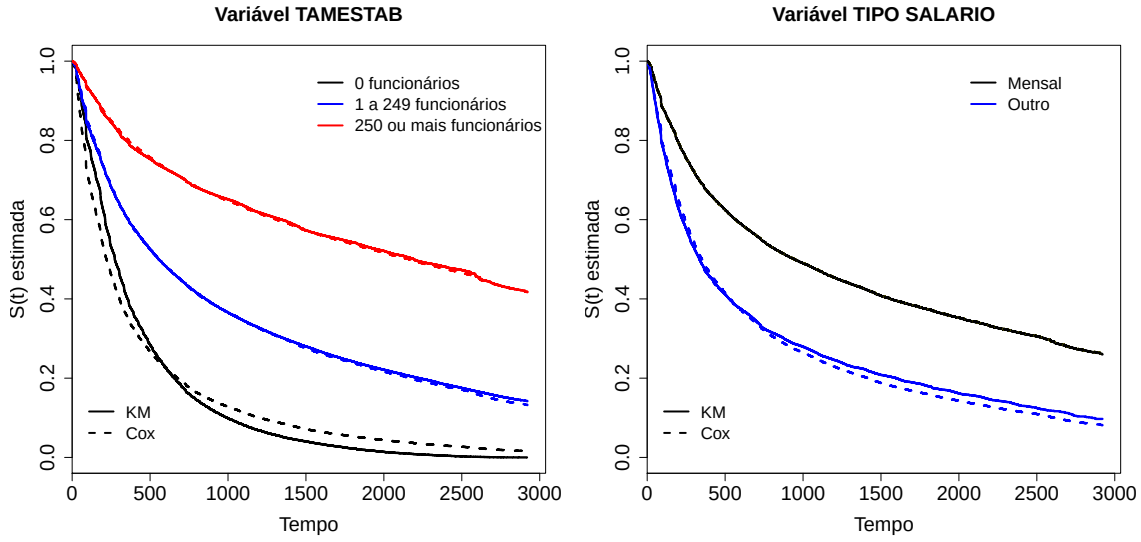

Figura 5.17: Curvas de sobrevivência das categorias das covariáveis estimadas pelo modelo de regressão de Cox e por Kaplan-Meier.

Por se tratar de um método não-paramétrico, a aproximação entre as curvas estimadas por Kaplan-Meier e por Cox é muito boa. Aparentemente, a suposição de riscos proporcionais não é violada na maioria das variáveis. No caso da **CLASSE CNAE**, a curva da Categoria 1 parece estar se aproximando da curva da Categoria 2. A curva que representa os trabalhadores que possuem mestrado ou doutorado, no gráfico da variável **GR INSTRUCAO**, teve um decaimento brusco no fim do período de acompanhamento e as curvas que representam os Ensinos Fundamental e Médio paracem entrar em contato uma ou mais vezes. Entre 3 das categorias variável **IDADE** ocorre o mesmo.

Para uma avaliação mais cuidadosa, foram utilizados os resíduos padronizados de Schoenfeld que apesar de envolverem conclusões subjetivas, foram a saída encontrada para o impasse da impossibilidade do uso de testes de hipóteses. Os gráficos se encontram a seguir, nas Figuras 5.18 e 5.19.

