ANÁLISE DO TEMPO DE PERMANÊNCIA DE TRABALHADORES NO MERCADO DE TRABALHO DO DISTRITO FEDERAL VIA MODELO DE RISCOS PROPORCIONAIS DE COX E LOG-NORMAL

Rayany de Oliveira SANTOS[1]
Eduardo Yoshio NAKANO[1]

- Resumo: Neste trabalho, o modelo de regressão log-normal e o modelo de regressão de Cox foram utilizados para analisar dados de sobrevivência relacionados a trabalhadores formais do Distrito Federal. O objetivo foi estimar, através de um conjunto de dados, a distribuição do tempo de permanência desses trabalhadores em um emprego. 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. Ambos os modelos considerados apresentaram ajustes adequados para o conjunto de dados considerado.

- PALAVRAS-CHAVE: Análise de sobrevivência; modelos de regressão; RAIS.

1 Introdução

Segundo Outhwaite e 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, porém 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 Secretaria de Trabalho do Distrito Federal (SETRAB-DF, 2013), a taxa de desemprego continua sendo a menor registrada desde 1992.

[1] Universidade de Brasília - UNb, Departamento de Estatística, CEP: 70910-900, Brasília, DF, Brasil. E-mail: *rayany_oliveira@hotmail.com*; *nakano@unb.br*

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 na região onde está localizada a capital do país possam ser consideradas ideais, para que o quantitativo de pessoas desempregadas diminua cada vez mais, 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.

Neste contexto, o objetivo do trabalho foi estimar a distribuição de probabilidades do tempo de permanência de um trabalhador em seu emprego mais recente, desde a admissão até a demissão, no Distrito Federal de acordo com as características da pessoa ou do trabalho. Isto permitirá, por exemplo, estimar o tempo médio de permanência de um particular trabalhador (ou perfil de trabalhadores) em seu emprego ou a probabilidade desse trabalhador permanecer um determinado período neste emprego. O estudo foi realizado dentro do contexto de Análise de Sobrevivência, utilizando o modelo de riscos proporcionais de Cox e o modelo de regressão log-normal. Toda a análise foi realizada pelo software livre R (R CORE TEAM, 2013).

2 Material e métodos

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. Financiado pelo Fundo de Amparo ao Trabalhador - FAT, o Ministério do Trabalho e Emprego - MTE oferece consulta via internet à base de dados da RAIS (MTE, 2014). 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.

No presente estudo, foram utilizadas as bases de dados dos anos 2002 a 2009 da RAIS, com informações apenas dos trabalhadores alocados na região do Distrito Federal. A limitação do período considerado no estudo foi devido à dificuldade de obter a base de dados em CD-ROM/DVD (base mais completa, utilizada para pesquisas mais aprofundadas), cujo acesso é feito somente após assinatura de um termo de compromisso firmado entre o MTE e a instituição interessada (MTE, 2014). Apesar de novas variáveis terem sido criadas e incluídas nas bases com o passar dos anos, somente as variáveis presentes simultaneamente nas bases de 2002 a 2009 foram consideradas. Foram adotadas neste estudo 11 covariáveis na análise, descritas a seguir:

1. **Classe CNAE:** classe da atividade econômica do estabelecimento informante, segundo a Classificação feita pela Comissão Nacional de Classificação (CONCLA, 2015);
2. **Grau de instrução:** Grau de instrução do trabalhador;
3. **Idade:** Idade do trabalhador na data de admissão;

4. **CEI vinculado:** Indica se o empregado/servidor está ligado ao Cadastro Específico do INSS;
5. **PAT:** Indicador de estabelecimento participante do Programa de Alimentação do Trabalhador;
6. **Ind. Simples:** Indicador de estabelecimento optante pelo Sistema Integrado de Pagamento de Impostos e Contribuições das Microempresas e Empresas de Pequeno Porte;
7. **Natureza jurídica:** Natureza jurídica da empresa;
8. **Portador de deficiência:** Indica se o trabalhador possui deficiência;
9. **Sexo:** Sexo do trabalhador;
10. **Tam. do estabelecimento:** Tamanho do estabelecimento baseado no número de trabalhadores;
11. **Tipo Salário:** Tipo de salário do empregado/servidor, de acordo com o contrato de trabalho.

Todas as variáveis consideradas no estudo foram categorizadas agrupando seus níveis de acordo com seus riscos relativos. O procedimento consistiu em agrupar aqueles níveis que apresentaram riscos relativos individuais semelhantes, respeitando a natureza das variáveis ordinais e o seu interesse prático (por exemplo, apesar da variável Idade ter apresentado apenas dois níveis distintos, optou-se representa-la através de quatro níveis). Maiores detalhes sobre este procedimento pode ser visto em Santos (2015).