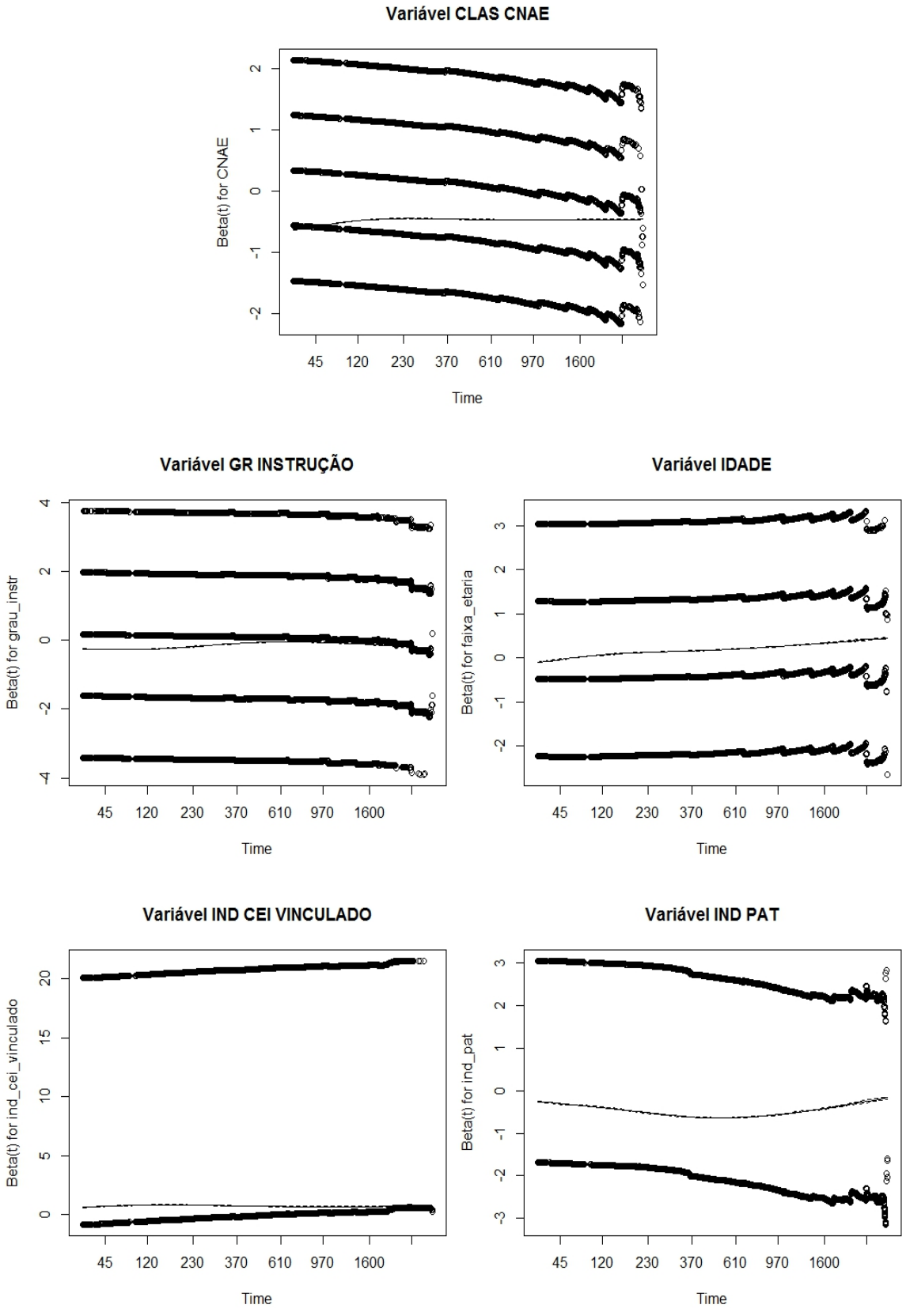

Figura 5.18: Resíduos padronizados de Schoenfeld *versus* os tempos.