Para a aplicação das técnicas de Análise de Sobrevivência, a variável de interesse foi definida como o tempo (em dias) compreendido entre a data de admissão e a data de demissão do trabalhador. Caso tenhamos, por algum motivo, perdido o acompanhamento do trabalhador durante o período de estudo ou o mesmo não foi demitido até o dia 31 de dezembro de 2009, o dado será considerado censurado.

A modelagem dos dados foi realizada considerando um modelo semi-paramétrico, através do modelo de riscos proporcionais de Cox (COX, 1972), e também através do modelo paramétrico log-normal. Mesmo sendo o tempo medido em dias, a utilização de um modelo contínuo é razoável devido a grande amplitude dos valores observados (NAKANO e CARRASCO, 2006). Trabalhos que apresentam aplicações do modelo de riscos proporcionais de Cox podem ser vistos em Martins (2003), Nakano e Cunha (2012), entre outros. E exemplos de modelos de sobrevivência que consideram tempos discretos podem ser vistos em Nakagawa e Osaki (1975) e Brunello e Nakano (2015).

Assim, o modelo de riscos proporcionais de Cox e o modelo log-normal foram ajustados com as 11 covariáveis consideradas no estudo. A inclusão das covariáveis no modelo log-normal foi feita considerando a função de ligação identidade (AGRESTI, 2007). Desta forma, o modelo de regressão log-normal apresenta a seguinte forma:

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(\frac{\log(t) - \mu(x'\beta)}{\sigma} \right) \right\}, \quad t > 0, \tag{1}$$

em que σ é o parâmetro de escala do modelo e $\mu(x'\beta) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$. Aqui, β_0 é o intercepto e β_i são os coeficientes de regressão associados às covariáveis x_i, i=1, 2,..., k.

A escolha do modelo log-normal foi feita através da comparação das estimativas empíricas e teóricas da função de sobrevivência sem a consideração de covariáveis. Para o modelo de Cox, a suposição de proporcionalidade dos riscos foi verificada através dos resíduos de Schoenfeld (COLOSIMO e GIOLO, 2006). O ajuste global dos modelos de Cox e log-normal será verificado pelo resíduo de Cox-Snell (COX e SNELL, 1968) e Deviance (COLOSIMO e GIOLO, 2006).

3 Resultados e discussão

A base de dados utilizada para a análise 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 até essa última data. Entretanto, percebeu-se que 675 deles apresentaram o mesmo dia de admissão e demissão (0 dias de permanência) 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 um dia e o maior tempo encontrado foi 2.921 dias, que representam aproximadamente 8 anos. A Tabela 1 apresenta as frequências das categorias das covariáveis após as recodificações descritas na Seção 2.

3.1 Modelo log-normal

Destaca-se a quantidade de censuras observadas nos dados, sendo que 53,9% dos tempos foram censurados. Este fato inviabilizou a seleção de potenciais modelos paramétricos através da curva do Tempo Total em Teste (curva TTT), pois a mesma não considera a censura nos tempos (COLOSIMO e GIOLO, 2006). Além disto, devido a grande quantidade de observações na base de dados (mais que 1,6 milhões), não foi viável a realização de qualquer teste de aderência do modelo, pois esse teste passaria a rejeitar qualquer mínimo desvio da distribuição teórica, mesmo esse desvio sendo insignificante no ponto de vista prático. Desta forma, a decisão pela escolha do modelo paramétrico foi embasada em procedimentos gráficos comparando a curva de sobrevivência estimada empiricamente e através de um modelo teórico.

O ajuste do modelo paramétrico foi realizado comparando as curvas de sobrevivência estimadas por Kaplan-Meier (KAPLAN e MEIER, 1958) e por alguns modelos paramétricos sem a presença de covariáveis. As estimativas paramétricas das curvas de sobrevivência foram obtidas pelo comando SURVREG da biblioteca *survival* do *software* R (R CORE TEAM, 2013). Os modelos considerados foram: exponencial, Weibull, log-normal e log-logística. Essa escolha foi devido ao fato desses modelos serem os principais modelos utilizados em análise de sobrevivência, capazes de ajustar formas de risco constantes (exponencial), monótonas (Weibull) e unimodais (log-normal e log-logística). A seleção do modelo que melhor se ajusta aos dados foi realizada segundo o Critério de informação de Akaike – AIC (AKAIKE, 1974) e o Critério de informação Bayesiano – BIC (SCHWARS, 1978) e entre os modelos considerados, o log-normal foi o que apresentou o melhor ajuste (Tabela 2). O bom ajuste do modelo log-normal pode ser observado na Figura 1.