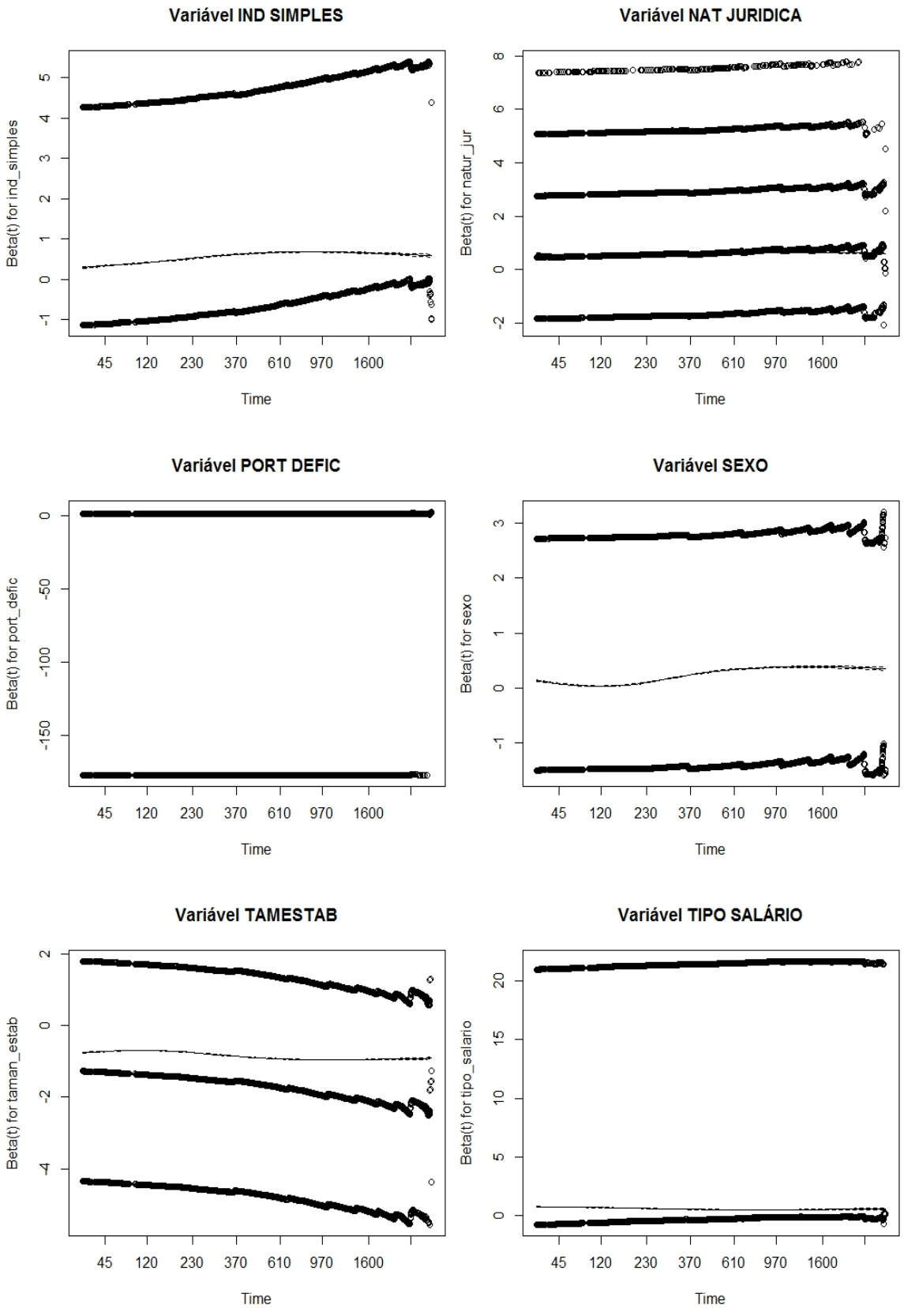

Figura 5.19: Resíduos padronizados de Schoenfeld *versus* os tempos.

Podemos observar que as variáveis que possuem duas categorias são as que mais aparentam atender a suposição de riscos proporcionais por seus resíduos não apresentarem inclinação. Já as variáveis com mais categorias chegam a apresentar resíduos com alguma tendência em tempos maiores. No entanto, não há evidências que alguma delas viole a suposição devido a grande quantidade de observações. Com isso, decidiu-se por manter as 11 covariáveis que permaneceram na base de dados após as modificações que são apresentadas no capítulo 4 e na análise descritiva desse capítulo.

Assim, o modelo de regressão de Cox foi ajustado e para o cálculo da função de sobrevivência temos que a estimativa da função de sobrevivência de base $\hat{S}_0(t)$ é expressa por:

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

A Tabela 5.3 apresenta as estimativas de $\hat{S}_0(t)$, sendo que o tempo está apresentado de 30 em 30 dias devido a grande quantidade de tempos distintos observados.

Tabela 5.3: Estimativas da função de sobrevivência de base $\hat{S}_0(t)$.