Tabela 1 - Frequências absolutas e relativas das covariáveis

		Frequência	Percentual
Classe CNAE[(1)]	Categoria 1	172.621	10,49%
	Categoria 2	889.139	54,04%
	Categoria 3	180.666	10,98%
	Categoria 4	45.027	2,74%
	Categoria 5	357.828	21,75%
Grau de instrução	Analfabeto	6.672	0,41%
	Ens. Fundamental	511.225	31,07%
	Ens. Médio	782.589	47,57%
	Ens. Superior	339.279	20,62%
	Mestrado/Doutorado	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%
CEI Vinculado	Não	1.582.262	96,17%
	Sim	63.022	3,83%
PAT	Não	1.059.462	64,39%
	Sim	56.022	35,61%
Ind. Simples	Não	1.303.198	79,21%
	Sim	342.086	20,79%
Natureza Jurídica	Administração pública	361.911	22,00%
	Entidades empresariais	1.171.355	71,19%
	Ent. sem fins lucrativos	96.465	5,86%
	Pessoas físicas	14.637	0,89%
	Inst. extraterritoriais	914	0,06%
Portador de deficiência	Sim	12.659	0,77%
	Não	1.632.625	99,23%
Sexo	Masculino	1.059.086	64,37%
	Feminino	586.198	35,63%
Tam. do estabelecimento	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%

A codificação das categorias da variável Classe CNAE pode ser vista no Apêndice.

Fonte: MTE (2014)

Tabela 2 - Comparação dos modelos pelos critérios AIC e BIC

Modelo	**AIC**	**BIC**
Exponencial	12.458.150	12.458.162
Weibull	12.370.306	12.370.330
Log-normal	12.285.383	12.285.408
Log-logística	12.317.024	12.317.048

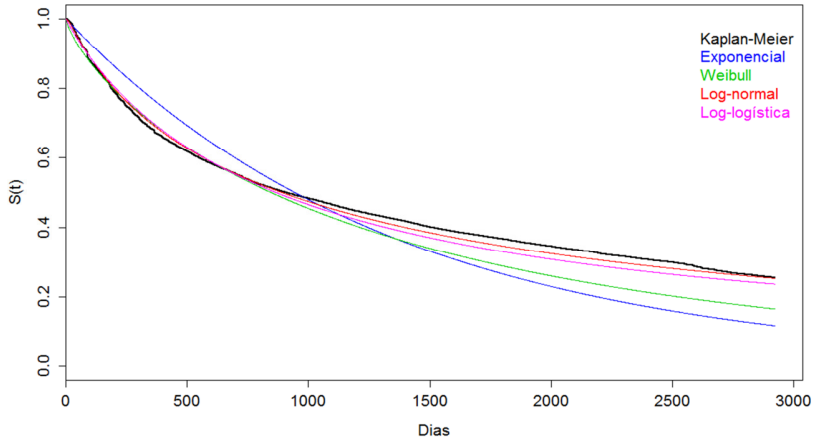

Figura 1 - Estimativas das funções de sobrevivência, sem a consideração de covariáveis, segundo Kaplan-Meier e modelos exponencial, log-normal, exponencial, Weibull e log-logística.

Devido a grande quantidade de observações, não foi conveniente utilizar os testes formais de seleção de covariáveis. Assim, foram utilizadas técnicas gráficas para decidir se algumas delas deveriam ser excluídas do modelo final. A decisão foi baseada na comparação das curvas de sobrevivência de cada covariável ajustada individualmente com a curva de sobrevivência empírica estimada por Kaplan-Meier. Desta forma, decidiu-se manter todas as 11 covariáveis no modelo log-normal final - a análise completa pode ser vista em Santos (2015). As estimativas de máxima verossimilhança dos parâmetros são apresentadas na Tabela 3.

Assim, de acordo com o modelo paramétrico log-normal (1), a probabilidade de um indivíduo cujo trabalho está relacionado na 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 permaneça mais de 990 dias no trabalho é de:

$$\hat{S}_{\text{lognormal}}(990 \mid \mathbf{x}) = 1 - \Phi\left(\frac{\log(990) - \hat{\mu}(\mathbf{x})}{\hat{\sigma}}\right) = 0{,}3996\,. \qquad (2)$$

Aqui, Φ(.) é a função de distribuição acumulada de uma normal padrão e as estimativas de $\hat{\mu}(\mathbf{x})$ e $\hat{\sigma}$ são apresentadas pela Tabela 2.