t	$\hat{S}_0(t)$	t	$\hat{S}_0(t)$	t	$\hat{S}_0(t)$
30	0,9760277	1020	0,5039101	2010	0,368392
60	0,9374294	1050	0,4986981	2040	0,3651482
90	0,8955609	1080	0,4928991	2070	0,3620675
120	0,8717997	1110	0,4870725	2100	0,3594284
150	0,8484549	1140	0,4815361	2130	0,3567725
180	0,8243789	1170	0,4760169	2160	0,3539873
210	0,796286	1200	0,4713667	2190	0,3503414
240	0,7745903	1230	0,4668198	2220	0,3464326
270	0,7538523	1260	0,4621987	2250	0,3439051
300	0,7346836	1290	0,4578063	2280	0,3412796
330	0,7155505	1320	0,4534833	2310	0,3387649
360	0,6999697	1350	0,4494207	2340	0,3360207
390	0,6840329	1380	0,4450335	2370	0,3331153
420	0,6716163	1410	0,4406323	2400	0,3300795
450	0,6593805	1440	0,4354172	2430	0,3278527
480	0,6477789	1470	0,429964	2460	0,3254967
510	0,6364787	1500	0,4255782	2490	0,3232313
540	0,6247634	1530	0,4217237	2520	0,3203981
570	0,6148157	1560	0,4179118	2550	0,3175133
600	0,605443	1590	0,4141243	2580	0,3137374
630	0,5963565	1620	0,4104933	2610	0,3085004
660	0,5876584	1650	0,4072407	2640	0,3039593
690	0,5791769	1680	0,4040395	2670	0,3010675
720	0,5702363	1710	0,4007437	2700	0,297247
750	0,5607125	1740	0,3975315	2730	0,29412
780	0,5520892	1770	0,3943313	2760	0,2909436
810	0,5451047	1800	0,391277	2790	0,2887195
840	0,5383299	1830	0,387397	2820	0,2864709
870	0,5318399	1860	0,3836726	2850	0,2841104
900	0,5257842	1890	0,3805639	2880	0,2819405
930	0,5197527	1920	0,3773255	2910	0,2796188
960	0,5140539	1950	0,374463		
990	0,5090241	1980	0,3714861		

A estimação dos parâmetros relativos as covariáveis foi feita pelo método de máxima verossimilhança parcial aproximado por Efron e se encontra na Tabela 5.4, assim como as estimativas do erro padrão, do risco relativo e do intervalo de 95% confiança de risco relativo.

Tabela 5.4: Estimativas dos parâmetros do modelo de Cox.

	$\beta \pm$EP	Risco Relativo (IC 95%)
CLASSE CNAE		
Categoria 1	0	1
Categoria 2	-0,452 ± 0,004	0,636 (0,631 ; 0,641)
Categoria 3	-0,789 ± 0,005	0,454 (0,45 ; 0,459)
Categoria 4	-1,194 ± 0,01	0,303 (0,297 ; 0,309)
Categoria 5	-0,956 ± 0,023	0,385 (0,368 ; 0,402)
GR INSTRUÇÃO		
Analfabeto	0	1
Ens. Fundamental	-0,119 ± 0,016	0,888 (0,86 ; 0,917)
Ens. Médio	-0,12 ± 0,016	0,886 (0,858 ; 0,915)
Ens. Superior	-0,06 ± 0,017	0,942 (0,912 ; 0,973)
Mestrado/Doutorado	-0,868 ± 0,032	0,42 (0,394 ; 0,447)
IDADE		
14 a 19 anos	0	1
20 a 29 anos	-0,029 ± 0,004	0,972 (0,964 ; 0,979)
30 a 54 anos	-0,139 ± 0,004	0,87 (0,863 ; 0,877)
mais de 55 anos	0,009 ± 0,008	1,009 (0,993 ; 1,025)
IND CEI VINCULADO		
Não	0	1
Sim	0,024 ± 0,006	1,025 (1,013 ; 1,037)
IND PAT		
Não	0	1
Sim	-0,264 ± 0,003	0,768 (0,764 ; 0,773)
IND SIMPLES		
Não	0	1
Sim	-0,012 ± 0,003	0,988 (0,982 ; 0,994)
NAT JURIDICA		
Administração Pública	0	1
Entidades Empresariais	0,881 ± 0,023	2,413 (2,308 ; 2,522)
Entidades sem fins lucrativos	0,470 ± 0,023	1,601 (1,53 ; 1,675)
Pessoas Físicas	0,777 ± 0,025	2,175 (2,071 ; 2,284)
Instituições Extraterritoriais	0,317 ± 0,06	1,372 (1,22 ; 1,544)
PORT DEFICIENCIA		
Sim	0	1
Não	0,452 ± 0,015	1,572 (1,525 ; 1,62)
SEXO		
Masculino	0	1
Feminino	0,125 ± 0,002	1,134 (1,128 ; 1,139)
TAMANHO ESTAB		
0 funcionários	0	1
1 a 249 funcionários	-0,641 ± 0,005	0,527 (0,522 ; 0,532)
250 ou mais funcionários	-0,776 ± 0,005	0,46 (0,455 ; 0,465)
TIPO SALARIO		
Mensal	0	1
Outro	0,421 ± 0,006	1,524 (1,507 ; 1,541)

[1]Nota: as categorias da variável **CLASSE CNAE** estão especificadas no Capítulo 4, na seção *Validação e correção dos dados*, na subseção *Recodificação da variável CLASSE CNAE*.