Tabela 3 - Estimativas dos parâmetros do modelo de regressão log-normal

		Estimativa(EP)	IC 95%
β_0		6,572 (0,039)	(6,495 ; 6,649)
$\log(\sigma)$		0,468 (0,001)	(0,466 ; 0,470)
Classe CNAE[(1)]	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)
Grau de 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 (0,010)	(–0,225 ; –0,185)
CEI Vinculado	Não	0	
	Sim	0,002 (0,008)	(–0,014 ; 0,018)
PAT	Não	0	
	Sim	0,378 (0,004)	(0,380 ; 0,394)
Ind. Simples	Não	0	
	Sim	0,067 (0,004)	(0,059 ; 0,075)
Natureza Jurídica	Administração pública	0	
	Entidades empresariais	–1,121 (0,025)	(–1,170 ; –1,071)
	Ent. sem fins lucrativos	–0,551 (0,026)	(–0,602 ; –0,500)
	Pessoas físicas	–1,026 (0,029)	(–1,084 ; –0,968)
	Inst. extraterritoriais	–0,336 (0,070)	(–0,473 ; –0,199)
Portador de deficiência	Sim	0	
	Não	–0,514 (0,018)	(–0,550 ; –0,479)
Sexo	Masculino	0	
	Feminino	–0,194 (0,003)	(–0,200 ; –0,188)
Tam. do estabelecimento	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 Salário	Mensal	0	

	Outro	–0,538 (0,008)	(–0,553 ; –0,523)

Notas: As categorias com coeficientes iguais a zero são os níveis de referência das variáveis; EP: Erro-padrão.
(1) A codificação das categorias da variável Classe CNAE pode ser vista no Apêndice.

3.2 Modelo de riscos proporcionais de Cox

O modelo de regressão de Cox foi também ajustado com mesmas covariáveis consideradas no modelo log-normal. A suposição básica para seu uso é que os riscos sejam proporcionais. Devido a grande quantidade de observações, não foi conveniente utilizar os tradicionais testes de tendências. Assim, a proporcionalidade dos riscos foi avaliada graficamente através dos resíduos padronizados de Schoenfeld (SCHOENFELD, 1982). O gráfico de resíduos de Schoenfeld foi construído considerando cada covariável separadamente e os resultados podem ser vistos em Santos (2015). Todas as covariáveis consideradas não apresentaram evidências de violar a suposição de riscos proporcionais. Desta forma, assim como no modelo log-normal, decidiu-se manter todas as 11 covariáveis no modelo de riscos proporcionais de Cox. Assim, o modelo de riscos proporcionais de Cox foi ajustado e a estimativa de sua função de sobrevivência base, $S_0(t)$ é apresentada pela Tabela 4.Devido a grande quantidade de tempos distintos, as estimativas de $S_0(t)$ foram apresentadas em intervalos de 30 dias.

Tabela 4 - Estimativas da função de sobrevivência base

t	$S_0(t)$	t	$S_0(t)$	t	$S_0(t)$	t	$S_0(t)$
30	0,9760	780	0,5521	1530	0,4217	2280	0,3413
60	0,9374	810	0,5451	1560	0,4179	2310	0,3388
90	0,8956	840	0,5383	1590	0,4141	2340	0,3360
120	0,8718	870	0,5318	1620	0,4105	2370	0,3331
150	0,8485	900	0,5258	1650	0,4072	2400	0,3301
180	0,8244	930	0,5198	1680	0,4040	2430	0,3279
210	0,7963	960	0,5141	1710	0,4007	2460	0,3255
240	0,7746	990	0,5090	1740	0,3975	2490	0,3232
270	0,7539	1020	0,5039	1770	0,3943	2520	0,3204
300	0,7347	1050	0,4987	1800	0,3913	2550	0,3175
330	0,7156	1080	0,4929	1830	0,3874	2580	0,3137
360	0,7000	1110	0,4871	1860	0,3837	2610	0,3085
390	0,6840	1140	0,4815	1890	0,3806	2640	0,3040
420	0,6716	1170	0,4760	1920	0,3773	2670	0,3011
450	0,6594	1200	0,4714	1950	0,3745	2700	0,2972
480	0,6478	1230	0,4668	1980	0,3715	2730	0,2941
510	0,6365	1260	0,4622	2010	0,3684	2760	0,2909
540	0,6248	1290	0,4578	2040	0,3651	2790	0,2887
570	0,6148	1320	0,4535	2070	0,3621	2820	0,2865
600	0,6054	1350	0,4494	2100	0,3594	2850	0,2841
630	0,5964	1380	0,4450	2130	0,3568	2880	0,2819
660	0,5877	1410	0,4406	2160	0,3540	2910	0,2796
690	0,5792	1440	0,4354	2190	0,3503		
720	0,5702	1470	0,4300	2220	0,3464		

750	0,5607	1500	0,4256	2250	0,3439

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

$$\hat{S}_{\text{Cox}}(t \mid \mathbf{x}) = \left[\hat{S}_0(t)\right]^{\exp\{\mathbf{x}'\hat{\beta}\}}, \tag{3}$$

onde $\hat{\boldsymbol{\beta}}$ é o vetor de parâmetros, cujas estimativas são apresentadas na Tabela 5.

Tabela 5 - Estimativas dos parâmetros do modelo de riscos proporcionais de Cox

		Estimativa(EP)	$e^{\boldsymbol{\beta}}$ **(IC 95%)**
Classe CNAE[(1)]	Categoria 1	0	
	Categoria 2	–0,452 (0,004)	0,636 (0,631 ; 0,641)
	Categoria 3	–0,789 (0,005)	0,454 (0,450 ; 0,459)
	Categoria 4	–1,194 (0,010)	0,303 (0,297 ; 0,309)
	Categoria 5	–0,956 (0,023)	0,385 (0,368 ; 0,402)
Grau de instrução	Analfabeto	0	
	Ens. Fundamental	–0,119 (0,016)	0,888 (0,086 ; 0,917)
	Ens. Médio	–0,120 (0,016)	0,886 (0,858 ; 0,915)
	Ens. Superior	–0,060 (0,017)	0,942 (0,912 ; 0,973)
	Mestrado/Doutorado	–0,868 (0,032)	0,420 (0,394 ; 0,447)
Idade	14 a 19 anos	0	
	20 a 29 anos	–0,029 (0,004)	0,972 (0,964 ; 0,979)
	30 a 54 anos	–0,139 (0,004)	0,870 (0,863 ; 0,877)
	Mais de 55 anos	0,009 (0,008)	1,009 (0,993 ; 1,025)
CEI Vinculado	Não	0	
	Sim	0,024 (0,006)	1,025 (1,013 ; 1,037)
PAT	Não	0	
	Sim	–0,264 (0,003)	1,025 (1,013 ; 1,037)
Ind. Simples	Não	0	
	Sim	–0,012 (0,003)	0,988 (0,982 ; 0,994)
Natureza Jurídica	Administração pública	0	
	Entidades empresariais	0,881 (0,023)	2,413 (2,308 ; 2,522)
	Ent. sem fins lucrativos	0,470 (0,023)	1,601 (1,530 ; 1,675)
	Pessoas físicas	0,777 (0,025)	2,175 (2,071 ; 2,284)
	Inst. extraterritoriais	0,317 (0,060)	1,372 (1,220 ; 1,544)
Portador de deficiência	Sim	0	
	Não	0,452 (0,015)	1,572 (1,525 ; 1,620)
Sexo	Masculino	0	
	Feminino	0,125 (0,002)	1,134 (1,128 ; 1,139)
Tam. do estabelecimento	0 funcionários	0	
	1 a 249 funcionários	–0,641 (0,005)	0,527 (0,522 ; 0,532)
	250 ou mais funcionários	–0,776 (0,005)	0,460 (0,455 ; 0,465)
Tipo Salário	Mensal	0	
	Outro	0,421 (0,006)	1,524 (1,507 ; 1,541)

Notas: As categorias coeficientes iguais a zero são os níveis de referência das variáveis; EP: Erro-padrão.
(1) A codificação das categorias da variável Classe CNAE pode ser vista no Apêndice.

Assim, de acordo com o modelo de riscos proporcionais de Cox (3), a probabilidade de um indivíduo cujo trabalho está relacionado na 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 permaneça mais de 990 dias no trabalho é de:

$$\hat{S}_{\text{Cox}}(990 \mid \mathbf{x}) = \left[\hat{S}_0(990)\right]^{\exp\{\mathbf{x}'\hat{\boldsymbol{\beta}}\}} = 0{,}516\,. \tag{4}$$

Aqui, as estimativas $\hat{S}_0(990)$ e $\hat{\boldsymbol{\beta}}$ podem ser encontradas, respectivamente, nas Tabelas 4 e 5.