[2]Nota: as classes com $\beta = 0$ são os níveis de referência das variáveis.

Assim, o modelo de regressão de Cox foi ajustado e a função de sobrevivência para um indivíduo com vetor de covariáveis $\mathbf{x} = (x_1, ..., x_{11})'$ é estimada por:

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

em que $\mathbf{x}'\hat{\boldsymbol{\beta}} = \hat{\beta}_{CLASCNAE} + \hat{\beta}_{GRINSTR} + \hat{\beta}_{IDADE} + \hat{\beta}_{CEI} + \hat{\beta}_{PAT} + \hat{\beta}_{SIMPLES} + \hat{\beta}_{NATJURIDICA} + \hat{\beta}_{PORTDEFIC} + \hat{\beta}_{SEXO} + \hat{\beta}_{TAMESTAB} + \hat{\beta}_{TPSALARIO}$.

Como exemplo, a partir das duas tabelas anteriores, temos que a probabilidade de um indivíduo com **CLASSE CNAE**=Categoria 2, **GR INSTRUÇÃO**=Ensino Médio, **IDADE**=14 a 19 anos, **IND CEI VINCULADO**=Não, **IND PAT**=Sim, **IND SIMPLES**=Não, **NAT JURIDICA**=Entidades empresariais, **PORT DEFICIENCIA**=Não, **SEXO**=Feminino, **TAMANHO ESTAB**=1 a 249 funcionários, **TIPO SALARIO**=Mensal permanecer no trabalho por mais que 990 dias é:

$$\hat{S}(990|\mathbf{x}) = [\hat{S}_0(990)]^{\exp\{\mathbf{x}'\hat{\beta}\}} = 0,5090241^{\exp\{-0,019\}} = 0,516$$

em que $\mathbf{x}'\hat{\beta} = -0,452 - 0,12 + 0 + 0 - 0,264 + 0 + 0,881 + 0,452 + 0,125 - 0,641 + 0 = -0,019$ e $\hat{S}_0(990)$ é dada pela Tabela 5.3.

Em outras palavras, essa é a probabilidade de um indivíduo cujo trabalho está relacionado a Categoria 2 da CNAE, que tenha Ensino Médio (completo ou incompleto), tenha entre 14 e 19 anos, não tenha CEI, não possua deficiência, seja mulher, receba salário mensal, que trabalhe em um estabelecimento que participa do PAT, não opte pelo SIMPLES, seja uma entidade empresarial e tenha de 1 a 249 funcionários permanecer no trabalho por mais que 990 dias.

Capítulo 6

CONCLUSÕES

Os resultados obtidos sugerem que o modelo de regressão log-normal é um modelo adequado para ajustar os dados sobre tempo de permanência no emprego dos trabalhadores do DF através das 11 variáveis explicativas selecionadas. O modelo de regressão de Cox também se mostrou adequado para esse mesmo fim. Os testes tradicionais de ajuste de modelos não puderam ser aplicados nesse trabalho devido ao grande número de observações na amostra. Como esperado, uma amostra de mais de 1.6 milhões de observações concedeu poder suficiente para rejeitar qualquer tipo de teste de ajuste do modelo ou de seleção de variáveis. Assim, todas as decisões de escolha do melhor modelo paramétrico, seleção e agregação dos níveis das covariáveis foram realizadas considerando técnicas gráficas e o tamanho do efeito (effect size) das estimativas, ao invés da significância estatística. Desta forma, uma comparação direta dos dois modelos apresentados não pôde ser realizada.