Análise de resíduos

Os diagnósticos do modelo de riscos proporcionais de Cox e do modelo log-normal foram realizados através do resíduo de Cox-Snell (COX e SNELL, 1968) e Deviance (COLOSIMO e GIOLO, 2006). Os resultados são apresentados, respectivamente, pelas Figuras 2 e 3.

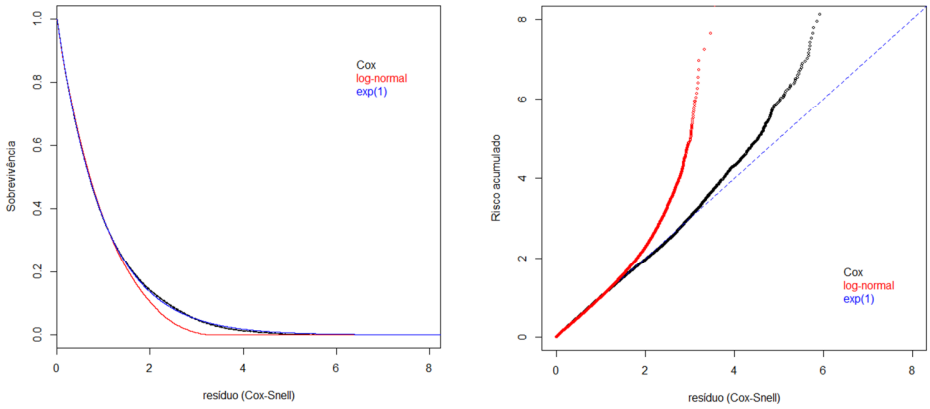

Figura 2 - Comparações das funções de sobrevivência e risco acumulado do resíduo de Cox-Snell com a Exponencial.

Segundo Lawless (2003), os resíduos de Cox-Snell vêm de uma população homogênea e devem seguir uma distribuição exponencial com média 1. A Figura 2 compara graficamente a função de sobrevivência e de risco acumulado dos resíduos calculados segundo o modelo de Cox e log-normal. Quanto mais próximo à função de sobrevivência dos resíduos se aproxima da função de sobrevivência da exponencial,

melhor é o ajuste. De forma equivalente, quando mais a função de risco acumulado dos resíduos se aproxima de uma reta diagonal, melhor é o ajuste do modelo. Segundo Klein e Moeschberguer (2003), essa interpretação deve ser feita com cautela, visto que a distribuição exponencial dos resíduos somente é válida quando os verdadeiros valores dos parâmetros são considerados. Desvios da distribuição exponencial podem ocorrer, em algumas situações, devido à incerteza no processo de estimação dos parâmetros e esses desvios podem ser maiores principalmente na cauda direita da distribuição e em casos de amostras pequenas.

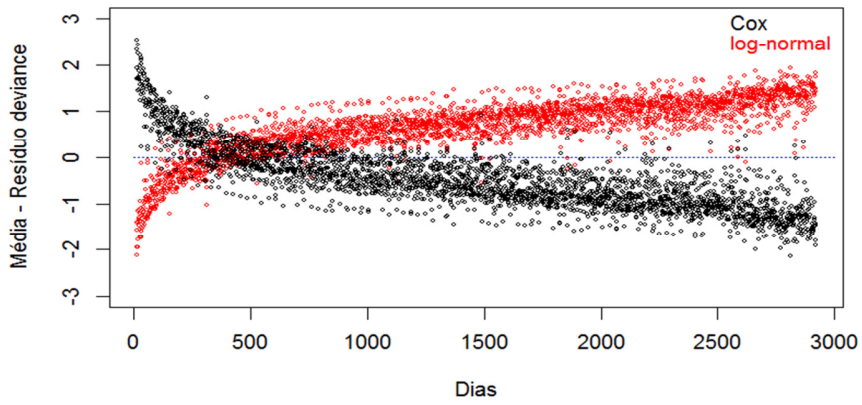

Figura 3 - Comparação das funções de sobrevivência e risco acumulado do resíduo de Cox-Snell com a Exponencial.

Observando a função de risco acumulada apresentada pela Figura2, nota-se a presença do fenômeno previsto por Moeschberger. Notam-se desvios na cauda da distribuição dos resíduos de ambos os modelos, quando comparados com a distribuição exponencial. No entanto, o gráfico da função de risco acumulado deve ser interpretado com cautela, pois a densidade dos pontos é superior para os valores pequenos dos resíduos. De fato, nota-se que apenas 0,12% dos resíduos do modelo de Cox são maiores que 3 (que é o ponto em que se começa notar um desvio significativo). Já no modelo log-normal, somente 2,8% das observações geraram resíduos maiores que 1,5 (que é o ponto em se começa notar um desvio significativo no modelo log-normal). Assim, entende-se que são "poucos" os valores dos resíduos que divergem da distribuição exponencial, podendo assim, considerar adequado o ajuste dos dois modelos. O gráfico da função de sobrevivência (Figura 2) corrobora esse fato, mostrando que há pouca divergência entre as funções de sobrevivência dos resíduos com a função de sobrevivência da distribuição exponencial.

A Figura 3 apresenta as médias dos resíduos Deviance para os modelos de riscos proporcionais de Cox e log-normal.

Visto que há muitos empates nos tempos (mais de 1,6 milhões de observações para menos de 3 mil tempos distintos), foi considerado aqui a média dos resíduos estimados para cada tempo. Os resíduos Deviance são negativos para observações com tempo de sobrevivência menores que o esperado (segundo o modelo ajustado) e positivos para

observações com tempo de sobrevivência maiores que o esperado. Para valores grandes do tempo, o resíduo do modelo de Cox foi, em média, negativo. Portanto, o modelo de Cox tende superestimar o tempo, isto é, o tempo de sobrevivência previsto pelo modelo é menor que o observado. Já o modelo log-normal apresentou um comportamento inverso, mostrando uma tendência de subestimar o tempo. Esse fenômeno pôde ser notado em (2) e (4), em que o modelo log-normal apresentou, em t = 990, uma estimativa de sobrevivência menor (e, portanto um tempo menor) que o modelo de Cox.

Conclusões

Os resultados obtidos sugerem que o modelo de regressão log-normal é um modelo adequado para ajustar a distribuição do tempo de permanência no emprego dos trabalhadores do DF através das 11 variáveis explicativas selecionadas. O modelo de riscos proporcionais 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. Neste trabalho, diagnósticos de multicolinearidade não foram realizados, visto que o objetivo foi estimar somente a distribuição do tempo de permanência dos trabalhadores e não os fatores de riscos que influenciam esse tempo.

Os resultados apresentados pela Figura 2 sugerem que, aparentemente, o modelo de Cox apresentou um melhor ajuste quando comparado com o log-normal. No entanto, devido ao grande numero de observações na amostra, não é possível verificar formalmente através de um teste estatístico se o ajuste do modelo de Cox é estatisticamente melhor que o modelo log-normal. A escolha do melhor modelo a ser utilizado, portanto, depende do objetivo do pesquisador. A escolha do modelo log-normal é preferível se o interesse é estimar a probabilidade de permanência de um trabalhador no mercado de trabalho por um tempo maior que os observados na amostra. Já o modelo de Cox, por ser um modelo semi-paramétrico, pode ser utilizado quando o foco da estimação está em tempos dentro da amplitude dos tempos observados. O modelo de Cox falha em obter estimativas para tempos superiores àqueles observados na amostra.

Evidencia-se também neste trabalho a dificuldade em se lidar com grandes bases de dados, devido à 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, como foi o caso do banco da RAIS.

Como propostas futuras, sugere-se considerar novos tipos de modelagens para abordar esse problema. Devido a grande quantidade de censuras observadas (aproximadamente 54%), um modelo de sobrevivência com fração de cura pode ser uma alternativa para modelar esse tipo de dados.

SANTOS, R. O.;NAKANO, E. Y. Analysis of job permanence time of formal workers on labor market in Federal District of Brazil via Cox proportional hazard and log-normal models. *Rev. Bras. Biom.*, São Paulo, v.33, n.4, p.570-584, 2015.

- *ABSTRACT: In this work, we propose to analyze survival data from formal labor market in Federal District of Brazil by log-normal regression and the Cox proportional hazard models. The objective was to estimate the distribution of job permanence time of these workers* through a dataset. *The large number of observations made unfeasible the use of traditional hypothesis testing and graphic techniques were used for decisions. Both log-normal and Cox models presented an appropriate adjustment for the considered data set.*

- *KEYWORDS: Survival analysis; regression models; RAIS.*

Referências

AGRESTI, A. *An introduction to categorical data analysis*. 2.ed. New Jersey: John Wiley and Sons, 2007. 400p.

AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, v.19, n.6, p.716-723, 1974.

BRUNELLO, G.H.V; NAKANO, E. Y. Inferência bayesiana no modelo Weibull discreto em dados com presença de censura. *TEMA - Tend. Mat. Apl. Comput.*, v.16, n.2, p.1-14, 2015.