Como visto em um exemplo ilustrando os dois modelos, as estimativas da função de sobrevivência para um certo tempo t apresentaram uma pequena divergência (aproximadamente 11%) para a específica combinação escolhida para as covariáveis. A escolha do melhor modelo a ser utilizado, portanto, depende do objetivo do pesqui-

sador. A escolha do modelo lognormal é preferível se o interesse é realizar previsões do tempo de sobrevivência de um trabalhador no mercado de trabalho, quando esse tempo é maior que os observados na amostra (extrapolação). Já o modelo de Cox, por ser um modelo não paramétrico pode ser o escolhido se a previsão é para um tempo dentro da amplitude de tempos observados (interpolação). Por ser um modelo não paramétrico, o modelo de Cox falha em prever tempos superiores àqueles observados na amostra.

Assim, evidencia-se neste trabalho a dificuldade em se lidar com grandes bases de dados, devido a impossibilidade do uso de procedimentos usuais para verificar a significância dos resultados. Além disso, dificuldades também são encontradas, principalmente, quando os dados observados são derivados de declarações e quando as informações disponíveis se encontram incompletas ou são falsas, como foi o caso do banco da RAIS.

Como propostas futuras sugerimos desenvolver uma metodologia de validação cruzada para modelos de sobrevivência, que permita avaliar o desempenho dos modelos e assim, poder confrontá-los entre si. Ainda, novos tipos de modelagem poderão ser considerados. Devido à grande proporção de censuras observadas (aproximadamente 54%), um modelo de sobrevivência com fração de cura pode ser uma alternativa para modelar esse tipo de dados.

Referências Bibliográficas

Agresti, A. (2007). *An introduction to Categorical Data Analysis*, (2 ed.). John Wiley and Sons, Inc.

BRASIL (1975). Decreto - lei nº 76.900, de 23 de dezembro de 1975. Institui a Relação Anual de Informações Sociais - RAIS, URL `http://www3.dataprev.gov.br/sislex/paginas/23/1975/76900.htm`. Acesso em 03 jun. 2014.

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

Colosimo, E. A. & Giolo, S. R. (2006). *Análise de Sobrevivência Aplicada*, (1 ed.). EDGARD BLUCHER.

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

Fernandes, A. M. R. (2010). Análise de dados em modelos multiestado. Technical report, Universidade do Minho.

IBGE (2003). *Classificação Nacional de Atividades Econômicas Fiscal*, (1.1 ed.).

IBM. *IBM SPSS Statistics 20 Command Syntax Reference*.

INFOMONEY (2005). Pis/pasep: saiba qual a diferença e para que servem. URL `http://www.infomoney.com.br/minhas-financas/noticia/17391/pis-pasep-saiba-qual-diferen-ccedil-para-que-servem`. Acesso em 03 jun. 2014.

Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, v.53, p.457-481.

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

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

MPAS (2014). Categoria de segurados. URL `http://www.previdencia.gov.br/informaes-2/categoria-de-segurados/`. Acesso em 25 jun. 2014.

MTE (2012). *Manual de Orientação da Relação Anual de Informações Sociais (RAIS)*.

MTE (2014). Dados e estatísticas: Relação anual de informações sociais - rais. URL `http://www3.mte.gov.br/rais/oquee.asp`. Acesso em 03 jun. 2014.

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

Neto, F. L., Mazicheli, J., & Achcar, J. A. (2002). *Introdução a Análise de Sobrevivência e Confiabilidade*. III Jornada Regional de Estatística.

OIT (2012). *Perfil do trabalho decente no Brasil: Um olhar sobre as Unidades da Federação*, (1 ed.). URL `http://www.oit.org.br/sites/default/files/topic/gender/doc/relatoriotrabalhodecentetotal_876.pdf`. Acesso em 06 mai. 2014.

Outhwaite, W. & Bottomore, T. (1996). *Dicionário do pensamento social do século XX*, (1 ed.). Zahar.

RCORETEAM (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

Santos, T. A. (2013). Modelo de regressão pertencente à família weibull com fração de cura. Technical report, Universidade de Brasília.

SEF-SP. *Simples Nacional*. URL `http://www.fazenda.sp.gov.br/educacao_fiscal/contents/Simples%20Nacional.pdf`. Acesso em 13 nov. 2014.

SETRAB-DF (2013). Mercado de trabalho aquecido no df. URL `http://www.trabalho.df.gov.br/noticias/item/2273-mercado-de-trabalho-aquecido-no-df.html`. Acesso em 01 abr. 2014.