COLOSIMO, E. A.; GIOLO, S.R. *Análise de sobrevivência aplicada*. 1.ed. São Paulo: Edgard Blucher, 2006. 370p.

CONCLA. *Classificação Nacional de Atividades Econômicas/CNAE*. 2015. [online] Disponível em http://www.receita.fazenda.gov.br/PessoaJuridica/CNAEFiscal/cnaef.htm. Acesso em abril de 2015.

COX, D.R. Regression models and life tables (with discussions). *J.R. Stat. Soc. B*, v.34, n.2, p.187-202, 1972.

COX, D. R.; SNELL, E. J. A general definition of residuals. *J.R. Stat. Soc. B*, v.30,n.2, p.248-275, 1968.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.*, v.53, n.282, p.457-481, 1958.

KLEIN, J.P; MOESCHBERGER, M. L. *Survival analysis: Techniques for censored and truncated data*. 2.ed. New York: Springer-Verlag, 2003. 536p.

LAWLESS, J. F. *Statistical models and methods for lifetime data*. 2.ed. New Jersey: John Wiley and Sons, 2003.664p.

MARTINS, M. S. *Previsão de insolvência pelo modelo de Cox: Uma contribuição para a análise de companhias abertas brasileiras*. 2003. 103f. Dissertação (Mestrado em Administração) – Escola de Administração da Universidade Federal do Rio Grande do Sul, Porto Alegre, 2003.

MTE. *Dados e estatísticas: Relação Anual de Informações Sociais - RAIS*. 2014. [online] Disponível em http://www3.mte.gov.br/rais/oquee.asp. Acesso em abril de 2015.

NAKAGAWA, T.; OSAKI, S. The discrete Weibull distribution. *IEEE Transactions on Reliability*, v.R-24, n.5, p.300-301, 1975.

NAKANO, E. Y.; CARRASCO, C. G. Uma avaliação do uso de um modelo contínuo na análise de dados discretos de sobrevivência. *TEMA - Tend. Mat. Apl. Comput.*, v.7, n.1, p.91-100, 2006.

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

OIT. *Perfil do trabalho decente no Brasil: Um olhar sobre as Unidades da Federação*. 2012. [online] Disponível emhttp://www.oitbrasil.org.br/node/880. Acesso em abril de 2015.

OUTHWAITE, W.; BOTTOMORE, T. *Dicionário do pensamento social do século XX*. 1.ed. Rio de Janeiro:JorgeZahar.1996. 974p.

R CORE TEAM (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

SANTOS, R.O. *Análise do tempo de permanência do trabalhador formal no mercado de trabalho no Distrito Federal*. Monografia(Bacharelado em Estatística) – Departamento de Estatística da Universidade de Brasília (UnB), Brasília, 2015.

SCHOENFELD, D. Partial Residuals for the proportional hazard regression model. *Biometrika*, v.69, n.1, p.329-241, 1982.

SCHWARZ, G. Estimating the dimensional of a model. *Annals of Statistics*, v.6, n.2, p.461-464, 1978.

SETRAB-DF. Mercado de trabalho aquecido no DF. 2013. [online] Disponível em <http://www.trabalho.df.gov.br/noticias/item/2273-mercado-de-trabalho-aquecido-no-df.html>. Acesso em abril de 2015.

Recebido em 09.05.2015

Aprovado após revisão em 25.08.2015

APÊNDICE – Classificação da variável CNAE

Categoria	Seção	Divisões	Descrição CNAE
Categoria 1	A	01 e 02	Agricultura, pecuária, silvicultura e exploração florestal
	F	45	Construção
Categoria 2	C	10 a 14	Indústrias extrativas
	D	15 a 33	Indústrias de transformação
	G	50 a 52	Comércio, reparação de veículos automotores, objetos pessoais e domésticos
	H	55	Alojamento e alimentação
	K	70 a 74	Atividades imobiliárias, aluguéis e serviços prestados às empresas
	O	90 a 93	Outros serviços coletivos, sociais e pessoais
Categoria 3	B	05	Pesca
	I	60 a 64	Transporte, armazenagem e comunicações
	M	80	Educação
	N	85	Saúde e serviços sociais
	P	95	Serviços domésticos
	Q	99	Organismos internacionais e outras instituições extraterritoriais
Categoria 4	E	40 e 41	Produção e distribuição de eletricidade, gás e água
	J	65 a 67	Intermediação financeira, seguros, previdência complementar e serviços relacionados
Categoria 5	L	75	Administração pública, defesa e seguridade social

Fonte: Concla (2015).